Can teachers' rewards improve educational outcomes? The role of financial and non-financial rewards

Martha Mechthilde Kluttig Vega

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

2018

ABSTRACT

Can teachers' rewards improve educational outcomes?

The role of financial and non-financial rewards

Martha Mechthilde Kluttig Vega

Inspired by the theoretical power of rewards in the labor market, to improve educational outcomes, this paper tests if giving a non-financial reward along with a financial one can result in higher student ex-post outcomes than just a financial incentive. The underlying mechanism by which non-financial reward might work is explored as well. The argument is based on Benabou and Tirole (2002)'s model, that non-financial reward may affect teachers' self-esteem and, with that, their effort, and thereby the student outcomes after the reward is given. This is accomplished by exploiting a discontinuity in the running variable used to assign the Teaching Excellence Award (AEP for its initials in Spanish). A Sharp Regression Discontinuity Design is used to identify the effect of AEP using data for more than 5,000 math and language teachers. The dataset includes the teaching evaluation score that AEP gives every year to their applicants, the corresponding standardized test score of more than 100,000 students, (SIMCE for its initials in Spanish), school characteristics, and information about motivation and self-perception that teachers self-report in a survey administrated by SIMCE along with the standardized test every year. The results show that rewarding teachers by giving a non-financial reward along with a financial one does not work in the intended way. I find a not statistically significant effect of giving a reward to teachers with outstanding teaching skills and pedagogical knowledge on student test scores, teaching practices, teacher's self-confidence in a window of three years after the certification process. Lastly, there is no evidence of teacher-student or teacher-school sorting as an ex-post effect of obtaining the certification.

Table of Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I thank Professors Thomas Bailey, Sarah Cohodes, Henry Levin, Jonah Rockoff, Carolyn Riehl, and Miguel Urquiola for comments and suggestions. I also thank Chilean Ministry of Education for sharing the dataset. Finally, I thank Sonja Karlsen and Kathryn McLellan for insightful comments and helpful edition, which greatly improved this manuscript. Special thanks to my nephews who helped me with the edition too. Financial support from the Chilean Government is gratefully acknowledged.

A special mention goes to my family. Without the unconditional everyday support of my admirable parents (Axel and Marta), my supportive, funny, and awesome siblings (Elisabeth and Christian), and my kind and smart nephews (Joaquin, Nicholas and Felipe), the dream would not have come true.

DEDICATION

First and foremost, this dissertation,

my work and heart are dedicated to Esteban and Pablo. You are the best!

Can teachers' rewards improve educational outcomes?

The role of financial and non-financial rewards

Chapter 1 - Introduction

There is a large body of evidence suggesting that teachers are a key component of the educational production function. A one standard deviation increase in teacher quality raises math achievement by 0.11-0.24 standard deviations per year and Reading achievement by 0.11-0.20 standard deviations per year (Rockoff, 2004, Rivkin et al., 2005, Aaronson et al., 2007, Kane and Staiger, 2008, Araujo et al., 2016). Moreover, teacher quality is a critically important determinant of later life outcomes as well (Chetty et al., 2011). Thus, there are many initiatives trying to foster teachers' effectiveness in order to improve education outcomes subject to the available resources.

Despite their importance, the question about how to enhance teachers' effectiveness has no trivial answer. A possible answer is to provide direct incentives tied to an observable outcome—student performance—in order to allow teachers to choose the best means to improve performance given their circumstances. In theory, a simple labor principal-agent model would predict higher effort in the presence of incentives. However, there are several reasons why incentives might not work. Neal (2011) argues that a bad design, such as setting a too low incentive or conditions too complicated to meet, might result on no effects of the incentive mechanism. Also, incentives may not work because teachers do not know how to improve (Springer, 2009). Others argue that teacher extrinsic incentives can decrease a teacher's intrinsic motivation—the individual's desire to perform the task for its own sake (Bénabou and Tirole,

2003)—which is referred as the crowding-out effect (Kreps, 1997, Deci and Ryan, 1975). Finally, incentives may lead to harmful competition between teachers in what some believe to be a collaborative environment (Moore Johnson, 1984, Firestone and Pennell, 1993, as cited in Fryer, 2013).

Empirically, there has been an increasing number of papers that study the effects of performance related pay for teachers, showing mixed results. The evidence from developing countries tends to support the effectiveness of teachers' incentives (Duflo et al., 2012, Glewwe et al., 2010, Muralidharan and Sundararaman, 2011), but these estimates are difficult to generalize to other countries due to large cross-country differences in educational systems (Imberman and Lovenheim, 2015). For developed countries, there is little evidence of performance-based pay's effects for teachers on student learning outcomes (Springer et al., 2011, Fryer Jr et al., 2012, Fryer, 2013). Moreover, financial teacher incentives have been proven to result in many unintended consequences such as cheating (Jacob and Levitt, 2003, Figlio and Winicki, 2005, Behrman et al., 2015) and teaching to the test (Holmstrom and Milgrom, 1991, Glewwe et al., 2010, Muralidharan and Sundararaman, 2011, Imberman and Lovenheim, 2015). Also, there is evidence that teachers narrow their effort toward a subgroup of students (Neal and Schanzenbach, 2010) and finally, the effects on student results fade-out (Glewwe et al., 2010, Springer et al., 2011).

A last and less explored answer to our question can be rewarding for teaching excellence. Theory derived from behavioral economics (Bénabou and Tirole, 2003) and psychology (Deci et al., 2001) supports the potential effectiveness of rewards to foster intrinsic motivation, and thereby productivity. In fact, rewards derive their intrinsic motivational power from a variety of

mechanisms including bolstering status, providing relative performance feedback, and enhancing self-esteem. Despite its potential, little is known about the empirical ex-post effect of rewards in the context of teacher-related policies.

In this context, where most of the literature has focused on the ex-ante effect of financial incentives—putting little attention on the ex-post effects of teaching rewards—, I test if giving a knowledge- and skills-based reward to outstanding teachers can result in higher ex-post student outcomes. Complementarily, I test if giving a non-financial reward along with a financial one results in higher effects than a financial reward only. At the same time, I test the underlying mechanism by which the rewards might work. To perform this analysis, I use a Chilean public program to certify teachers' quality, which is similar to National Board Certification (NBPTS) implemented in the US. I do so by exploiting a discontinuity in the running variable used to assign the Teaching Excellence Award (AEP for its initials in Spanish). I use a Sharp Regression Discontinuity Design (SRD) to identify the effect of AEP using data for more than 5,000 math and language teachers.

The AEP rewarded teachers who showed outstanding teaching skills and knowledge after they voluntarily applied to a standards-based assessment. The program followed the professional teaching standards established in the *Marco para la Buena Enseñanza* (Ministry of Education, 2004), which was developed by the Ministry of Education (MINEDUC) in collaboration with teachers' union and public schools' administrators (municipalities). The AEP reward had two components: a yearly bonus on top of their salaries (8% of salary increase) for a period that could range from two to ten years, and a non-financial component (public merit acknowledgement, and a pin). Teachers were awarded in a public ceremony hosted by the

MINEDUC and received a diploma and a program pin that they could wear in their everyday activities. Teachers may have also received public acknowledgement at their schools.

Since the certification was tied to bonuses (financial component of the reward) and public recognition (non-financial component of the reward), the program had the potential to improve school performance. The financial component could have also positively affected teacher's decisions on the optimal level of effort. At the same time, the recognition of teacher's quality may have worked as a reward, fostering teacher's self-confidence and motivation and thereby permanently increasing their effort and productivity.

All these likely effects of the AEP have not been causally addressed. Despite the absence of quantitative and causal research, a qualitative study carried out in 2012 raises questions about the effectiveness of the program and pointed out the value of the merit acknowledgement component of the program, and the importance of the monetary award (Araya, Taut, Santelices, and Manzi, 2011). In addition, the program seems to have results that might have important and unexpected consequences.

In more general terms, studying whether AEP and its financial and non-financial components had the power to increase teachers' effort, and teachers' productivity has the potential to make a significant contribution to the literature for at least four reasons. First, despite the theoretical potential benefits of rewards, there is little empirical evidence on their effects. So far, financial incentives as a mechanism to motive teachers have been widely studied showing mixed evidence on their effectiveness. Second, there is no previous research comparing financial with non-financial rewards. Third, the fact that rewards may result in persistent effects makes it interesting to study the ex-post effect of a merit reward whose influence might never end. It is a novel topic

considering that the literature is focused on the behavioral changes either before the incentive is delivered or after the incentive has been removed. In fact, the effects of financial incentives do not seem to last after the incentive has been removed. Fourth, testing whether the effect of the program fades-out over time adds new evidence to the literature that has found that the effect of monetary incentives tends towards zero once the program is over. Notice that the non-financial component of the program never ended, despite the fact that was given only once. Ultimately, the objective of this research is to identify and test the underlying mechanisms of the program and then showing whether programs that acknowledge teacher merit are effective increasing educational quality.

This document is divided into eleven chapters. The first one is this introduction. The second chapter explains the characteristics of the AEP. The third chapter is the literature review, while the fourth discusses the theoretical framework that links the program design to expected outcomes by an underlying theory. The fifth chapter aims to clearly state the research questions that are empirically addressed as explained in Chapters 7 and 8 with the data described in Chapter 6. Chapter 9 shows how the data is used and variables constructed. Finally, the estimation results and their discussion are presented in Chapters 10 and 11 respectively. Complementary appendixes are included as well.

Chapter 2 - Program description

In Chile, large efforts have been done to improve educational outcomes by fostering teachers' effectiveness through a different set of strategies of recruitment, assignment, compensation, evaluation, promotion, and retention. One of these strategies has been the recognition of teacher quality. One example of this was the Teaching Excellence Award[1] program (AEP for its initials in Spanish); that was implemented by the Chilean Ministry of Education (MINEDUC) in 2002. This program had a political and strategic purpose in the context of the educational system reform and teacher-related policies. As explained by World Bank 2018), before implementing a mandatory program for all teachers in 2004, the Chilean administration introduced AEP as a voluntary individual assessment and incentive system that set a precedent for teacher evaluation. Because these steps allowed time for adjusting and gaining support for the new system, they were key to its success. In fact, in 2012 the program went through a great transformation to keep progressing with the introduction of incentive/evaluation models. It was modified in order to reward a larger number of teachers by providing different certification levels with smaller and differentiated bonuses. Later, the AEP has been transformed again to become the base of the new teacher career ladder that the MINEDUC designed in 2016.

The AEP was a national, free, confidential, voluntary, standards-based, and multi-method teacher evaluation program that was implemented to reward the excellence of public-funded school's teachers from pre-school to high school. Specifically, the program aimed:

[1] Asignación de Excelencia Pedagógica.

... to strengthen the quality of education and to recognize and highlight the merit of teachers, fostering their retention at teaching and helping identify those that show knowledge, skills and competencies of excellence. (Law N19715, 2001, Art. 14).

In 2002 the program was created by law and their characteristics were established. First, the application requirements stated that only teachers with at least two years of experience who were teaching a minimum of 20 hours in public-funded schools could voluntary apply to the program. Second, the application process was confidential. Only the applicants would know the results of the process, and only the certified teachers would be publicly announced.

As mentioned before, the AEP was standards based following the professional teaching standards established in the Framework for Good Teaching[2] (FGT for its initials in Spanish), which defines good teaching performance (Ministry of Education, 2004). The FGT was based on Daniel-son's Framework for Teaching (Danielson, 1996), which is also used for teacher evaluation in various school districts across the United States (Heneman III et al., 2006) as well as in the Measures of Effective Teaching (MET) Project (Kane and Cantrell, 2010, Kane and Staiger, 2012, Cantrell and Kane, 2013).

The program combined two instruments to evaluate applicants: a structured portfolio, and a pedagogical knowledge written test. The instruments covered different aspects of teaching as defined by the FGT. The score for each of the two assessment instruments was continuous and took values between 1 and 4. The scores of each instrument were weighted (30% for the test and 70% for the portfolio) to obtain a final score. If this final score was equal to or greater than the

[2] Marco para la Buena Enseñanza in Spanish.

cutoff score, the teacher was classified as certified by AEP. In this case, the teacher can easily access her certificate from the program web site (see Appendix A). Regardless of the final results of the evaluation process, the evaluated teachers received the same descriptive and extensive report detailing their results (see the Results Letter in Appendix B).

The AEP worked as both evaluation and reward for those teachers who voluntarily applied for the program. On one hand the program assessed teaching quality and on the other hand rewarded teachers with a financial reward and public recognition—a non-financial reward. The economic benefit was a yearly bonus on top of their salary for a period that could range from two to ten years, if a) the teacher continued to teach in subsidized schools, and b) kept getting good teacher evaluations while the certification was valid.

The AEP financial reward was likely to be perceived as attractive. The bonus represented an 8% salary increase, which can be considered significant, especially in a context of low salaries as in Chile. According to the 2014 OECD Indicators, teachers' salaries in Chile were among the lowest for all OECD countries (OECD, 2014). Chilean upper secondary teachers earned 77% of what other tertiary-educated full-time workers earned, compared with an average salary gap of 92% across all OECD countries. Moreover, a Chilean upper secondary teacher with 15 years of experience earned USD 26,195 a year while the OECD average was USD 42,861.

The non-financial component included a certification ceremony where the teacher received a diploma and a pin (see Appendix C for more details). With great solemnity, the MINEDUC organized the ceremony. As an example of the importance put on the ceremony, in 2003, the President personally gave the certification to the first cohort of certified teachers. Afterwards the Minister of Education handed out the certificates in the ceremony held in the capital, Santiago. In

addition, in some cases, local authorities and schools also organized reward ceremonies. Importantly, in 2007 and 2008 the non-financial component of the program was not given to teachers. Certified teachers during those years did not receive public recognition. No awarding ceremonies were held, and no pins were handed out. The reason for this had to do with the bureaucratic issues that prevented the government those years from a timely purchase of needed pins. Later in 2009, both financial and non-financial rewards were given again.

The AEP successfully became the nationwide program to reward teaching quality publicly. Until 2011, more than 16,000 Chilean teachers who worked in publicly funded schools had voluntarily applied to the program and 24% of them were awarded the certification (Table 2.1).

Table 2.1. AEP applicants.

Application Year	AEP Applicants	AEP Certified	%
2002	1,906	313	16%
2003	935	409	44%
2004	1,621	522	32%
2005	1,834	632	34%
2006	2,215	626	28%
2007	1,666	341	20%
2008	1,661	315	19%
2009	1,815	319	18%
2010	1,499	258	17%
2011	1,316	272	21%
Total	16,468	4,007	24%

Despite its importance, the program effects have not been fully and causally addressed Interestingly, Araya et al. (2011) carried out a qualitative study to develop and test the program underlying action theory. In the absence of a formal theory of action, based on inputs given by key informants, the expected results of the program were set at three levels. First, at the educational system level, it was expected that the program would increase teaching salaries

according to individual performance, retain good teachers in the classroom, and encourage the collaboration with peers. Second, at the school level, certified teachers were expected to encourage their peers to be-come certified and to model good practices among their colleagues. In the medium term, certified teachers were expected to contribute to installing the concept of quality within schools. Third, at the individual level, certificates were expected to increase their salary, lead to social recognition and increase their professional self-esteem, while in the long term, certification was expected to stimulate teachers' reflection on their own teaching practices. With all this, in the long term, the AEP was expected to contribute to the improvement of educational quality. To empirically verify the action theory of the program, Araya et al. (2011) carried out interviews and focus groups with principals and teachers who applied to the program. It was intended to check if the expected effects were taking place, as well as identify effects—positive or negative—that were not originally anticipated by those who designed the AEP.

The qualitative study shows that:

- Regarding self-perception, certified teachers experienced greater professional self-esteem. All the teachers interviewed said that obtaining the certification improved their professional assessment of themselves and their sense of self-efficacy. They stated that they felt more confident in their abilities, feel responsible for doing well, and had a desire to support their peers. In addition, they experienced higher levels of conscientiousness and self-commitment, desire to improve themselves, greater interest in innovating teaching practices and adapting to the learning needs of their students, as a result of obtaining certification. The teachers were proud of themselves for having obtained a certification that is recognized in the education system as a demanding and arduous process. A certified teacher stated *"It is a personal effort, also the*

desire to want to improve, to change. That's why for me this has meant a positive change" (Araya et al., 2011, p. 311, certified teacher quote).

- Regarding perceived social recognition, teachers felt more recognized by external and distant actors in their communities than by their own schools. From the perception of certified teachers, the reaction of the school community was ambivalent: when the results of the certification process were known, peers and principals publicly congratulated the successful applicants, however, there were also signs of envy and ignorance regarding the achievement obtained. Even though teachers pointed out that they felt respected by colleagues belonging to their closest network, almost all of them reported feeling isolated or criticized by their peers, some of whom cast doubt on the legitimacy of the certification or the real merits of a certified teacher. One-third of the principals agreed with this perception. In summary, certified teachers reported having received an initial recognition from their local school community, but it was superficial, ambiguous and not long-lasting. This finding is similar to that found by the National Research Council and others (2008) in the USA. According to this research, teachers certified by the NBPTS made considerable efforts to minimize the distinctions between them and their non-certified colleagues, sometimes even hiding the fact that they have earned the credential. In this way, certified teachers avoid transgressing the egalitarian tradition that dominates the teaching field in order to avoid isolation in their schools (National Research Council and others, 2008). This evidence suggests that teachers might have increased their self-esteem, but in weak fashion.
- Certified teachers did not systematically report a perception of higher effort and improvement. One third of the principals and most of teachers interviewed perceived that the certification did not imply improved teacher performance. A certified teacher stated, "*it simply*

confirms a performance that was already good before the certification" (Araya et al., 2011, p. 312, certified teacher quote). However, one third of principals and some certified teachers registered changes in teachers' level of professionalization to the extent that they used a more technical language and had greater reflection on their own pedagogical practice.

- Certified teachers found that the financial reward was welcome, although the bureaucratic obstacles in receiving it detract from the benefits. They described difficulties to obtain it, so that it ends up losing its value as a reward. All the certified teachers who participated in the study noted permanent arrears in the payment of the bonus, attributing it to the bureaucracy and to the disinformation at all levels of the administrative system.

These findings reveal that the program might not have worked as expected, which casts doubts over the program's effectiveness to enhance teachers' behavior and performance and thereby the quality of the educational system. In fact, Araya et al. (2011) state that important questions remain unanswered, especially those regarding the effects of AEP at the educational system level, for example assessing the impact of AEP on students' achievements, taking into consideration these qualitative findings.

Chapter 3 – Incentives and rewards for teachers: are they effective? A theoretical answer from a literature review

3.1. Introduction

We know teachers matter. However, the question about how to enhance teachers' effectiveness has no easy answer. The first answer would be to pay wages to teachers based on their observable characteristics, as most school systems do. This idea has been proven to be ineffective since teacher characteristics rewarded in this way (such as experience and having a master's degree in education) are poor predictors of better student outcomes (Rockoff, 2004, Rivkin et al., 2005). A second answer would be to pay teachers based on their effort. This is especially difficult considering the moral hazard problem that makes their effort unobservable and unmeasurable for the principal. It would be prohibitively expensive to write complete contracts that specify desired actions for each potential classroom setting and then to provide the monitoring required to ensure these desired actions are taken (Neal, 2011). A third possible answer is to give teachers extra resources to allow them to improve their teaching. Even though extra resources might be a possible condition for improvements, it may not be sufficient if teachers lack the motivation to increase their effort. If teachers lack motivation to put effort into the education production functions (e.g., lesson planning, parental engagement), direct incentives to teachers may have a positive impact by motivating teachers to increase their effort, subject to the available resources. Thus, in order to allow teachers to choose the best means to improve performance given their circumstances, the best possible answer is to provide these direct incentives tied to an observable outcome such as student performance.

One way to create incentives for teachers is performance-related pay, which can be considered an extrinsic incentive because it is a recompense that is tangible or physically given to the teacher for accomplishing something positive. This kind of incentive has been increasingly implemented, showing mixed results. In fact, the implementation in the U.S. of test-based accountability systems such as No Child Left Behind (NCLB), and the Race to the Top education initiative, coupled with the poor relative performance of American students on international math and science tests, have stimulated interest in performance-related pay policies (Podgursky and Springer, 2007). However, while the idea of using incentive pay schemes for teachers as a way of improving school performance is increasingly making its way into policy, it might not be as promising as expected.

The most recent empirical evidence with improved identification of the causal impact of teacher incentives on the effectiveness of such policies shows that the evidence from developing countries tends to support their effectiveness (Duflo et al., 2012, Glewwe et al., 2010, Muralidharan and Sundararaman, 2011). However, these estimates are difficult to generalize to a context of a more developed country. This is due to large cross-country differences in educational systems (Imberman and Lovenheim, 2015). Indeed, for developed countries, there is little evidence of performance-based pay's effects for teachers on student learning outcomes (Springer et al., 2011, Fryer et al., 2012, Fryer, 2013). In addition, monetary teacher incentives have been proven to result in many unintended consequences, such as cheating and teaching to the test.

The mixed evidence on the effectiveness of teacher incentives can be discussed from a theoretical perspective. This help rationalize the empirical results that have been found. A labor

economics model, such as the theory of piece-rate compensation, would predict higher effort in the presence of incentives Lazear (2000). In this model of agency theory, introducing extrinsic incentives cannot lower effort levels. Without extrinsic incentives, effort is necessarily at the lowest possible level (Kreps, 1997). However, more sophisticated models state something different by taking into consideration the fact that jobs are not always based on relatively straightforward, observable, and measurable tasks. As shown in Holmstrom and Milgrom (1991), increased rewards for measurable outcomes can lead to either increased or decreased effort on other unobserved outcomes, depending on whether different types of effort are complements or substitutes in the production of those outcomes. Theoretically, teachers could narrowly direct their effort only at increasing scores on the formula used to determine teacher rewards. Others argue that teacher incentives can decrease a teacher's intrinsic motivation (Bénabou and Tirole, 2003), or lead to harmful competition among teachers in what some believe to be a collaborative environment (Moore Johnson, 1984, Firestone and Pennell, 1993, as cited in Fryer, 2013).

In this context, we have to wonder how the design of teacher incentive programs can be improved to assure their effectiveness and to minimize the unintended consequences. One way of addressing this question is to study the mechanisms underlying the effect of programs rather than by only estimating their treatment effect. This is the path undertaken recently by Barlevy and Neal (2012), Fryer et al. (2012), Goodman and Turner (2013) and Imberman and Lovenheim (2015). For example, Goodman and Turner (2013) examines a group-based teacher incentive scheme implemented in New York City. They investigate whether specific features of the program (i.e., the number of teachers' tests, the different degrees of accountability pressure, the teachers' lack of understanding of the bonus program's complex goals) contributed to its

ineffectiveness. Another way of addressing the question is to think about how the monetary incentives can be complemented to increase their efficacy. However, it may be that monetary incentives alone are not sufficient to foster enough teacher motivation to obtain significant positive results that endure over time.

Rewards for what teachers know and do, as an alternative to performance-based incentives, can also be considered a valid strategy to increase student outcomes after rewarding high-skilled teachers. Moreover, they can be a cost-effective tool to foster teachers' motivation and education quality. In this sense, this chapter contributes to existing literature by providing a theoretical and empirical discussion on how teacher incentives could or could not result in enhanced motivation and higher effort. Complementarily, this chapter discusses how a reward could also be an effective alternative policy compared with expensive monetary-incentive policies.

Importantly, the focus of analysis is the effect of giving a knowledge- and skills-based reward to high-performing teachers on ex-post outcomes. This is very different to the most common approach that is to estimate the effect of incentives programs focused on ex-ante or contingent outcomes, before the incentive is given. In fact, incentives are frequently offered to people as an inducement to engage in a behavior in which they might not otherwise engage (Deci et al., 2001). The offer or probability of receiving the compensation is what works as an incentive for all potential winners, while competence-based rewards work fostering motivation on those that have received the reward with probability one.

The purpose of this chapter is to systematize what can be learned from empirical and theoretical literature regarding initiatives to enhance teachers' performance, which are primarily focused on monetary incentives given to affect the teachers' performance before the incentive is

given. The goal is to understand the theory that supports the effectiveness of monetary incentives in school settings. The second goal is to review whether the empirical literature agrees with the theory. The empirical evidence is analyzed to see if teacher incentives have unintended consequences that compromise their potential benefits. Later, theories are presented to explain the empirical results. Lastly, in light of what we know about teacher incentives, the evidence supporting that rewards potential is discussed.

3.2. In theory: why should incentives work?

Why would we need incentives for teachers? The answer is moral hazard. This implies that workers' efforts are not perfectly observable. Thus, the administrator's problem is to find a contract that induces the highest effort. This is a principal-agent problem with asymmetric information. Teachers often work in an environment in which most of their actions are hidden from their supervisors. The contextual information that determines the efficient choice of actions at any point in time is also hidden to their supervisors. In this configuration, it is prohibitively expensive to write contracts that specify the desired required actions for each potential classroom setting and then to provide the monitoring required to ensure these actions are taken (Neal, 2011). Faced with these monitoring problems, educational authorities can pursue one of two strategies. They can pay teachers flat salaries and try to shape their effort through professional development and the processes used to identify and recruit educators, or they can link incentives to teacher performance. In the latter context, performance pay systems are designed to encourage greater teacher effort.

The theory underlying teacher incentive programs seem to be straightforward. It is to be expected that teachers, as with other occupations, are extrinsically motivated to some extent by

income. This is borne out by studies of teachers' responsiveness to relative salary levels (Zymelman and DeStefano, 1989, Eide et al., 2004, Chevalier and Dolton, 2004, Glewwe and Kremer, 2006. as cited in Levačić, 2009). If teachers lack the motivation to put additional effort into important inputs to the education production function, financial incentives tied to student achievement may have a positive impact by motivating them to increase their effort (Fryer et al., 2012). All this is assuming teachers have the knowledge and skills to effectively improve.

From a theoretical perspective, the labor supply model can be used to explain how incentives can increase effort. The simplest way of doing this is to analyze performance-based pay (a piece-rate salary) against flat salaries (a time-rate salary) by using the model proposed by Lazear (1986) and explained in its simplest case by Borjas (2000). From this analysis, the simple static labor supply model shows important implications given that teachers' effort cannot be directly observed, and it is not possible to closely monitor their activities. Performance-based salaries may be able to induce effort and to affect sorting of teachers among schools. Specifically, a performance-based system that ties pay to performance may be able to attract the most able teachers, elicit higher levels of effort, and increase a school's productivity.

3.3. Do incentives work? Evidence on incentives' effectiveness

It is a central theme of economics that incentives promote effort and performance, and there is much evidence they actually do (Bénabou and Tirole, 2003). However, there is a body of empirical research that does not support incentives' effectiveness for teachers. As a result, the evidence on financial incentives effectiveness is mixed. In order to provide some clarity to the discussion, the evidence can be classified according to three variables: the kind of empirical methodology used (experimental/non-experimental), economic development of the country

where the program was implemented (developed/developing), and measure used to give the incentive (relative/absolute). This helps organize the large body of available evidence as shown in Figure 3.3. There, it is possible to observe that there are no unequivocal results.

First, the evidence can be organized based on the type of methodology used, it gives us an idea of the internal validity of the results. The research on the effectiveness of teacher incentives has used experimental and non-experimental methodologies, with mixed results, regardless of method. For example, Glewwe et al. (2010), Goodman and Turner (2013) and Fryer (2013) all study school-based incentive programs using experimental methodology, but they arrive at different results. Goodman and Turner (2013) and Fryer (2013) examine a New York City school-based teacher incentive pay experiment where close to 200 high-poverty schools were randomly selected to participate in an incentive scheme that allowed the schools to choose how to allocate incentive payments. Each participating school could earn USD 3,000 per unionized teacher (3% - 7% of annual teacher pay), which the school could distribute at its own discretion if the school met the annual performance target set by the Department of Education.

Yet, despite this apparent flexibility, the vast majority of schools chose to distribute the rewards evenly. Fryer (2013) finds there were not any effects on student achievement or teacher behavior. If anything, there was a negative impact especially in larger schools where free riding may have been an issue. Complementarily, Goodman and Turner (2013) find that program had only a little effect on student achievement in schools where incentives to free ride were weakest. In contrast, Glewwe et al. (2010) show that treatment scores increased by 0.14 standard deviations relative to controls in the second year of the experimental implementation of a Kenyan school-based program that gave bonuses to schools for either being the top scoring

school or for showing the most improvement. Importantly, scores on exams not linked to incentives did not increase significantly.

Table 3.1. Summary of Evidence.

		Relative measure to assign rewards	Absolute measures to assign rewards	
		Rank-tournament Incentive Pay Program	**Test-Based Accountability System**	**Performance Based System**
Developed Country	Experimental	Roland G. Fryer et al. (2012) (0)	Fryer (2013) (0)	
		Springer et al. (2011) (0)	Goodman et al. (2013) (0)	
	Non-Experimental	Imberman et al. (2015) (+)	Neal et al. (2010) (0)	Atkinson et al. (2009) (0)
		Ladd (1999) (0)		Dee et al. (2015) (+)
				Jackson (2010) (+)
Developing Country	Experimental	Glewwe et al. (2010) (+)		Muralidharan et al. (2011) (+)
				Behrman et al. (2015) (0)
				Duflo et al. (2012) (+)
	Non- Experimental	Lavy (2002) (+)		
		Lavy (2009) (+)		

Note: (+) Program had positive and significant effects, (0) Program had no significant effects

While the first category was related to internal validity of the results, the second has more to do with the external validity of the results. The evidence can be classified by the development level of the country where the program was implemented. This variable is important since the results might be linked to the setting and context of program implementation, which affects the generalizability or external validity of the results. The evidence for developing countries is difficult to generalize in a more developed country context due to large cross-country differences. For instance, Glewwe et al. (2010) argues that agency problems between public school teachers and education authorities are often much more severe in developing countries. In

many developing countries, teachers are often absented from school and even absent from their classrooms even when they attend school. They summarize evidence from a number of developing countries and made a compelling case that educational settings are very different in many dimensions. Therefore, the effect found from an incentive program in a developing country might not be easily extrapolated to developed countries. Having said that, we can observe some kind of pattern here. As stated by Imberman and Lovenheim (2015) for developed countries, the results from several recent randomized controlled trials suggest that linking teacher pay to their students' academic performance does little to raise student achievement (Fryer, 2013, Goodman and Turner, 2013, Fryer et al., 2012, Springer et al., 2011). They do not find any significant impact of teacher incentives on student performance on average. In contrast, programs in developing countries have shown better results, but they have also shown mixed results. For instance, Behrman et al. (2015) and Muralidharan and Sundararaman (2011) implemented experimental performance pay incentive programs in developing countries–Mexico and India, respectively. Both studies were first in their purpose: Muralidharan and Sundararaman (2011) conducted the first randomized evaluation of teacher performance pay in a representative sample of schools, and Behrman et al. (2015) studied the first randomized control trial that incorporated incentive payments to both students and teachers.

Behrman et al. (2015) evaluates the impact of the ALI program, which is a large-scale social experiment that was designed to promote math achievement through performance-based monetary incentives. The program randomly assigned 88 Mexican high schools with over 40,000 students into three different incentive schemes for students, teachers, and a control group. Treatment 1 was for students and Treatment 2 was for math teachers. The annual bonus was

between 10% and 15% of the annual teacher salary in a federal high school. Treatment 3 gave both individual and group incentives to students, teachers, and school administrators, thus rewarding cooperation among all actors in the school. The third treatment included bonuses based on the performance of peers and for non-math teachers and administrators. Payments to math teachers were based on the performance of the students in their classes and on the performance of the students in all other math classes. Payments to non-math teachers and school administrators were based on the performance of all the students in the school. This mix of treatments allowed the authors to compare student and teacher incentives. They find the largest average effects for the school incentive, smaller impacts for student incentives and no impact for teacher incentives. In order to rationalize the non-significant effect of the teacher incentive from the ALI experiment, Behrman et al. (2015) developed a model of student and teacher effort choice. The model suggests that teacher incentives are not enough to make the difference; students need incentives as well. It also suggests that under specific assumptions, a teacher bonus alone may not be sufficient to induce enough students who, without the bonus, were supplying minimum effort, to supply above-minimum effort in response to an increase in teacher effort (given complementarity of student and teacher effort). A student bonus alone, given that it directly affects student incentives, can induce such a response and will also increase teacher effort. Once a student bonus is in place and students are supplying above-minimum effort, an additional teacher bonus can further augment both teacher and student effort.

On the other hand, Muralidharan and Sundararaman (2011) finds positive effects in rural India. They evaluated the effects of teacher performance incentives and school input interventions by comparing test score outcomes in treatment schools over a subsequent two-year

period to outcomes in a group of control schools. They randomly allocated schools to four treatment groups and to one control group with 100 schools in each group. One of the treatments was a performance incentive paid to teachers on the basis of the average improvement in their students' test scores. Taken as a whole, they find that paying teachers for test score increases in math and reading increased test scores without any evidence of any adverse consequences.

Finally, the studies have been categorized depending on how a program is aimed at allocating incentives. The monetary incentives can be allocated based on either relative or absolute teacher performance. Among the programs based on absolute standards to identify incentive beneficiaries, we find both assessment-based accountability systems and performance-pay systems. Assessment-based accountability systems are promoted as vehicles for holding public schools accountable for their use of public funds. These systems define students' achievement standards and then measure schools' performance. This is by using metrics that describe the degree of discrepancy between the standards set by the accountability systems and student achievement in various schools. Furthermore, these systems include a set of sanctions and rewards that school administrators and teachers face if their students fail or do not meet the performance targets set by the accountability system (Neal and Schanzenbach, 2010). Finally, the primary objective of most accountability systems is performance measurement rather than performance incentive, which explains why they typically contain rewards and sanctions that are either not spelled out in detail or are less than credible because they cannot be enforced ex post (Neal and Schanzenbach, 2010). In contrast, performance-pay systems are more explicitly focused on incentive provision and often contain precise mappings between student performance and educator compensation and employment status (Neal, 2011). These performance-pay

systems have shown more positive results. For instance, Dee and Wyckoff (2015) studies the IMPACT Program, the teacher-evaluation system introduced in the District of Columbia's public schools. IMPACT implemented uniquely high-powered incentives linked to multiple measures of teacher performance (i.e., several structured observational measures as well as test performance). The regression discontinuity results indicate that financial incentives further improved the outcomes of high-performing teachers.

In the group of programs that are based on relative performance to identify their beneficiaries, we find rank-order tournament incentive pay programs. While test-based accountability systems and performance-based systems assign recompenses to those who accomplish an absolute goal, the rank-order tournament incentives assign the rewards to the best performances in a distribution of outcomes. The reliance on relative performance measures means that some teachers will win, and others will lose by design. So relative performance may encourage competition among teachers resulting in increased outcomes or detrimental collaborative behavior among them (Neal, 2011).

There have been three studies of programs that involve both competition among educators for a fixed set of prizes and the use of value-added models (VAM) to rank schools or teachers. Ladd (1999), Lavy (2002) and Lavy (2009), all contain evaluations of experimental relative performance pay schemes. All three studies find that these programs generated significant increases in measured achievement among students, but all of them also report significant heterogeneity in estimated treatment effects for different sub-populations. For instance, in Ladd (1999) pass rates on standardized reading and math tests increased significantly, but only for white and Hispanic students. Black students do not exhibit significant gains relative to untreated

schools. However, despite the fact that Lavy (2002) and Lavy (2009) employed several empirical strategies that attempted to pin down the causal impacts of these programs, none of these programs involved random assignment of schools or teachers to treatment.

Experimental evidence shows no effects of a rank-tournament incentive pay program. First, we have Springer et al. (2011) who evaluated a 3-year pilot initiative conducted in the Metropolitan Nashville School System from the 2006-2007 school year through the 2008-2009 school year. Middle school math teachers who volunteered to participate were randomly assigned to the treatment or the control group. Teachers in the treatment group could earn bonuses of USD5,000, USD10,000, or USD15,000 for surpassing the 80%, 90%, and 95% threshold, respectively, in the historic distribution of value-added scores. Springer et al. (2011) finds there was not any significant treatment effect on student achievement or on measures of teachers' response such as teaching practices. Second, Fryer et al. (2012) conducted an experiment in nine schools in Chicago Heights, IL, to study the effect of teacher incentives when the timing and framing of the reward payment varied significantly. They find similar results. At the beginning of the school year, teachers were randomly selected to participate in a pay-for-performance program. Performance was incentivized according to the "pay for percentile" method developed by Barlevy and Neal (2012), in which teachers were awarded according to how highly their students' test score improvement ranked among peers with similar baseline achievement and demographic characteristics. The expected value of the reward (USD4,000) was equivalent to approximately 8% of the average teacher salary in Chicago Heights. Consistent with much of the literature, they find no significant impacts of this incentive.

Having covered the most recent research regarding the effect of financial teacher incentives allows us to understand the meaning of mixed results. Regardless of the nature of the incentive program, whether it is the setting that conditions its results, or the methodology applied to evaluate the program, there is not any systematic evidence supporting the unequivocal effectiveness of teacher incentive programs. Moreover, as discussed in the next section, these programs have additionally proved to have many other consequences that go beyond their primary purpose.

3.4. Unintended consequences

Going beyond the average treatment effect of the program, the incentives have shown to have unintended consequences. The literature explores many ways that schools may seek to inflate their assessment scores without actually increasing all students' subject mastery (Neal and Schanzenbach, 2010). This section covers the most important unintended consequences. This discussion has attracted much attention because these unintended consequences might have significant impact on the overall functioning of the educational system, making it essential to address them in program design to prevent their appearance.

The first unintended consequence is that the program effect does not reach all the students. How do teacher incentives influence the decision of "on whom" to put the effort? Depending on how the incentives are set, teachers might decide to concentrate their effort on specific groups of students, reducing their interest in helping all their children learn. To study this, Neal and Schanzenbach (2010) uses the introduction of two separate accountability systems in Chicago Public Schools. There was a district-wide system implemented in 1996, and the introduction of NCLB in 2002. They use this to investigate how the rules that accountability systems use to turn

student test scores into school performance rankings determine how teachers allocate their efforts. Specifically, they analyze how teachers may have incentives to concentrate on subsets of students. In their case, it was students near cutoff values that determined whether school-level goals were met. They show that as performance measures, the use of proficiency counts provides strong incentives for schools to shift more attention to students who are near the proficiency standard. However, it provided weak incentives to devote extra attention to students who are either already proficient or to those who have little chance of becoming proficient in the near term. The authors show that this evidence is consistent with previous research (Gillborn and Youdell, 2002, Booher-Jennings, 2005, Reback, 2008, Springer, 2008).

The second unintended consequence of teacher incentive programs is that the program effects do not include the acquisition of broad skills. In this case, teachers decide whether they respond to incentives by promoting broad human capital acquisition or by narrowly focusing on skills and actions that raise scores on the formulas. The latter is what (Koretz, 2002) refers to as coaching. Coaching involves activities that improve scores on a given assessment without improving student mastery of a subject (Koretz, 2002). This behavior can be understood from a theoretical perspective by using the agency model with multiple tasks developed by Holmstrom and Milgrom (1991). Based on their model, they warn that when workers perform complex jobs involving many tasks, pay-for-performance schemes based on objective measures of output often create incentives for workers to shift effort among the various tasks they perform in ways that improve their own performance rating but hinder the overall mission of the organization. Applied to the educational context, teachers could narrowly direct effort at increasing scores on the formula used to determine teacher rewards at the expense of effort aimed at broader, longer term

increases in their students' human capital. Holmstrom and Milgrom (1991) notes that coaching in response to test-based accountability systems is an obvious example of this phenomenon.

Much of the existing empirical literature on assessment-based accountability focuses on whether the test score increases that might follow the introduction of such systems represent actual increases in subject mastery. If coaching is the case, we should observe that teachers might respond to incentives by devoting more class time to topics listed in the curriculum and stressed on related high-stakes assessments. Scores on these assessments may then rise substantially while scores on broader assessments of the same subject may show only modest improvements. This is exactly what Glewwe et al. (2010) finds while investigating a school-based teacher incentive experiment in rural Kenya. They find that the program created large score gains on government tests, but no improvements on the low stakes exams. These results are consistent with the fact that teachers responded to the program by increasing the number of test preparation sessions held for students while there is no evidence of improvements in teacher attendance or classroom practice. Another piece of evidence is provided by Imberman and Lovenheim (2015). They use of a series of non-incentivized exams to test whether the impacts of the incentive program occur only on the directly incentivized exams. This analysis indicates that teachers may be teaching to the test rather than increasing general knowledge of the students. In the case of India, Muralidharan and Sundararaman (2011) provides evidence of coaching behaviors surrounding the positive effects of the incentive program. Teachers in treated schools assigned more work and conducted classes beyond regular school hours, and part of the extra class time was devoted to taking practice tests.

The third unintended consequence of the teacher incentive programs is that the program diverts teacher efforts from producing higher learning towards wasteful activities. Although coaching—the second unintended consequence—is typically not an optimal allocation of teacher effort, some forms of coaching may generate some lasting human capital gains for students. If coaching activities reflect a reduction of teaching time losses[3] on the part of teachers rather than reductions in effective teaching time, it is possible that these incentive schemes are improving educator performance (Neal, 2011). Nonetheless, the literature has documented other ways that some teachers respond to assessment-based incentive schemes that are almost certainly unproductive from a social perspective such as cheating. Some researchers have found that high-stakes testing can lead teachers or administrators to engage in cheating behaviors (Jacob and Levitt, 2003, Figlio and Winicki, 2005). Jacob and Levitt (2003) provides evidence that some teachers or principals in Chicago actually changed student answers after high stakes assessments in the 1990s. They estimated that 4-5% of Chicago elementary school teachers help their pupils cheat, and that this cheating increased after the introduction of high-stakes testing. In the context of the introduction of teacher incentives in Mexico, Behrman et al. (2015) finds evidence of cheating as well. A comparison of the ALI program impact estimates to those of prior studies reveals that the treatment effects associated with the ALI treatments in which students received incentives are quite large, especially for the treatment in which both students and teachers receive incentives. However, close examination of the textbook answer patterns shows that part of the reason for higher test scores in the treatment group was a higher rate of cheating (in the

[3] This means that coaching may imply that teachers reduce those lost periods during the day. For instance, teachers may allocate an excessive and unnecessary number of hours on hygiene, bathroom and feeding activities.

form of student copying) than in the control group, particularly in higher grades and in later years of the program.

The fourth unintended consequence of the teacher incentive programs is that the program effect fades out. The purpose of an incentive program is to enhance behavior and foster the accumulation of human capital, which, if real, should endure over time. However, the evidence has shown that the effects of incentives tend to fade. Glewwe et al. (2010) conducted a randomized trial over a 2-year period that provided incentives to primary school teachers based on student performance on district-level exams in seven subjects. Students in treatment schools had higher test scores in the second year of the program, but the gains dissipated by one year after the program ended. Thus, the test preparation sessions and other activities that generated the measured improvements in high-stakes test performance during the program did not generate lasting improvements in test-taking skills or knowledge specific to the government exams. They interpret these results as being consistent with teachers expending effort toward short-term increases in test scores but not toward long-term learning. Complementarily, Springer et al. (2011) evaluated a 3-year pilot initiative on teacher incentives conducted in the Metropolitan Nashville School System. The program involved 5th through 8th grade math teachers. There was some evidence of achievement gains in 5th grade math in years two and three, but these gains did not persist over the next school year. Along the same line, Rothstein (2010) and Carrell and West (2010) find that teachers' impacts on test scores fade out very rapidly in subsequent grades. All this evidence is consistent with the fact that rewards (extrinsic motivation) might have a limited impact on current performance and may reduce the agent's motivation to undertake similar tasks

in the future, as suggested by the Intrinsic Motivation Theory developed by Bénabou and Tirole (2003).

Theoretical and practical solutions have been suggested to address some of these unintended consequences. Barlevy and Neal (2012) provides a theoretical work on optimal teacher incentive pay. They develop the “pay per percentile” mechanism: ordinal comparisons between each of a teacher’s students and the student’s peers. Because this is a relative performance system where performance thresholds are endogenously determined through peer comparisons, they argue that such a scheme does not promote focusing on a specific group of students, but on enhancing the learning of all students. In addition, the pay-per-percentile mechanism allows for free employment of assessments without repeated items and common formats. Much research demonstrates that, while repeated items and common formats make scale integrity possible in theory, these features also invite the coaching behaviors that undermine scale integrity in practice. The policy recommendation made by Barlevy and Neal (2012) follows from this argument. The assessment used to calculate the "pay per percentile” should be a test with simpler psychometric characteristics, making it hard to be anticipated by teachers. This reduces the chances of coaching and cheating. For instance, tests with no repeated items can make it very hard for a teacher to produce tests that prepare and train the students.

3.5. Why financial incentives don’t seem to work as expected?

While simple labor supply models predict that incentives should increase effort, the empirical evidence is ambiguous. Why teacher incentives fail to operate in the desired manner? For instance, teachers may not know how to increase student achievement, or the production function has important complementarities outside their control. Additionally, the incentives are either

confusing or too weak, and teacher incentives may not have any impact on achievement. Even though there can be many reasons behind the ineffectiveness of financial incentives, those that have been studied from an empirical or theoretical perspective are discussed below, by grouping the reasons into two types: form and substance. In terms of form, the design of the teacher incentive program can determine its potential effectiveness. In terms of substance, a theoretical model might highlight how the design of a program matters and explain why incentives are not so effective as expected.

Why financial incentives don't seem to work as expected? Answers related to the program design

Many "form" aspects of program design might influence effectiveness. However, only a few of the design features have been studied. One important feature regarding the design of a teacher incentive program is how the prize is allocated: individual or school based. The theoretical prediction of the relative effectiveness of individual versus group teacher incentives is ambiguous. As explained by Muralidharan and Sundararaman (2011), school incentives could induce free riding and thus be less effective than individual incentives (Holmstrom, 1982). Free riding occurs in group-based incentives because each worker has a temptation to reduce effort and consume more leisure in response to the expected benefit received from the effort of others in the group (Imberman and Lovenheim, 2015). However, social norms and peer monitoring (which may be feasible in small groups of teachers) may enable community enforcement of the first-best level of effort. In this case the costs of free riding may be mitigated or eliminated (Kandel and Lazear, 1992, Kondor, 1992). Finally, if there are gains for cooperation or complementarities in production, then it is possible that group incentives might yield better

results than individual incentives (Itoh, 1991, Hamilton et al., 2003). The relative effectiveness of group and individual teacher performance pay is therefore an empirical question, which has been directly addressed by Muralidharan and Sundararaman (2011), Fryer et al. (2012), and Behrman et al. (2015). They study school and individual incentives simultaneously by using experimental methodology, obtaining mixed results. Muralidharan and Sundararaman (2011) studies group and individual incentives in the same field experiment over two full academic years. The study was conducted by randomly allocating incentive programs across a representative sample of 300 government-run schools in rural Andhra Pradesh, India, where 25% of teachers were absent on any given day. They find that school-level group incentives and teacher-level individual incentives performed equally well in the first year, but the individual incentive schools outperformed group incentive schools after two years of the program. In the same vein, Fryer et al. (2012) conducted an experimental pay-for-performance program in which they randomly assigned teachers to receive either individual or team rewards. Examining the individual and team treatment separately, the estimated effects were identical. Conversely, Behrman et al. (2015) finds that individual incentives had no effect, but group incentives for teachers have an effect on students test scores only when accompanied by student incentives.

Within group-based incentives, a focus of study has been on whether the effect of the group-based incentives vary with the strength of the incentive, which has been measured either as the group size or as the percentage of students in a group that a teacher instructs. Both measures have shown to be important in determining the effectiveness of group-based teacher incentives programs. Goodman and Turner (2013) used the variation in the number of math and English teachers in each school in a school-level randomized teacher incentive pay experiment in New

York City, to examine the effect of group size. They present suggestive evidence that the group-based structure of the program may have been detrimental in the majority of schools where the number of teachers was large. A lack of monitoring, as well as the diffusion of responsibility for test-score gains, may have diluted the incentives of the opportunity to earn bonuses. Conversely, the program improved math achievement in schools that had fewer teachers responsible for tested students or that had a more cohesive group of teachers. Another focus of study has been whether the effect of the incentives varies with the percent of students in a group a teacher instructs. Under a group incentive scheme, the share of students instructed by a teacher is a strong proxy for incentive strength because as the teacher's share increases, the teacher's impact on the probability of award receipt rises and free-rider incentives decline. Specifically, Imberman and Lovenheim (2015) focuses on whether teachers who are responsible for a larger share of students in each grade and subject generate more achievement gains after implementation of the award system than those who are responsible for teaching fewer students. Taking advantage of the variation in the share of students in a subject-grade that a teacher instructs, which proxies for incentive strength, they identify how the effect of this share changes when the incentive pay program is implemented. This was using a difference-in-differences methodology. They find that achievement on incentivized exams improved when incentives were strengthened.

Reward timing has been also studied. Fryer et al. (2012) conducted the first field experiment on teacher incentives that exploited the power of framing it as the presence of loss aversion. The experiment gave some teachers individual-based award bonuses and other teachers fixed cash payouts prior to the school year that needed to be returned if performance was low. Consistent with much of the literature on the effectiveness of teacher financial incentives in developed

countries, they do not find any significant impacts from the first group, but they do find improvements from the second group, when teachers have to pay back an earlier bonus payment for poor performance. The math test scores increased between 0.2 and 0.4 standard deviations. This is equivalent to increasing teacher quality by more than one standard deviation. Thus, the authors suggest that loss aversion is a more powerful incentive than standard pay for performance.

Many important features of the incentive programs are still not studied. First is the size of the reward. As explained by Glewwe et al. (2010), larger incentives could lead to more efforts focused on broad acquisition of human capital, but, of course, larger incentives could also induce wasteful or even harmful signaling effort such as cheating on tests or forcing weak students to drop out. Additionally, individual-level teacher incentives might undermine cooperation within schools. The second unstudied feature is how the standards are set. Political forces often create pressure for "high standards" in education, but these pressures can be counterproductive. Although it is clearly wasteful to set standards too low, standards well beyond what is possible may not induce any additional effort from teachers (Neal, 2011). A good example of this issue would be the POINT program that allowed math teachers in the 5th through 8th grades in Nashville, TN, to volunteer for a performance pay program. This program was studied by Springer et al. (2011). They find there was no significant treatment effect on student achievement or on measures of teachers' response, such as teaching practices. According to Neal (2011), this finding might be explained by the fact that POINT may have set targets so high that teachers responded optimally by doing roughly what they had done before. Around half of the teachers in the experiment faced less than a 20% chance of winning a bonus based on their past

performance. The third unstudied factor is relative or absolute standards. As discussed above, the distinction between the relative and absolute distribution of the rewards can be critical for the design of the teacher incentive programs. It is simply not clear whether rank-order tournament incentive pay programs, assessment-based accountability systems, or performance pay systems are the most effective. In fact, there is no empirical research comparing the effectiveness of relative versus absolute teacher incentives systems.

The evidence provided so far makes clear how sensitive the effectiveness of the programs is to the conditions of its application and its features. The conjunction of school characteristics, the education production function, and the incentive design determine the potential effectiveness of such programs. Importantly, more research is needed to get a deeper understanding of whether specific program features contributed to its ineffectiveness. Despite the lack of research, theories have been developed to explain why incentive programs do not always bring the expected results.

Why financial incentives don't seem to work as expected? Answers related to the theory of change behind the incentive programs

So far, the "form" reasons behind the inconsistent effectiveness of financial incentives have been discussed. Now, the reasons that have been studied from a theoretical perspective to address that inconsistency are discussed below. Financial incentives can work as extrinsic sources of motivation, which may imply that the worker is willing to make a larger effort to reach the award linked to the incentive plan. However, teachers also have intrinsic sources of motivation (the individual's desire to perform the task for its own sake), which may be substituted by the extrinsic sources of motivation brought by the financial incentive. This substitution can be

crucial since teacher performance is strongly predicted by the level of intrinsic motivation. The relationship between the extrinsic incentive and intrinsic motivation of the teacher can help explain why financial incentives do not always positively affect performance in the long term.

By using a meta-analysis focusing on the inter-relationship among intrinsic motivation, extrinsic incentives, and performance, Cerasoli et al. (2014) shows that intrinsic motivation is positively related to performance. Intrinsic motivation is a medium-to-strong predictor of performance. They also show that the relation between intrinsic motivation and performance is stronger for quality-type tasks than for quantity-type tasks. Tasks emphasizing performance quality, such as teaching, will have a strong link to intrinsic motivation. The reason is that quality-type tasks tend to require a higher degree of complexity and engagement of more skills, which commands greater personal investment.

Having stated the great importance of intrinsic motivation for teaching performance, Kreps (1997), gives a more intuitive idea about how incentives may harm teachers' intrinsic motivation to teach, which is consistent with the null effects found in empirical research. He explains that imposing extrinsic incentives changes the individual's utility for the work. If an employee undertakes some effort without the spur of some extrinsic incentive, he will rationalize his effort as reflecting his enjoyment of the task. Since he enjoys it, he works harder at it. But if extrinsic incentives are put in place, he will attribute his efforts to those incentives, developing a distaste for the required effort. In this case, the extrinsic incentive is "crowding-out" the intrinsic motivation.More recently, Bénabou and Tirole (2003) inspired by psychology literature, formalizes Krep's intuition and developed a theoretical model on why incentives may not work, and if they do, why the effect fades out in the long run. Bénabou and Tirole (2003) examines a

motivational crowding out phenomena in a theoretical principal-agent model. Discussed here is one of their interpretations, where the agent is a teacher and the principal is a school authority. The teacher has imperfect knowledge about his own ability. His own self-confidence is defined as the belief the teacher has about the probability of succeeding in a task where effort and ability are complements. Meanwhile, the principal has knowledge about the ability of the teacher to succeed in the task and wants the teacher to pursue the task. However, the teacher will pursue the task and resist distractions (such as planning a class rather than having more leisure hours) only if he has high enough self-confidence that he will succeed in the task. Of course, in the short run, the principal can motivate the teacher to do so by giving an incentive for success. The incentive influences the teacher's motivation through two channels. First, the incentive increases the direct payoff the teacher has from succeeding in the task. Second, the incentive affects the self-confidence of the teacher via an inference process, where he takes the incentive as a signal of the principal's knowledge of his ability. Here, a large incentive is bad news for the teacher, because he understands that the principal would offer a lower incentive if he were more able. That is, a higher incentive reduces the teacher's self-confidence and thereby his intrinsic motivation, which, in turn, lowers his effort once the incentive is no longer given.

A substantial body of experimental and field evidence, but not within the teacher labor market, indicates that extrinsic motivation can conflict with intrinsic motivation. On one side, we have Deci and Ryan (1975)'s experiment where college students were either paid or not paid to work for a certain time on an interesting puzzle. Those in the no-pay condition played with the puzzle significantly more in a later unrewarded "free time" period than paid subjects, and they also reported a greater interest in the task. More recently, Visaria et al. (2016) finds that

incentives might have significant effects on student motivation. In an experiment in non-formal schools in Indian slums, a reward scheme for attending a target number of school days increased average attendance when the scheme was in place, but it had heterogeneous effects after it was removed. Among students with high baseline attendance, the incentive had no effect on attendance after it was discontinued, and test scores were unaffected. Among students with low baseline attendance, the incentive lowered post-incentive attendance, and test scores decreased. For these students, the incentive was also associated with lower interest in school material, and lower optimism and confidence about their ability. Nonetheless, more recent empirical studies have found evidence that financial incentives do not always result in decreasing intrinsic motivation. Bettinger (2011) studies the experimental implementation of a pay-for-performance program for primary school children in Coshocton, Ohio. Even though the primary focus of the paper was on measuring the effects of Coshocton's program on student achievement, one of its aims was to reconcile some of the recent findings in economics with the established literature on psychology on the impacts of external incentives. The author finds that students' intrinsic motivation was not significantly lower as a result of participating in the program. Cerasoli et al. (2014) also shows that the intrinsic motivation was less important to performance when the incentive was directly tied to performance, including some evidence of the "crowding out" effect of the extrinsic incentives. When incentivized, the relationship between intrinsic motivation and performance is negatively moderated (weakened) by the presence of directly performance-salient incentives.

As previously stated, there are many possible reasons that can result in ineffective teacher incentives. From the empirical evidence, we can see that conditional on the school setting, the

program's design and features can make the difference. Furthermore, from the theoretical perspective, we can learn that extrinsic incentives might be ineffective due to the crowding out of self-confidence and intrinsic motivation, which are highly related to teaching performance in the long run. Also, we can learn that extrinsic incentives might have negative consequences due to the multitasking nature of teaching. In consequence, and taking theoretical and empirical evidence into consideration, one could suggest studying reward programs with specific features that at least theoretically limit the chances of ineffectiveness. Thus, in the following section is a model that suggests reward that might be effective in increasing teachers' effort and therefore learning. Also provided is empirical evidence to support this argument. This evidence comes from fields other than the teacher labor market, underscoring the research gap in the literature.

3.6. Why rewards may have an ex-post effect? The value of status, self-esteem, and feedback

Having summarized the evidence on the effectiveness of incentives for teachers, we still need to address how to motivate teachers. Specifically, how can teachers be extrinsically incentivized to exert higher effort in the long term without crowding out their intrinsic motivation and creating unintended consequences? Taking into consideration that psychologists and management scholars have long recognized that workers are motivated by more than just monetary extrinsic rewards (Etzioni 1964, Deci and Ryan 1975, as cited in Blanes i Vidal and Nossol (2011)), a potential answer to the question is excellence teacher reward. Many additional reasons support this answer. They are less likely to reduce intrinsic motivation (Frey, 2007) due to the value is given to status, self-esteem, and feedback. They might even increase intrinsic motivation (Deci et al., 2001). Because of that, their effects on effort might not fade out over time. Non-financial

rewards are also an attractive solution because they are potentially cost-effective (Levitt et al., 2012), especially when it is more difficult to formulate specific contracts ex ante, and to monitor *ex post* (Frey, 2007, Frey and Gallus, 2017).

As an extrinsic motivation, rewards can take many forms, such as a merit recognition or award. Awards are designed to give recognition to those who are thought to best exemplify the norms/goals promoted by the award giver (Frey and Gallus, 2017). For instance, awards play a significant role in the arts, sports, as well as in the business sector. In the educational setting, one could suggest giving a prize to teachers with high performance.

Rewards, and especially non-financial ones, derive their intrinsic motivational power from a variety of mechanisms including bolstering status, providing relative performance feedback, and enhancing self-esteem. For individuals who care about status and a positive self-image, non-pecuniary awards carry additional utility when they remind oneself and others of one's own special achievements (Huberman et al., 2004, Ariely et al., 2009). Formally, and for workers in general, Besley and Ghatak (2008) studies the theoretic role of the preference for status as a source of motivation. To do so, they developed a model with moral hazard and limited liability that limits the ability of an organization to achieve its desired effort level using monetary incentives. The model allows the principal to introduce a purely nominal award, which works as a positional good to the agent in the event he produces high output for the principal. This could be a job title change (promotion from associate to full professor in academia) calling some employees "employee of the week," or an award to those teachers who certify their excellence in teaching. The model assumes that giving the worker a positional good has a zero marginal-cost and that the positional good is valued by the agent. However, the extent of the conveyed status

depends on how scarce the award is, and it requires a well-defined rule that awards only the deserving. The implications of the model are that status incentives increase effort while reducing the optimal level of financial incentives. The model also predicts that the case for status incentives may be stronger when the problem of measuring the worker's output is more severe, such as in schools. This has not been empirically proven.

The second mechanism by which non-financial incentives channel their influence goes through the performance feedback that they provide. As mentioned by Blanes i Vidal and Nossol (2011), Deci (1972) and Anderson et al. (1976) find that positive feedback increases intrinsic motivation implying that the feedback component of the rewards can enhance motivation as well. In a schooling context, Azmat and Iriberri (2010) and Tran and Zeckhauser (2009) argue that learning about relative performance leads to higher student effort. Azmat and Iriberri (2010) shows how relative performance feedback raises high school students' educational attainment using data from a naturally occurring change in Spanish schools.

Perhaps more importantly, the third underlying mechanism of non-financial rewards is its power on teachers' self-esteem. They consider self-confidence as a source of intrinsic motivation which, if enhanced, can increase effort in the long run.

The psychology literature with the Cognitive Evaluation Theory (CET), developed by Deci and Ryan (1980), and explained for the educational context in Deci et al. (2001), states that "*The underlying intrinsic motivation is the innate psychological need for competence and self-determination*" (Deci et al., 2001, p. 3). Thus, if rewards influence the people's perception of competence and self-determination, they will also affect motivation. This seems to be completely feasible considering the theoretical model developed by Compte and Postlewaite (2004). They

demonstrate that in a world where performance depends on emotions, higher self-esteem enhances welfare. The authors compared the two effects of confidence: the harm of being overly confident and the benefits of being optimistic about one's own capacity. It will be the case that agents with biased perceptions will have excessively optimistic beliefs and consequently will be induced to undertake activities they should not have. On the projects they undertake, however, their optimism leads to higher performance. They showed that when confidence affects performance, it is no longer true that correct perceptions maximize long-term payoffs. In theory, having some degree of optimism is preferable to correct perceptions, showing the power of manipulating the self-esteem of agents.

Bénabou and Tirole (2005) provides the first formal theoretical model in economics that investigated the maintenance and enhancement of self-confidence. Simply put, Bénabou and Tirole (2005) suggests that an overly positive view of one's ability may be an important motivational factor, because ability and effort are complementary factors in educational production. Greater self-confidence makes teachers believe their effort will be very productive, which in turn enhances their motivation to study/teach.

Wang and Yang (2003) and Filippin and Paccagnella (2012) further apply these ideas. Wang and Yang (2003) theoretically investigates the notion of self-confidence in an economic model of education where students care both about their grades and about their own perception of their ability. The grading system determines how much information a grade conveys about ability. This thereby influences self-confidence, which in turn affects the choice of effort through the complementarity described above. If students care primarily about their perceived ability, then strong competition induced by relative grading may actually lead to low effort, even from high-

ability students. Competition limits the number of good grades, making it less likely that a student gets favorable feedback about her ability if she works hard. To protect a prior positive self-image, a student can put in low effort, which makes the grades relatively uninformative about ability and allows the student to maintain her self-image no matter what happens. Filippin and Paccagnella (2012) explores in a theoretical model how a small initial difference in self-confidence can result in diverging patterns of human capital accumulation, even when students start off with the same level of initial ability.

A more recent branch of behavioral economics has explored the effectiveness of non-financial rewards showing their potential. Symbolic awards have been studied by Kosfeld and Neckermann (2011) who hired students to enter data for three weeks as part of a non-governmental organization project. The treatment was to honor the best performance publicly with a symbolic award. They find that the award treatment increased performance by 12%.

These types of non-pecuniary benefits may be particularly potent in the context of recognition for school performance. Levitt et al. (2012) directly compares the effects of financial and non-financial rewards on short-term student effort and performance. They also investigate the effectiveness of low and high financial incentives (USD10, USD20), and compare these to the impact of non-monetary rewards, specifically, an achievement trophy. These incentives were presented in either the gain or the loss domain and were offered either immediately after the test or with a delay of a month after the test. They find that giving primary school students a trophy lead to increased performance as did financial rewards in the range of USD10 to USD20. Complementarily, Jalava et al. (2015) examines the effects of non-financial rewards on student effort on a math test. They conducted a randomized field experiment on more than a thousand

sixth graders in Swedish primary schools, finding significant differences in test scores between the intrinsically motivated control and the extrinsically motivated treatment groups. Test performance is significantly higher for students receiving a symbolic reward. The motivational strengths of the non-financial rewards differ across the skill distribution and with respect to gender. Specifically, only girls were motivated by a symbolic reward. They find that extrinsic non-financial incentives played an important role in motivating highly-skilled students to exert more effort, but symbolic rewards tended to crowd out intrinsic motivation for low-skill students.

Consequently, a reward program such as the Teaching Excellence Award[4] (AEP for its initials in Spanish), with its non-financial component, may enhance teachers' self-confidence resulting in a higher effort. This assumes the teacher knows which kind of efforts are productive in terms of increasing students learning.

Nevertheless, the final ex-post effect of rewards can also be unexpected. In making predictions about reward effects on intrinsic motivation, the CET analyzes whether the reward is likely to be experienced as informational or controlling. The informational aspect of the rewards conveys self-determined competence, and this enhances intrinsic motivation. In contrast, the controlling aspect—if rewards are experienced as the reason for doing the task—prompts a low perceived self-determination, which undermines intrinsic motivation. In consequence, depending on the features of the extrinsic incentive, this can result in the crowding out of intrinsic motivation. For instance, we have the case of performance-contingent rewards. They have strong

[4] Asignación de Excelencia Pedagógica.

control aspect that undermines intrinsic motivation. However, performance-contingent rewards can also include positive information when the reward signifies excellent performance. In those cases, rewards give information that affirms competence and, thus, offsets some of the negative effects of control. Deci et al. (2001), via meta-analyses, finds that tangible rewards significantly and substantially undermined intrinsic motivation for the rewarded activity. According to CET, another factor that is expected to influence the effects of performance-contingent rewards is the interpersonal context. In terms of education, it suggests that when rewards are used in the classrooms and schools, it is important that the climate of the classroom be supportive so that the students and teachers are less likely to experience the rewards as controlling.

So far in this chapter, theoretical and empirical arguments have been provided to suggest that teacher effort can increase after a reward is given. There are three plausible theoretical channels by which non-financial rewards can have influence. The fact is that workers' value status creates the opportunity to motivate them by offering a public valued reward. Receiving positive feedback about one's performance can increase workers' motivation and thereby increase effort. More importantly, non-financial rewards can help teachers increase their self-confidence and their effort. Furthermore, the empirical evidence emerging also supports the potential of non-financial rewards. However, the review of the literature on non-financial rewards for teachers shows that they have not been explored despite its potential.

3.7. Final remarks

The effectiveness of financial teacher incentives remains unresolved. Moreover, and despite the fact that this kind of program has made its way into the education system, the fine-tuning of teacher incentive programs is still inconclusive. These topics remain important for researchers

and especially for policy makers. The incentive programs for teachers can be expensive taking into consideration they represent an increase in the cost of school staff, which accounts for the majority of spending on in-school education—on average 80% in OECD countries (OECD, 2015). Beyond analyzing how the features of a financial incentive program can be tuned, there is a need to look for cost-effective mechanisms to improve teaching. In this context, the non-financial rewards for teachers seem to be a tool worthy of being considered as potential way to make incentives for teachers more effective.

Chapter 4 - Theoretical framework. Why AEP may (not) work?

The Teaching Excellence Award (AEP for its initials in Spanish) can be seen a reward program that involves to components: one financial and another non-financial. To improve our understanding of the program's net effect on ex-post outcomes is necessary to study how each of them may work. The bonus, based on the theoretical and empirical evidence reviewed, is expected to have an ambiguous effect on effort. Based on the Cognitive Evaluation Theory (CET) (Deci et al., 2001) and Motivational Theory (Bénabou and Tirole, 2002), the non-financial component of the program may work in a positive and negative direction as well. These models are described in the following section, making clear that the effect of both components of the program have the potential of increasing effort, but it is not equivocal. In consequence, the effect of each component and the total net ex-post effect of the full program is ultimately an empirical question, whose answer comes to narrow a gap in the existing reward literature.

4.1. The Cognitive Evaluation Theory

Using the Cognitive Evaluation Theory (CET), AEP can be considered a performance-contingent reward because it is linked to teachers' performance, with strong controlling and informational aspects. In this theoretical framework, a program like AEP,

> *...has the potential to affect intrinsic motivation in two ways, one quite positive and one quite negative. Performance-contingent rewards can maintain or enhance intrinsic motivation if the receiver of the reward interprets it informationally, as an affirmation of competence. Yet, because performance-contingent rewards are often used as a vehicle to control not only what the person does but how well he or she does it, such rewards can easily be experienced as very controlling, thus undermining intrinsic motivation. According to CET, it is the relative salience of the informational versus controlling aspects of performance-contingent rewards which determines their ultimate effect on intrinsic motivation.* (Deci et al., 2001, p. 12).

Specifically, the controlling and informational aspects of the AEP program may have affected productivity as well. Overall, teachers had to meet a standard to obtain the reward, increasing the controlling aspect of AEP, and thus, reducing self-determination and intrinsic motivation. However, these negative effects may have been offset by the strong informational aspects of the program. In addition, the nature of the program (voluntary application, a standard-based evaluation, excellence standards to obtain the reward, and the rigorous process of objective evaluation of the portfolio and test submitted by the applicant) made it more likely that teachers perceived self-determination when obtaining the reward. As has been found, this net effect on intrinsic motivation, in turn, can have affected the level of effort made by the teacher after the reward was given (Bénabou and Tirole, 2003).

To better analyze the potential of effect of the program, each component of the program can be studied separately. The two components—financial and non-financial—worked based on different mechanisms and may have had different results. Regarding the financial component, as suggested by the empirical (Deci et al., 2001, Cerasoli et al., 2014), and theoretical evidence (Lazear, 1986, Bénabou and Tirole, 2003), it may have induced a crowding out process of the intrinsic motivation. From psychology, it is known that the positive effect of a short-term raise of extrinsic motivation by rewards (or punishments) might prove costly due to the possibility of a "crowding out" of self-confidence and intrinsic motivation in the long run, Koch et al. (2015). Therefore, the performance-contingent feature of the program and the fact that one component of the program is monetary, may have resulted in a reduction of intrinsic motivation, effort, and productivity after the program was given. Adding the public recognition component to the program could reinforce the information aspect of the reward, which increased the perception of

self-determined competence. In consequence, AEP, with its features, could also increase the perception of self-determination that the reward brings, thus increasing the chance of having a positive net effect of the program on teachers' effort and effectiveness. Consequently, the total effect of the program depended on which effect took precedence over the other on the intrinsic motivation, effort and productivity.

4.2. The Motivational Model

To further understand the relationship between an extrinsic non-financial reward, as one component of the AEP program, and the certified teachers' self-confidence and effort, a theoretical framework, based on the paper "*Self-Confidence and Personal Motivation*" written by Bénabou and Tirole in 2002, is described in this section.

Bénabou and Tirole (2002) developed a framework, the Motivational Model (MM), that unifies themes from psychology literature, and brings to light some of their economic implications. This MM can be applied to teachers in school settings to understand how they build their self-esteem, and how the AEP non-financial reward could have helped protect it and keep their intrinsic motivation at the highest possible level.

Let's consider a teacher who has imperfect knowledge of his abilities, or more generally, of the eventual costs and payoffs of his actions. He every day determines an optimal endogenous value of self-confidence that accounts for both "can-do" optimism and "defensive" pessimism when faced with situations and incentives. This self-esteem determination process responds to an optimization of the benefits obtained from preserving his effort and motivation against the risk of becoming overconfident.

The solution of this optimization process takes into consideration that the teacher has a demand and supply of self-confidence. The self-esteem-demand is based on the fact that he, like any other person, may prefer positive self-views to accurate views, which enhances the individual's motivation. This preference comes from the fact that ability and effort interact in determining performance, in most instances, they are complements, so that higher self-confidence enhances the motivation to act. In consequence, anyone with a vested interest in his performance has an incentive to build up, maintain and, ultimately, demand self-esteem.

The supply side of the self-esteem problem is based on the power of self-deception to achieve a positive self-assessment. Most often, the relevant issue is how the teacher deals with the good and especially the bad news concerning his performances and abilities. This is where the mechanisms of defensive denial, so prominently emphasized in psychology, come into play.

The interaction between self-confidence demand and supply in the presence of new information is modeled with a game-theoretic model of endogenously selective memory. The basic idea is that the teacher can, within limits, affect the probability of remembering a given piece of data. This is the motivation part. On the other hand, the teacher is assumed as rational and realize that he has a selective memory. This is the cognition part. The resulting structure is that of a game of strategic communication between the teacher's temporal selves. In deciding whether to try to repress bad news, the he weighs the benefits from keeping his motivation level versus the cost associated to make decisions being overconfident. Later on, however, the teacher appropriately discounts the reliability of optimistic recollections and rationalizations. Solving the model, a multiple interpersonal equilibrium may arise, ranging from systematic denial to

complete self-honesty. In practice, the model solution is a set of perfect Bayesian equilibria, which depends on the teacher's degree of time inconsistency and memory repression costs.

Using the Bénabou and Tirole (2002) model, one can argue how the non-financial reward linked to AEP could have been capable of protecting or enhancing teacher's self-confidence and motivation to give his best. The symbolic AEP award might have worked as a fixed memory making the self-deception process harder, which may increase teacher self-confidence. Moreover, information that negatively affects teacher's self-image after the program was given, is more likely to be repressed to maintain a desired and optimal level of self-confidence. Consequently, the non-financial component of the reward might have been effective in increasing, or at least keeping, teachers' efforts even after the reward has been given. This has the potential to reduce the likelihood of observing the "fade-out" phenomena of the effect of financial incentives on teacher behavior.

The following is an excerpt and simplified adaptation of Bénabou and Tirole (2002)'s model to provide a theoretical framework for understanding the relationship among a non-financial reward, self-confidence, and effort. Below the model is described with details and with small adaptations as presented by the authors in Bénabou and Tirole (2002)'s pages 878 to 879 (demand side), 884 to 889 (supply side), 889 to 892 and 894 (equilibrium and solution).

The Bénabou and Tirole (2002) Motivational Model

The motivation problem. The demand side

Consider a risk-neutral individual with a relevant horizon of three periods: t = 0, 1, 2. At date 0, he selects an action that may affect both his flow payoff u_0 and his date-1 information structure. At date 1, he decides whether to undertake a task/project (exert effort, which has disutility cost *c*

> 0) or not (exert no effort). With some probability, which defines his ability, the project will succeed and yield a benefit V at date 2; failure generates no benefit. The individual's beliefs over θ (defining his self-confidence or self-esteem) are described by distribution functions $F(\theta)$ at date 0 and $F_1(\theta)$ at date 1. In the intervening period, new information may be received, or previous signals forgotten; focus here is on the first, more standard case, and turn to memory when modeling $\bar{\theta}_1 \equiv \int_0^1 \theta dF_1(\theta)$ will be a sufficient the supply side. Note that with risk-neutrality the mean statistic for F_1. For brevity, the model also refers to it as the agent's date-1 self-confidence.

Now, assume that the individual's preferences exhibit time inconsistency, due to quasi-hyperbolic discounting. There is indeed considerable experimental and everyday evidence that intertemporal choices exhibit a 'salience of the present', in the sense that discount rates are much lower at short horizons than at more distant ones. Denoting u_t and $E_t(\bullet)$ the flow payoffs and expectations at $t = 0, 1, 2$, the intertemporal utility perceived by the individual as of date 1 is:[5]

$$u_1 + \beta\delta E_1 [u_2] = -c + \beta\delta \bar{\theta}_1 V \quad (4.1)$$

when he undertakes the activity, and 0 when he does not. By contrast, intertemporal utility conditional on the same information set at date 1, but evaluated from the point of view of date 0 is:[6]

$$u_0 + \beta E_0 [\delta u_1 + \delta^2 u_2 | \bar{\theta}_1] = u_0 + \beta\delta [-c + \delta \bar{\theta}_1 V] \quad (4.2)$$

[5] Equation 1, page 879, in Bénabou and Tirole (2002)

[6] Equation 2, page 889, in Bénabou and Tirole (2002)

if the activity is undertaken at date 1, and u_0 otherwise. Whereas is a standard discount factor, reflects the momentary salience of the present. When β<1 the individual at date 0 ('Self 0') is concerned about his date 1 ('Self 1's') excessive preference for the present, or lack of willpower, which leads to the under provision of effort (procrastination). Indeed, Self 1 only exerts effort in the events where $\bar{\theta}_1 > c/\beta\delta V$, whereas, from the point of view of Self 0, it should be undertaken whenever $\bar{\theta}_1 > c/\delta V$. Note that while the model focuses here on the case where the individual's intrinsic ability is unknown, it could equally be the expected payoff in case of success *V*, the 'survival' probability δ, or the task's difficulty, measured by the cost of effort c. All that matters for the theory is that the individual be uncertain of the long-term return to effort $\theta\delta V/c$.

The supply side of the self-esteem problem

The model now turns to the supply side of the self-esteem problem. Given that a positive self-assessment may be desirable, the question is what are the means through which it can be achieved, or at least pursued? The answer would be self-deception. The model shall capture this process with an intertemporal setting to reconcile the motivation and cognition aspects of self-deception within a standard information-theoretic framework. The motivation part says that, under time-inconsistency, there is an incentive to try to recall signals that help sustain long-term goals and forget those that undermine them. The individual can, within limits, affect the probability of remembering a given piece of data. On the other hand, the cognition part says that people know that they have a selective memory. This means that the model maintains the rational inference process.

Assumption 1 (memory or awareness management)

The individual can, at a cost, increase or decrease the probability of remembering an event or its interpretation. Formally, let $\lambda \in [0,1)$ denote the probability that a piece of information received at date 0 will be recalled at date 1. The model defines the natural rate of recall $\lambda_N \in [0,1)$ as that which maximizes the date 0 flow payoff u_0. Increasing or decreasing λ thus involves a 'memory cost' $M(\lambda)$, i.e., a reduction in u_0, with $M(\lambda_N) = 0$, $M'(\lambda) \leq 0$ for $\lambda < \lambda_N$ and $M(\lambda) \geq 0$ for $\lambda > \lambda_N$.

Assumption 2 (metacognition)

While the individual can manipulate his conscious self-knowledge, he is aware that incentives exist that result in selective memory. If a person has a systematic tendency to forget, distort, or repress certain types of information he will likely become aware of it, and not blindly take at face value what comes to his mind when thinking about his past performances and the feedback he received. Instead, using (some) rational inference, he will realize that what he may have forgotten are non-random events. Formally, this introspection or skepticism with respect to the reliability of one's own self-knowledge is represented by Bayes' rule, which implies that a person cannot consistently fool himself in the same direction. Less sophisticated inference processes lead to similar results, so long as they are not excessively naive.

The game of self-deception

Let the agent receive, at date 0, a signal σ about his ability θ. To make things simple, let σ take only two values: with probability 1-q the agent receives bad news, $\sigma = L$, and with probability q

he receives no news at all, σ= ∅. In other words, 'no news is good news'. Let[7]

$$\theta_L \equiv E[\theta|\sigma = L] < E[\theta|\sigma = \emptyset] \equiv \theta_H \quad (4.3)$$

Since σ is informative about the return to date-1 effort, the agent's Self 1 would benefit from having this signal. If it is ego-threatening, however, Self 0 may have an interest in suppressing it. The recollection at date 1 of the news will be denoted $\hat{\sigma} \in \emptyset$, L. The model assumes that memories can be lost but not manufactured ex nihilo, so σ= ∅ always leads to $\hat{\sigma} = \emptyset$. A signal σ= L, on the other hand, may be forgotten due to natural memory decay or voluntary repression. Let λ denote the probability that bad news will be remembered accurately:[8]

$$\lambda \equiv Pr[\hat{\sigma} = L|\sigma = L] \quad (4.4)$$

As explained earlier, the agent can increase or decrease this probability with respect to its 'natural' value $\lambda_N \leq 1$; choosing a recall probability involves a 'memory cost' $M(\lambda)$. The model now analyzes the equilibrium in several stages.

The equilibrium

1. Inference problem of Self 1

Faced with a memory $\hat{\sigma} \in$ L, ∅, Self 1 must first assess its credibility. Given that memories cannot be invented, unfavorable ones are always credible. When Self 1 does not recall any adverse signals, on the other hand, he must ask himself whether there was indeed no bad news at date 0, or whether it may have been lost or censored. If Self 1 thinks that bad news is recalled

[7] Equation 9, page 889, in Bénabou and Tirole (2002).

[8] Equation 10, page 889, in Bénabou and Tirole (2002).

with probability λ^*, he uses Bayes' rule to compute the reliability of a "no recollection" message as,[9]

$$r^* \equiv \Pr[\sigma = L|\hat{\sigma} = \emptyset; \lambda^*] = \frac{q}{q+(1-q)(1-\lambda^*)} \quad (4.5)$$

His degree of self-confidence is then[10]

$$\theta(r^*) \equiv r^* \theta_H + (1 - r^*) \theta_L \quad (4.6)$$

2. Decisions and payoffs

The model normalizes the payoff in case of success to $V = 1$ and assume that the cost of date 1 effort is drawn from an interval $[c, \bar{c})$, with probability distribution *(c)* and density *φ(c) > 0*. The model assumes that $c > \beta\delta\theta_H > \beta\delta\theta_L > c$, which means that at date 1 there is always a positive probability of no effort, and a positive probability of effort. Given a signal σ at date 0 and a memory $\hat{\sigma}$ at date 1, Selves 0 and 1 respectively assess the productivity of date 1 effort as E[θ|σ) and $E[\theta|\hat{\sigma})$. Self 1 only works when the realization of the effort cost is c<$\beta\delta E[\theta|\hat{\sigma})$. So Self 0's payoff is:[11]

$$\beta\delta \int_0^{\beta\delta E[\theta|\hat{\sigma})} (\delta E[\theta|\sigma) - c) d\Phi(c) \quad (4.7)$$

3. Costs and benefits of selective memory or attention

Focusing on the 'bad news' case, denote as $U_C(\theta_L|r^*)$ is the expected utility of Self 0 (gross of memory-management costs) when the adverse information is successfully forgotten, and as $U_T(\theta_L)$ the corresponding value when it is accurately recalled. The subscripts *C* and *T* stand for

[9] Equation 11, page 889, in Bénabou and Tirole (2002).

[10] Equation 12, page 889, in Bénabou and Tirole (2002).

[11] Equation 13, page 890, in Bénabou and Tirole (2002).

'censored' and 'truth' respectively. Hiding from Self 1 the signal σ = *L* raises his self-confidence from θ_L to (*r*), leading him to exert effort in the additional states of the world where $_L < c <$ (*r*). As with ex ante ignorance, this has both costs and benefits; thus, if *r* is high enough that (*r*) > $_L$, the net gain or loss from self-deception is:[12]

$$U_C(\theta_L|r^*) - U_T(\theta_L) = \beta\delta \int_{\beta\delta\theta_L}^{\delta\theta_L} (\delta\theta_L - c)d\Phi(c) - \int_{\delta\theta_L}^{\beta\delta\theta(r^*)} (c - \delta\theta_L)d\Phi(c) \quad (4.8)$$

4. Strategic memory or awareness management

Faced with a signal σ = L that is hurtful to his self-esteem, Self 0 chooses the recall with probability λ so as to solve:[13]

$$max_\lambda \lambda U_T(\theta_L) + (1-\lambda)U_C(\theta_L|r^*) - M(\lambda) \quad (4.9)$$

Given the convexity of $M(\lambda)$, the optimum is uniquely determined (given r^*) by the first-order condition, which involves comparing the marginal benefit from self-deception, $U_C(\theta_L|r^*) - U_T(\theta_L)$, with the marginal cost, $M(\lambda)$. Finally, the Bayesian rationality of Self 1 means that he is aware of Self 0's choosing the recall strategy opportunistically according to Equation 4.9 and uses this optimal λ in his assessment of the reliability of memories (or lack thereof). A Perfect Bayesian Equilibrium of the memory game is a pair $(\lambda^*, r^*) \in [0,1)\ x\ [q,1)$ solves Equations 4.5 and 4.9.

5. The solution of the model

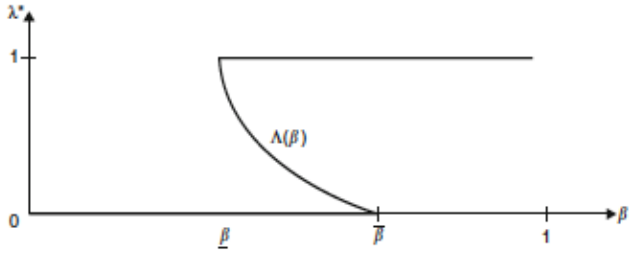
First assume that the manipulation of memory is costless, $M \equiv 0$. When $M \equiv 0$, there exist $\underline{\beta}$

[12] Equation 14, page 891, in Bénabou and Tirole (2002).

[13] Equation 15, page 891, in Bénabou and Tirole (2002).

arid $\overline{\beta}$ in (0,1), with the following properties. For low degrees of time inconsistency, $\beta > \overline{\beta}$, the unique equilibrium involves minimum repression ($\lambda^* = 1$); for high degrees, $\beta < \underline{\beta}$, it involves maximum repression ($\lambda^* = 0$). For intermediate degrees of time inconsistency, $\beta \in [\beta, \overline{\beta}]$, as shown in Figure 4.1, there are three equilibria, including a partially repressive one: $\lambda^* \in 0, \Lambda(\beta)$ decreases from 1 to 0 as β rises from $\underline{\beta}$ to $\overline{\beta}$.

Figure 4.1. Model solution for momentary salience of present (β) and recall rate (λ).



Note: Figure extracted from Bénabou and Tirole (2002), page 893.

Now assume that there is a cost of memory or awareness management. The memory cost function is:[14]

$$M(\lambda) = a + (1 - ln\,\lambda) + b(1 - ln(1 - \lambda)) \quad (4.10)$$

with $a > 0$ and $b \geq 0$. It minimized at the 'natural' recall rate $\lambda_N = a/(a+b)$ N = a/(a+b) and precludes complete repression. When b > 0 perfect recall is also prohibitively costly, and M is U-shaped. As to the distribution of effort costs, the model takes it to be uniform, $\varphi(c) = 1/\overline{c}$ on $[0; \overline{c}\,)$, with $\overline{c} > \beta\delta\theta_H$. In this case, for any (*a, b*) there are again either one or three equilibria. One can go further, and obtain explicit comparative statics results, by focusing on the

[14] Equation 17, page 894, in Bénabou and Tirole (2002)

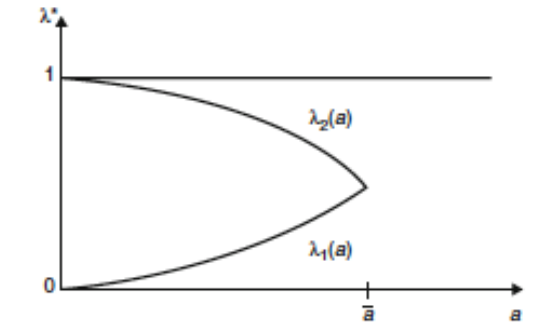
simpler case where recall is costless, but repression is costly. The graphic solution can be seen the following figures. Formally, for any given β there exist thresholds $\underline{a}$ and $\overline{a}$ with $0 \leq \underline{a} \leq \overline{a}$, and continuous functions $\lambda_1(a),\ \lambda_2(a),$ respectively increasing and decreasing in a, such that: (i) for $a\ \in (0, \underline{a})$, the unique equilibrium corresponds to $\lambda^* = \lambda_1(a)$; (ii) for $a\ \in (\underline{a}, \overline{a})$, there are three equilibria: $\lambda^* \in \lambda_1(a), \lambda_2(a), 1$; (iii) for $a\ \in (\overline{a}, \infty)$, the unique equilibrium corresponds to λ^*= 1. Importantly, note that small changes in awareness cost can induce large changes in self-esteem and behavior.

6. Increasing effort

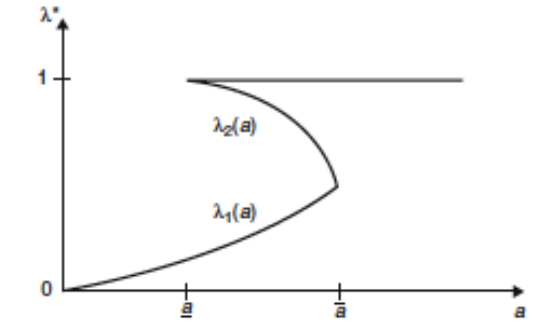
Having reviewed the MM formally, I can state also formally how the non-financial incentive as part of AEP may result in higher levels of effort. Assume that AEP can alter the awareness/repression' technology $M(\lambda)$ in a way that increases the cost of manipulating memory. Thus, AEP increases the first part of the cost function (*a* in Equation 4.10) because AEP gives a merit-reward to teachers, equipping them with pin that can always wear while teaching. To understand the consequences of this technological change the time consistency level of the teacher and the initial parameter cost *a* need to I need to be considered. Taking the simpler case that has a unique equilibrium, where $\beta < \beta_1$, where the teacher has a high degree of time inconsistency—teacher lack of willpower, which might lead to under-provision of effort—and the cost parameter is in a low range, the higher a, the more likely = 1, which means the minimum repression. The probability that bad news will be remembered accurately approximates one. With that *r* = 1, which means that Self 1 considers the memory process completely reliable. Thus, $\theta(r^*) = \theta_H$, the self-confidence of the teacher is the highest possible. Consequently, non-financial incentives might be effective increasing teachers' effort. Nonetheless, for different

levels of cost parameters and different levels of time consistency, multiple equilibriums arise. Thus, the effectiveness of non-financial incentives on effort is ultimately an empirical question.

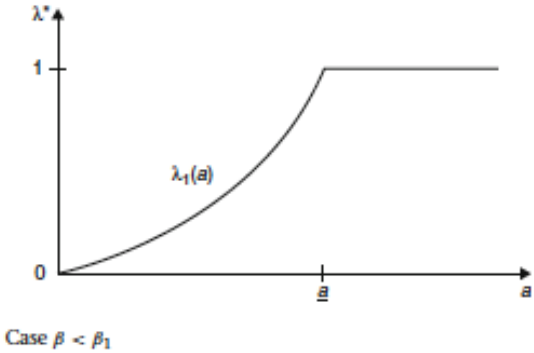
Figure 4.2. Model Solution for the parameter cost (*a*) and the recall rate (λ).



Case $\beta_2 < \beta < \beta_3$



Case $\beta_1 < \beta < \beta_2$



Case $\beta < \beta_1$

Note: Figures extracted from Bénabou and Tirole (2002), pages 895 and 896.

4.3. Final remarks

Since the certification was tied to bonuses (the financial component of the reward) and public recognition (the non-financial component of the reward), the program has the potential of improving school performance. The financial component could positively affect teacher's decision on the optimal level of effort as a principal-agent model states; however, it could crow-out intrinsic motivation decreasing it. Complementarily, based on Deci et al. (2001), the CET suggests that adding a public recognition component to the AEP program could reinforce the information aspect of the reward, which increases the perception of self-determined competence. In same sense, based on Bénabou and Tirole (2003)'s MM, the recognition of teacher's quality may have fostered teacher's self-confidence and motivation and thereby increasing their effort and productivity (student outcomes) in a permanent fashion. Taking all together, the net ex-post effect of the program on teachers' effort and effectiveness could be positive.

At the same time, the MM argues that multiple equilibria may exist, while CET states that the performance-contingent feature of the program and the fact that one component of the program is monetary, may result in a reduction of intrinsic motivation, effort, and productivity.

Consequently, the total ex-post effect of the program depends on which effect takes precedence over the other on self-confidence or intrinsic motivation, effort, and productivity. Thus, the effectiveness of financial and non-financial rewards on effort is ultimately an empirical question.

Chapter 5 - Program effect and research questions

As discussed previously, the rewards given by the Teaching Excellence Program[15] (AEP for its initials in Spanish) to certified teachers had the potential of increasing their productivity. Taking the previous literature into account, the research on teachers' incentives, the program's objectives, and the findings made by qualitative research focused on the AEP, the program's potential consequences on teachers' effects can be divided into four groups.

1. **Full program effect.** The program aimed to increase the quality of the educational system. Even though approximating and measuring "quality" is part of a more extensive discussion as explained in Chapter 9, a common practice has been to define it as the observed student learning outcomes. Thus, the direct effect of the full program can be measured by ex-post students' academic achievement (SIMCE for its initials in Spanish) within three years after the teacher's application to the program, which is the year that the teacher receives certification and the two following. The full program effect including financial and non-financial components is estimated for teachers who applied to AEP from 2003-2006 and 2009-2011. This means those years, when the non-financial reward was not given, are not included in the estimation of the full program effect.
2. **"Fade-out" of the full program effect.** The program's effect may have faded out which would be consistent with the literature on teacher incentives (Springer et al., 2011, Glewwe et al., 2010) and with recent studies that have found that the impact of being

[15] Asignación de Excelencia Pedagógica.

assigned to a more effective teacher declines by half or more between end-of-year test scores and ones two years later (Jackson et al., 2014). Thus, the full program effect is estimated for the second and third year after the teacher applied to the program.

3. **Unbundled effect.** The public recognition component (ceremony and pin) and the financial component (annual bonus) could independently explain ex-post outcomes. We have seen that from a theoretical perspective, the effects of the financial and public recognition components might even cancel each other out. On the one hand, as explained above, the latter may have improved the self-confidence or the perception of competence, which has the potential to increase teachers' intrinsic motivation, and thus increase effort and improve teacher's performance (Bénabou and Tirole, 2002). Thus, frequently wearing the pin may have decreased the cost of increasing self-confidence, which makes teachers believe their efforts will be productive. This, in turn, enhanced their motivation and effort to teach. On the other hand, we have the effect of giving an unconditional bonus to certified teachers every year. The bonus works as an extrinsic motivator, which may have had a limited impact on current performance and may have reduced the teacher's intrinsic motivation to undertake similar tasks in the future. Consequently, the net full program's effect on teachers' effort and performance after the certification was given is not unequivocal. To understand the direction and magnitude of the effects of each component, the bonus and pin effect are unbundled, taking advantage of a difference in the program implementation between years. In 2007 and 2008, AEP only included the financial incentive; thus, estimating the program's effect for those years is equivalent to isolating the effect of the financial component. Importantly, the pin's effect is never

observed in isolation, only the effect of the non-financial component on the top of the financial one is under analysis. Therefore, the comparison between full program effect and the unbundled effect sheds some light on the effect of the non-financial component complementing a financial incentive.

4. **Underlying mechanism.** Following the theories of Bénabou and Tirole (2002), I explore the relationship among the program, teachers' self-confidence, and efforts. This is to provide evidence on the underlying mechanisms of the program's direct effects on student learning after the certification was given. Specifically, the effort levels and teacher self-confidence are studied as intermediate results. These results are obtained from the survey answered by the teachers associated with SIMCE. To study this underlying mechanism, a *ceteris paribus* assumption for the Motivational Model (MM) is also tested. This means testing the existence of an alternative mechanism by which the program may have resulted in significant sorting of students and schools to certified teachers. If the program affected the matching process between AEP teachers and students or schools after the program was given, the program's effect on outcomes such as test scores could be explained by this fact at least in part, limiting the chances of testing the MM.

Based on the theoretical framework previously described and for the sake of simplicity, the analysis of unbundled effects assumes that the full program's effect can be represented by a function with perfect substitution and no complementarities between its two inputs/components. This means that the program's effect function is linear and additive on the program's components.

Importantly, the estimated program effect corresponds to the change in outcomes after the reward is given, which makes a difference compared with most of the previous research on incentives that estimates their impact on outcomes before the incentive is given.

Summing up, while exploring the four groups of effects, there are five ***hypotheses*** that are tested:

***Question** #1*, Does the AEP affect ex-post teachers' performance? (Full program effect)

***Question** #2,* Does the ex-post effect of the AEP fade-out? (Fade-out of the full program effect).

***Question** #3,* Does each component of the AEP affect ex-post teachers' performance? (Unbundled program effect)

***Question** #4,* Does the AEP affect ex-post teachers' behavior? (Underlying mechanism of the full program)

***Question** #4.1,* Does the AEP affect ex-post teachers' self-confidence?

***Question** #4.2,* Does the AEP affect ex-post teachers' effort?

***Question** #5,* Does each component of the AEP affect ex-post teachers' behavior? (Underlying mechanism of each program component)

***Question** #5.1,* Does each component of the AEP affect ex-post teachers' self-confidence?

***Question** #5.2,* Does each component of the AEP affect ex-post teachers' effort?

In Chapter 8, the empirical strategy developed to answer these research questions is presented. In the following Chapters 6 and 9, the data, the variables, and measures used to implement this strategy are presented and discussed regarding their interpretation.

Chapter 6 - Data

Before presenting the theoretical and empirical approach used to estimate and understand the program's effect, this chapter presents the data available for implementing the empirical strategy. It specifically explains how the data is accessed and merged into the main dataset, and how the final number of observations used for the estimations is obtained. Complementarily, Chapter 9 explains the operationalization of the data to define and measure the outcomes of interest and covariates, while Appendix E describes the specific datasets that make up the main dataset used.

Overall, the main dataset includes math and language teachers' evaluation scores on a scale from 1 to 4, the corresponding standardized test score (SIMCE for its initials in Spanish), and the SIMCE survey data for teachers and students within the three years following the program application. Characteristics of schools and students, and information about motivation and self-perception that teachers self-report are also in the dataset. The data is available for application year (t), the year when the incentives are given (t + 1) (April), and the year when the students take the SIMCE (November) at some point (t + l) within three years (l = 1, 2, 3). The timing of data is anchored to the application year t. This means that data for year t + l is data for l years after the application. The final sample of teachers that is used to estimate the program's effect includes those who: were applying for certification in the math or language categories between 2003 and 2011, were teaching math or language when their students took the math or reading SIMCE respectively, were applying for the first time, and who were assessed as non-cheaters. Teachers who applied in 2002 are excluded from the sample because their scores in the dataset only had one decimal instead of two, as the scores normally do. Lastly, the dataset used to estimate the specification checks in year t, as described in Section 8.2, is the same group of

teachers, except those teachers who applied in 2003 have been excluded. This is done because there was no information available on where they taught, making it impossible to merge their application data with their teaching data to check the validity of the regression discontinuity design.

6.1. Merging the data

Specifically, the main dataset is merged by linking six databases:

1. AEP DATASET (database of AEP applicants that identifies the applicants and their scores),
2. STCS DATASET (database of teachers working in publicly-funded schools identifying the school where they work, the classroom they teach in, and the subject that they teach),
3. SIMCE DATASET (database of students and their scores on the SIMCE test),
4. SIMCE TEACHER SURVEY (database of teacher's answers to the SIMCE survey, which is used to obtain information about teachers' characteristics, teaching attitudes, and expectations),
5. TEACHER CENSUS DATASET (database of administrative data on teachers that includes teachers' characteristics, experience, position within the school, and so on), and
6. SCHOOL SIMCE DATASET (database of schools' administrative data).

The Ministry of Education in Chile (MINEDUC), through the Statistics Department and the Center of Teaching Training and Improvement (CPEIP in Spanish), systematically collects the above datasets. Some are publicly available, and others need a special request to access and merge. The datasets are merged by school, teacher, and student id respectively. Appendix E describes content of the datasets and keys used in the merge.

Creating the consolidated dataset involves several steps. The first step is to merge the AEP DATASET with the STCS DATASET in order to identify those teachers teaching language or math within three years after their application. For this group of teachers, I also identify the school and classroom where they teach. Panel A, columns 4 and 8, of Table F.1 shows that 28%-27% of the teachers were teaching language or math in the same year that they applied to the Teaching Excellence Award[16] (AEP for its initials in Spanish), correspondingly. Table F.1 shows the number of applicants that were teaching math or language one, two and three years after the application, respectively Panels B, C, D. Since the STCS DATASET is available from year 2004 onward, I cannot identify where 2003 applicants taught when they applied to the AEP.

Now that I have identified where the applicant teachers taught for years t, $t + 1$, $t + 2$ and $t + 3$, as a second step, I identify the teachers who applied to be certified in math or language. Columns 6 and 10 of Table F.1 show the percentage of teachers who applied for language or math certification and teach language or math respectively. The percentage is calculated from the number of teachers that teach language or math. In general, a high percentage (between 79% and 87%) of teachers that teach a given subject apply for certification in that subject (Panel A of Table F.1).

Then, as a third step, I merge the teacher data with the SIMCE DATASET in order to add the math and reading SIMCE scores of their students in case that they took the test in years t, t + 1, t + 2 and t + 3. Column 4 of Panel A in Tables 6.1 and 6.2 show that 27%/29% of applicants

[16] Asignación de Excelencia Pedagógica.

teaching language/math and applying for matching certification in year t can be linked to a reading/math SIMCE Score in t + l. There is an important fact to consider in Tables F.1, 6.2 and 6.4. The applicants in the year 2006 tend to have fewer observations; this occurs because of a problem with the STCS DATASET for that year. The MINEDUC recognizes that there might be a problem with the teachers' ids and matching them with other datasets in 2006.

Tables 6.1 and 6.2 show the results of the merging process, which is the number of students of the subgroup of teachers identified previously in the third step. From Table F.1, we can see that, 10,621 teachers were teaching language in t+1, t+2 or t+3 and applied for a language certification in t (see total in the tenth column). From them 3,144 teachers (see total in the third column of Panel E of Table 6.1) can be linked to student's test scores within three years after teachers' application. There were 2,616 teachers (first-takers and non-cheaters) who applied for language certification, taught language and had reading test scores.

This merged dataset is combined with the SIMCE TEACHER SURVEY and the TEACHER CENSUS DATASET to obtain information about teachers' characteristics, teaching attitudes, experience, position within the school, and students' attitudes towards learning, among other data. Lastly, the merged dataset is complemented with school data by merging with SCHOOL SIMCE DATASET. This supplies school characteristics, such as type of administration.

Having done the merge process for each subject separately, the last step is to combine the math and language datasets, in order to have one single dataset for math and language teachers.

It is important to mention again the fact that the dataset includes data for one, two, and three years after the teachers apply to the AEP. In practice, within the datasets each observation is stacked with three years following the application year. Stacking the datasets allows me to

evaluate the effect of the program on current and future outcomes by estimating Equation 8.6 for each value of *l* separately to allow a more flexible specification. Using data for the three years following the application year enables me to increase the sample size, and, with that, the power of the estimations.

6.2. Final sample

The result of this merging process is a dataset in which a sample of applicants are linked to their AEP score, the classrooms in which they teach, their students' test scores, the parents' report, and their own answers to the SIMCE teacher survey. This data for each applicant is available for a window of three years after the application.

However, the definition of the final sample needs to take three issues into account. First, the evaluation process carried out by the government was aimed at finding applicants who cheat using material produced by other applicants. These teachers did not receive certification in $t + 1$ because evidence of copying or plagiarism was found by the AEP. Therefore, it is reasonable to keep only non-cheating applicants in the sample since the rule used for treatment assignment does not apply to the cheaters. The exclusion of this group of applicants is unlikely to bias the global estimates because they are 26 cases and represent 0.8%/0.9% of the total number of teachers with math or reading test scores in the data set (see Tables 6.1 and 6.2).

Table 6.1. SIMCE and math teachers in t, t + 1, t + 2, and t + 3.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Teachers applying and teaching math and having math test scores		Teachers cheating in t+1		Re-takers		Teachers (first-takers and non-cheaters) applying to math certification, teaching math and having math test scores	
Application Year t	Applicants #	Teaching and Test Scores								
Panel A. In t	N	Year t	N	%	N	%	N	%	N	%
2002	1,906	2002								
2003	935	2003								
2004	1,621	2004	129	17	0	0.0	12	9	117	16
2005	1,834	2005	113	20	1	0.9	22	19	90	16
2006	2,215	2006	119	29	0	0.0	27	23	92	23
2007	1,666	2007	197	38	2	1.0	41	21	154	30
2008	1,661	2008	133	30	0	0.0	13	10	120	27
2009	1,815	2009	165	38	4	2.4	22	13	139	32
2010	1,499	2010	134	33	4	3.0	16	12	114	28
2011	1,316	2011	160	42	4	2.5	29	18	127	33
TOTAL	16,468	TOTAL	1,150	29	15	1.3	182	16	953	24
Panel B. In t+1	N	Year t+1	N	%	N	%	N	%	N	%
2002	1,906	2003								
2003	935	2004	98	18	0	0.0	2	2	96	18
2004	1,621	2005	148	21	3	2.0	25	17	120	17
2005	1,834	2006	146	28	1	0.7	25	17	120	23
2006	2,215	2007	128	33	0	0.0	24	19	104	27
2007	1,666	2008	146	30	0	0.0	25	17	121	25
2008	1,661	2009	159	40	2	1.3	19	12	138	34
2009	1,815	2010	124	31	1	0.8	12	10	111	28
2010	1,499	2011	157	41	3	1.9	20	13	134	35
2011	1,316	2012	124	35	3	2.4	21	17	100	28
TOTAL	16,468	TOTAL	1,230	30	13	1.1	173	14	1,044	25
Panel C. In t+2	N	Year t+2	N	%	N	%	N	%	N	%
2002	1,906	2004								
2003	935	2005	103	20	0	0.0	25	24	78	16
2004	1,621	2006	195	30	1	0.5	20	10	174	27
2005	1,834	2007	191	40	2	1.0	30	16	159	34
2006	2,215	2008	98	30	0	0.0	20	20	78	23
2007	1,666	2009	172	40	1	0.6	31	18	140	32
2008	1,661	2010	118	31	0	0.0	16	14	102	27
2009	1,815	2011	158	41	1	0.6	24	15	133	35
2010	1,499	2012	126	35	3	2.4	21	17	102	28
2011	1,316	2013								
TOTAL	16,468	TOTAL	1,161	33	8	0.7	187	16	966	27
Panel D. In t+3	N	Year t+3	N	%	N	%	N	%	N	%
2002	1,906	2005								
2003	935	2006	106	22	0	0.0	17	16	89	19
2004	1,621	2007	217	36	1	0.5	34	16	182	30
2005	1,834	2008	132	31	1	0.8	19	14	112	26
2006	2,215	2009	88	29	0	0.0	21	24	67	22
2007	1,666	2010	117	29	2	1.7	19	16	96	23
2008	1,661	2011	154	43	1	0.6	17	11	136	38
2009	1,815	2012	110	31	0	0.0	18	16	92	26
2010	1,499	2013								
2011	1,316	2014								
TOTAL	16,468	TOTAL	924	31	5	0.5	145	16	774	26
Panel E. In $t + l$	N	Year t+1/t+2/t+3	N	%	N	%	N	%	N	%
2002	1,906	2002/2003/2004								
2003	935	2003/2004/2005	307	33	0	0.0	44	14	263	17
2004	1,621	2004/2005/2006	560	35	5	0.9	79	14	476	24
2005	1,834	2005/2006/2007	469	26	4	0.9	74	16	391	27
2006	2,215	2006/2006/2008	314	14	0	0.0	65	21	249	24
2007	1,666	2007/2008/2009	435	26	3	0.7	75	17	357	27
2008	1,661	2008/2009/2010	431	26	3	0.7	52	12	376	33
2009	1,815	2009/2010/2011	392	22	2	0.5	54	14	336	29
2010	1,499	2010/2011/2012	283	19	6	2.1	41	14	236	32
2011	1,316	2011/2012/2013	124	9	3	2.4	21	17	100	28
TOTAL	16,468	TOTAL	3,315	20	26	0.8	505	15	2,784	26

Table 6.2. SIMCE and language teachers in t, t + 1, t + 2, and t + 3.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Teachers applying and teaching language and having reading test scores		Teachers cheating in t+1		Re-takers		Teachers (first-takers and non-cheaters) applying to language and teaching language and having reading test scores	
Year t	Applicants #	Teaching and Test Scores								
Panel A. In t	N	Year t	N	%	N	%	N	%	N	%
2002	1,906	2002								
2003	935	2003								
2004	1,621	2004	125	15	0	0.0	6	5	119	15
2005	1,834	2005	101	17	1	1.0	19	19	81	14
2006	2,215	2006	2	7	0	0.0	0	0	2	7
2007	1,666	2007	186	35	4	2.2	37	20	145	27
2008	1,661	2008	154	32	0	0.0	10	6	144	30
2009	1,815	2009	162	36	2	1.2	17	10	143	31
2010	1,499	2010	137	33	2	1.5	20	15	115	28
2011	1,316	2011	124	35	5	4.0	19	15	100	28
TOTAL	16,468	TOTAL	991	27	14	1.4	128	13	849	23
Panel B. In t+1	N	Year t+1	N	%	N	%	N	%	N	%
2002	1,906	2003								
2003	935	2004	96	17	0	0.0	4	4	92	17
2004	1,621	2005	143	19	3	2.1	22	15	118	16
2005	1,834	2006	157	29	0	0.0	33	21	124	23
2006	2,215	2007	13	36	0	0.0	4	31	9	25
2007	1,666	2008	133	28	0	0.0	18	14	115	24
2008	1,661	2009	143	33	2	1.4	23	16	118	27
2009	1,815	2010	135	32	1	0.7	21	16	113	27
2010	1,499	2011	153	40	2	1.3	18	12	133	35
2011	1,316	2012	154	48	6	3.9	27	18	121	38
TOTAL	16,468	TOTAL	1,127	29	14	1.2	170	15	943	24
Panel C. In t+2	N	Year t+2	N	%	N	%	N	%	N	%
2002	1,906	2004								
2003	935	2005	108	21	0	0.0	27	25	81	16
2004	1,621	2006	224	32	1	0.4	21	9	202	29
2005	1,834	2007	204	40	1	0.5	40	20	163	32
2006	2,215	2008	5	21	0	0.0	2	40	3	13
2007	1,666	2009	161	36	3	1.9	28	17	130	29
2008	1,661	2010	136	33	1	0.7	18	13	117	29
2009	1,815	2011	120	30	0	0.0	17	14	103	26
2010	1,499	2012	187	53	3	1.6	30	16	154	44
2011	1,316	2013								
TOTAL	16,468	TOTAL	1,145	34	9	0.8	183	16	953	28
Panel D. In t+3	N	Year t+3	N	%	N	%	N	%	N	%
2002	1,906	2005								
2003	935	2006	110	22	0	0.0	19	17	91	18
2004	1,621	2007	207	31	1	0.5	33	16	173	26
2005	1,834	2008	131	29	0	0.0	20	15	111	24
2006	2,215	2009	10	38	0	0.0	2	20	8	31
2007	1,666	2010	108	26	2	1.9	15	14	91	22
2008	1,661	2011	138	37	1	0.7	25	18	112	30
2009	1,815	2012	168	47	0	0.0	34	20	134	37
2010	1,499	2013								
2011	1,316	2014								
TOTAL	16,468	TOTAL	872	31	4	0.5	148	17	720	26
Panel E. In t+1	N	Year t+1, t+2, t+3	N	%	N	%	N	%	N	%
2002	1,906	2002/2003/2004								
2003	935	2003/2004/2005	314	20	0	0.0	50	16	264	17
2004	1,621	2004/2005/2006	574	27	5	0.9	76	13	493	23
2005	1,834	2005/2006/2007	492	33	1	0.2	93	19	398	26
2006	2,215	2006/2006/2008	28	33	0	0.0	8	29	20	23
2007	1,666	2007/2008/2009	402	30	5	1.2	61	15	336	25
2008	1,661	2008/2009/2010	417	34	4	1.0	66	16	347	29
2009	1,815	2009/2010/2011	423	36	1	0.2	72	17	350	30
2010	1,499	2010/2011/2012	340	46	5	1.5	48	14	287	39
2011	1,316	2011/2012/2013	154	48	6	3.9	27	18	121	38
TOTAL	16,468	TOTAL	3,144	31	27	0.9	501	16	2,616	26

Table 6.3. Number of teachers by application order.

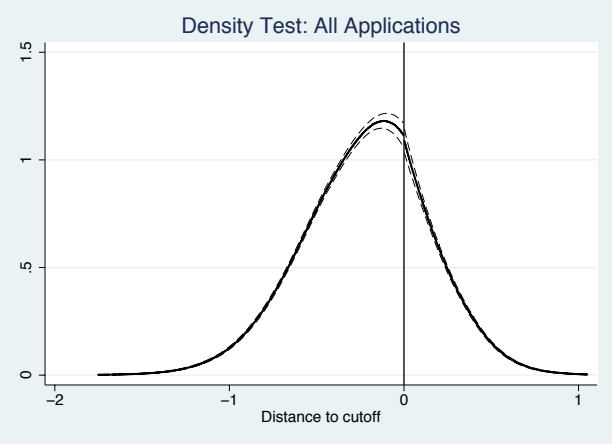
Number of application	Number of Teachers	%
1	14,703	89.3%
2	1,632	9.9%
3	124	0.8%
4	9	0.1%
Total	16,468	100%

The second issue has to do with the fact that the sample of teachers includes takers. As seen in Table 6.3, there are 16,468 applications, of which 14,703 are first applications (p=1). This means that 10.7% of the applicants are applying for a second (p=2), third (p=3), or fourth time (p=4).

Teachers could apply many times to the AEP, which may have resulted in a manipulation of the running variable. For instance, a teacher who knew that his first application was very close to reaching the cutoff could improve exactly what was needed in order to obtain a higher score. If this happened, the validity of the identification strategy would be threatened. Even though there is no evidence of manipulation, as shown in Figures 6.1, 6.2 and 6.3, the test is clearer for p=1. Therefore, it is reasonable to keep only first-time applicants in the sample. Doing this, as shown in Tables 6.1 and 6.2 results in losing 15%/16% of data for language and math teachers with SIMCE scores in the main data set.

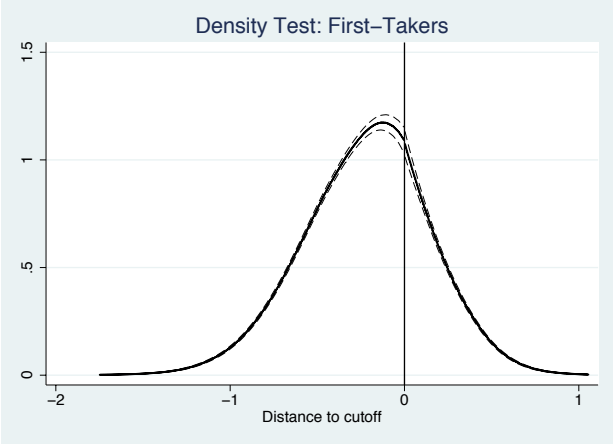
The third issue that needs attention in order to define the final sample of applicants is related to the completeness of the dataset made available by the Ministry of Education. As mentioned above, there is no data to identify the school and classrooms where the applicants taught in 2003. In consequence, teachers who applied to the program in 2003 are dropped from the sample used to run the specification checks in year t.

Figure 6.1. MacCrary test for any p.



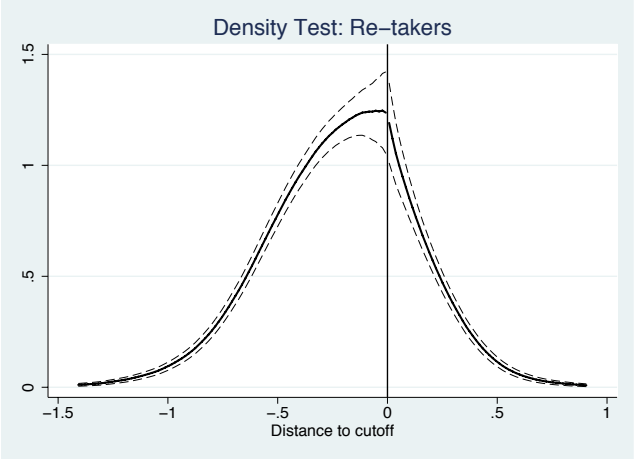
Notes: The figure shows the plot calculated by the DCdensity STATA command for the score distance to the program cutoff score. Only data for non-cheaters who applied in 2003-2011 is included. The coefficient estimated by the McCrary test is -0.014, with a standard deviation of 0.04, which gives us a t-statistic of -0.39.

Figure 6.2. MacCrary test for p=1.



***Notes*:** The figure shows the plot calculated by the DCdensity STATA command for the score distance to the program cutoff score. Only data for non-cheaters and first-applicants who applied in 2003-2011 is included. The coefficient estimated by the McCrary test is -0.007 with a standard deviation of 0.04, which implies a t-statistic of -0.175.

Figure 6.3. MacCrary test for p>1.



***Notes*:** The figure shows the plot calculated by the DCdensity STATA command for the score distance to the program cutoff score. Only data for non-cheaters who applied more than once in 2003-2011 is included. The coefficient estimated by the McCrary test is -0.013 with a standard deviation of 0.113, which implies a t-statistic of -0.119.

Consequently, the program's effect, as defined in the methodology section, is estimated using data for teachers applying for math or language certification in year t (2003-2011) who taught

math or language, and whose students took the math or reading SIMCE test in t + 1, t + 2 and t + 3. As shown in Table 6.4, 2,616 language teachers are in the final sample. They in turn can be linked to 93,393 students with reading test scores. Meanwhile, the final sample is comprised by 2,784 first-applicants and non-cheaters who applied for math certification in t, taught math in (t + l) and have math test scores in the same year (t + l). This group of teachers is linked to 100,545 students with math test scores. In fact, as shown in Table 6.4, the program's effect is estimated with data for teachers by using close to 200,000 observations at student level. For example, Table 6.4 shows the effect of program for is estimated with 193,938 student observations for 5,400 teachers. Lastly, the specification checks are run for those 1,802 teachers applying for math or language certification in year t (2004-2011) who teach math or language, and whose 64,548 students took the math or reading SIMCE test in t, as well (see Panel A of Table 6.4).

Table 6.4. Final sample of math and language teachers and students.

			(1) Final teachers' sample in t+1, t+2, and t+3	(2) Final students' sample in t+1, t+2, and t+3		(3) Final teachers' sample in t	(4) Final students' sample in t
Year t	N	Year t+1/t+2/t+3	N	N	Year t	N	N
2002	1,906	2003/2004/2005			2002		
2003	935	2004/2005/2006	527	20,586	2003		
2004	1,621	2005/2006/2007	969	37,290	2004	236	10,839
2005	1,834	2006/2006/2008	789	32,499	2005	171	5,135
2006	2,215	2007/2008/2009	269	7,923	2006	94	2,705
2007	1,666	2008/2009/2010	693	25,020	2007	299	11,299
2008	1,661	2009/2010/2011	723	26,041	2008	264	9,629
2009	1,815	2010/2011/2012	686	23,879	2009	282	9,379
2010	1,499	2011/2012/2013	523	14,266	2010	229	8,088
2011	1,316	2012/2013/2014	221	6,434	2011	227	7,474
TOTAL	16,468	TOTAL	5,400	193,938	TOTAL	1,802	64,548

6.3. Descriptive statistics of applicants and certified teachers.

The AEP program was open to teachers who comply with specific requirements. Teachers working in a teaching position for at least 20 hours a week, in a public-funded school, and who have at least two years of experience, could apply to the program. Thus, the minimum requirements to apply, together with the volunteer feature of the program, resulted in having a pool of applicants that is a subsample of teachers who do not necessarily represent Chilean teachers as a whole. This fact, in turn, may have significant consequences concerning the external validity of the results obtained in this research. In this case, attention should be paid towards the possibility of extrapolating the results of the study carried out on the analyzed teacher sample, as compared to the average teacher in the labor market.

To describe the teacher's sample, in this section, besides the description of the application process, two kinds of comparative descriptions are presented: applicants versus non-applicants and certified versus non-certified teachers. The first comparative exercise allows an understanding of what makes the applicants different relative to those teachers that did not participate in the AEP. The second exercise gives an idea of the global average differences between teachers that showed having the teaching skills and knowledge to obtain the certification and those who did not.

Application process characterization

Two facts can characterize the application process. The first one is the high level of self-selection along the application process, represented by a high drop-out rate between the registration stage and the portfolio submission six months later. As stated by an applicant, *"They knew that this was a*

difficult thing; that is, that it was made that way because it was a selective process, not everyone was going to reach the end" (Araya et al., 2011, p. 218, certified teacher quote).

Every year, thousands of teachers registered for the program, and high rate of them asked for the portfolio. However, only a small portion of them became valid applications, submitting the portfolio and taking the test to allow their assessment. At the registration stage, more than 65,446 registrations for the program were received. However, the percentage of teachers that completes the application process averaged 25% and ranged between 19% and 30% for the period between 2002 and 2011. Teachers reported that the main reasons behind their drop-out decision were an excessive work-load and limited time to work on the portfolio (Falck et al., 2015).

Table 6.5. Application process statistics.

Application Year	Total registered teachers	% of valid applicants
2002	8,638	22%
2003	4,147	23%
2004	5,392	30%
2005	6,927	26%
2006	7,473	30%
2007	6,941	24%
2008	6,298	26%
2009	7,314	25%
2010	7,696	19%
2011	4,620	28%
Total	65,446	25%

The second fact that characterized the application process is the small share of the total number of potential applicants of teachers who finished the application process (valid applicants). Until 2011, more than 16,000 Chilean teachers applied voluntarily to the program.

Even though the program was free and confidential, for the period of 2002-2011, on average, only 12% of eligible teachers applied to the program.[17]

It is also relevant to study if the teachers that finally applied to the program come from a diverse set of schools. This can be informative and demonstrates how special or representative the applicants are. In Table 6.6, we observe that the number of schools with at least one teacher applying for AEP every year is slightly smaller than the number of applicants, which implies that most applicants come from a different school. Furthermore, the number of schools that are part of the program are closer to 11% per year. In the same sense, 38% of all public-funded schools have at least one teacher who applied to AEP during the years 2003-2011. All of this suggests that there is no particular concentration of applicants across the schools.

Table 6.6. Distribution of applicants across the schools.

Application Year	AEP Applicants	Number of schools with at least one AEP applicant	Total number of public-funded schools	% Public-funded schools with at least one AEP applicant
2003	935	785	10,568	7.40%
2004	1,621	1,325	11,448	11.60%
2005	1,834	1,495	11,673	12.80%
2006	2,215	1,723	11,724	14.70%
2007	1,666	1,379	11,795	11.70%
2008	1,661	1,362	11,922	11.40%
2009	1,815	1,386	12,050	11.50%
2010	1,499	1,243	12,112	10.30%
2011	1,316	1,116	12,066	9.20%

[17] *"This percentage is calculated on the basis of the 137,729 teachers with more than 2 years of work experience (counted since their graduation date) who work for at least 20 classroom hours in municipal or private subsidized schools according to the database of the Professional Recognition Bonus Program of the Ministry of Education in September 2011"* (Falck et. al, 2015, p.58).

Applicants' characteristics

In Table 6.7, I present a set of characteristics of the teachers that did apply and did not apply to the program per year. Since teachers instruct at more than one school, variables such as "only teaches at school" are dummy variables equal to 1 if the teacher declares to only instruct in at least one of the schools where he/she teaches.

In terms of teacher characteristics, applicants are on average 41.9 years old, have nearly 15 years of experience, and 73.2% of them are female. In most of the cases, teachers who apply to the program tend to teach at the school and do not have other responsibilities within their school. For instance, 97.2% of the applicants only teach at their school, while 3.5% have other responsibilities. Thus, the applicants' main responsibility is teaching. As shown in Table 6.7, the teachers work in a few different schools, while 66% of applicants have a tenure contract. Lastly, almost all applicants have a bachelor's degree in teaching and graduate education, which ranges from a diploma to a doctoral degree. However, their graduate education is mostly in areas other than education.

In comparing applicants and non-applicants, we can see that in terms of the number of schools where they teach, there are no relevant differences between them. However, there are more important differences in the types of degrees. Applicants for the program are more likely to have an undergraduate degree in education (97% vs 92%) than non-applicants and are more likely to have some graduate education than non-applicants (38% vs 35). Furthermore, applicants are younger, less experienced and contain more female participants than non-applicants.

Some differences between applicants and non-applicants can be explained by the nature of the program. For instance, to apply for the program, the teacher has to have at least 20 hours of

teaching, which reduces the chances of having other responsibilities at the school. Also, as stated by Falck et al. (2015), the fact that females are overrepresented in the applicants group can be explained by the way the program was implemented. The certification categories by subject-grade were progressively expanded, starting by subject-grades that were dominated by females, such as general-primary-education. However, subject-grades that were dominated by males, including science-high-school, were opened four years after the program started and in turn, this made female teachers have a greater chance to apply for the program than males.

Table 6.8 shows some differences in the types of schools where applicants and non-applicants teach. Applicant teachers tended to come more frequently from non-vocational schools than non-applicants (36% vs 29%), are less likely to teach in the metropolitan region (28% vs 34%) and are more likely to teach in public or private voucher-schools than non-applicants, 51% vs 48% and 55% vs 46%, respectively. Moreover, 10% of the teachers changed the pools of schools where they were teaching within the application period (movers) and almost no teachers re-entered the teaching profession after not being a teacher in a public-funded school during the previous period.

Overall, there are some differences between applicants and non-applicants. Applicants are younger and with more specialized degrees than non-applicants, they tended to teach more often and are less involved in administrative duties and were more likely to work in public schools. Moreover, as was described above, the application process is long, and many teachers start the process but do not finish it; thus, it is likely that other factors, such as perseverance, are correlated with the final decision to apply to the program. Therefore, one must be careful when

examining the external validity of the results, since there are some small, but relevant, observable differences between applicants and non-applicants to the program.

Table 6.7. Differences in teacher characteristics between applicants and non-applicants.

Variable	Did not apply	Applied
Teacher characteristics		
Age	43.96	41.89
	(11.81)	(9.135)
Female	0.716	0.732
	(0.451)	(0.443)
Years of experience	16.32	14.80
	(12.36)	(9.559)
Teacher working conditions		
Has other responsibilities besides teaching	0.162	0.0353
	(0.369)	(0.184)
Only teaches at school	0.846	0.972
	(0.361)	(0.165)
Teaches in the school where he/she works the most	0.840	0.968
	(0.366)	(0.176)
Number of schools where the teachers work	1.098	1.155
	(0.348)	(0.416)
Tenure contract	0.670	0.657
	(0.470)	(0.475)
Teacher education level		
Bachelor's degree in education	0.920	0.973
	(0.272)	(0.163)
Some graduate education	0.967	0.989
	(0.178)	(0.106)
Some graduate education with specialization in teaching/education	0.346	0.384
	(0.476)	(0.486)
Where the teacher teaches		
Teaches in non-vocational high school	0.288	0.357
	(0.453)	(0.479)
Teaches in vocational high school	0.0660	0.0876
	(0.248)	(0.283)
Teaches in metropolitan region	0.352	0.294
	(0.478)	(0.456)
Teaches in public voucher-school	0.475	0.511
	(0.499)	(0.500)
Teaches in private voucher-school	0.458	0.550
	(0.498)	(0.498)
Teaches in a rural school	0.146	0.128
	(0.353)	(0.334)

Notes: (i) Standard errors in parentheses, (ii) this table shows the information for all teachers working in public-funded schools.

Certified and non-certified teachers

According to the relative results of the assessment process, AEP is a highly selective process. Up until 2011, 24% of the applicants were awarded this certification (see Table 2.1). The certificated group consists of 4,007 different teachers, who represent 3% of the total number of potential applicants for the program[18].

Next, I examine the differences within the applicants and between teachers who obtained the certification and teachers who did not. Table 6.9 reports difference-in-means of the outcome variables and pre-intervention covariates for the restricted final sample that is used in the specification checks described in Section 8.2 and whose results are presented in Chapter 10. The variables under analysis are grouped into balance and falsification. The former lists the variables that, in the next chapter, are used to run the balance checks to support the validity of the Sharp Regression Discontinuity (SRD) design. While the falsification variables are those outcomes that are analyzed in next chapter as a result of the program after certification is granted, they are also analyzed in the application year to see if there is a treatment effect on placebo outcomes. It is noteworthy that, in this case, the descriptive statistics do not restrict the sample to only observations that are within a pre-determined distance to the cutoff or bandwidth.

[18] This rate was recalculated taking into consideration the AEP applicants until 2011 and the number of eligible teachers in that year (137,729). They are *"teachers with more than 2 years of work experience (counted since their graduation date) who work for at least 20 classroom hours in municipal or private subsidized schools according to the database of the Professional Recognition Bonus Program of the Ministry of Education in September 2011"* (Falck et. al, 2015, p.58).

Table 6.8. Summary statistics and difference-in-means within AEP applicants.

	Non-Certified Group			Certified Group			Difference-in-means		
	N	Sample Mean	Std. Err	N	Sample Mean	Std. Err	Diff-in-means	Std. Err	p-value
Panel A. Balance Variables									
Panel A.1 Teachers' characteristics									
Experience	1,166	14.1	0.30	590	13.0	0.39	-1.2	0.49	0.02
Female	1,168	0.8	0.01	590	0.7	0.02	-0.1	0.02	0.01
Age	1,167	41.9	0.28	589	39.7	0.38	-2.2	0.47	0.00
Spec. Degree	1,168	0.2	0.01	590	0.2	0.02	0.1	0.02	0.00
Tenure	1,168	0.7	0.01	590	0.6	0.02	0.0	0.02	0.10
Metropolitan Region	1,168	0.3	0.01	590	0.3	0.02	0.1	0.02	0.02
Rural Area	1,168	0.2	0.01	590	0.2	0.02	0.0	0.02	0.02
Teacher is a mover	1,036	0.1	0.01	487	0.1	0.01	0.0	0.01	0.91
Teacher is a re-entry	1,036	0.0	0.01	487	0.0	0.01	0.0	0.01	0.28
Panel A.2 Students' Characteristics									
Female Student	41,585	0.49	0.00	22,963	0.50	0.00	0.0	0.00	0.06
Mother's Schooling>12	36,350	0.19	0.00	20,463	0.21	0.00	0.0	0.00	0.00
Book at home>50	37,289	0.19	0.00	20,932	0.24	0.00	0.0	0.00	0.00
Family Income Top Quintile	37,289	0.18	0.00	20,932	0.21	0.00	0.0	0.00	0.00
Panel A.3 Schools' Characteristics									
Public School	1,168	0.50	0.02	590	0.46	0.02	0.0	0.03	0.19
Low-Medium SES School	1,168	0.52	0.02	590	0.42	0.02	-0.1	0.03	0.00
Panel B. Falsification Variables									
Well Prepared	926	0.67	0.01	489	0.75	0.01	0.1	0.02	0.00
Topics Covered	930	0.37	0.01	491	0.41	0.01	0.0	0.02	0.01
Class Preparation>5	804	0.55	0.02	416	0.63	0.02	0.1	0.03	0.01
Test Scores	41,269	-0.03	0.01	22,841	0.18	0.01	0.2	0.01	0.00

Notes: The restricted sample of math and language teachers includes those who applied for math or language certification and were teaching that subject in year t (2004-2011). Only first-time applicants and non-cheaters are considered.

Table 6.9 shows that certified teachers were likely to be significantly less experienced, younger, have some specialized degree, as well as being more likely to teach students from a higher socioeconomic background in higher SES schools. These certified teachers tended to feel better prepared at a higher frequency than teachers who did not obtain certification. Also, those certified in the future were more likely to cover a greater part of the curriculum and spent more time on class preparation than those who remained non-certified. Importantly, the students of certified teachers outperformed students of non-certified teachers. Student test scores of certified

teachers are, on average, 0.2 standard deviations higher than those of students whose teachers received certification the next year. Similarly, Bravo et al. (2008) shows that in 2002, the students whose teachers were certified obtained an average test score that was 0.16 SD and 0.27 SD higher than those obtained by students whose teachers were non-certified or did not apply for the program, respectively. However, certified teachers also taught students whose parents also had more years of education than the parents of students of non-certified teachers and non-applicants. Even though this evidence is informative in terms of how representative the AEP applicants are compared to that of the average teacher in Chile, it also gives an idea that the observed differences in student's test scores cannot be assumed as differences in teacher performance or quality.

Finally, I study the determinants of the AEP scores. To do this, I regress the final score on several teacher and school characteristics, which allows me to calculate partial correlations. Table 6.10 presents the results of this descriptive exercise. I found that teachers who teach in primary school, have more years of experience and have a specialized degree in education performed better in the certification, while older teachers and teachers that teach in vocational schools performed worse.

Taken altogether, all the presented evidence suggests that applicants and certified teachers cannot be considered a representative group of all teachers in Chile. The fact that certified teachers are different, on average, compared to their non-certified counterparts suggests that using a local approximation to estimate the effect of the program is the right strategy, since the global approximation shown suggests that there are significant differences between treated and control teachers in the global sample. This also reinforces the idea that the program effects that

are found in Chapter 10 represent the local average treatment effect (LATE), and its extrapolation to all applicants would need further assumptions. Moreover, the extrapolation of results is also limited due to the significant instance of applicant's self-selection that results in applicants who are systematically different than non-applicants in endogenous variables, such as working conditions and education levels. This might limit the external validity of the results found in this research, which means that the ex-post effect of the program that is estimated cannot be assumed, as the average effect that giving a performance recognition would have if given to an average teacher.

Table 6.9. Determinants of the AEP score. Partial correlations.

Variables	AEP Score
Teacher characteristics	
Age	-0.009***
	(0.00)
Female	0.015
	(0.01)
Years of experience	0.004***
	(0.00)
Teacher education	
Specialized degree in education	0.055***
	(0.01)
Where the teacher teaches	
Teaches in Primary	0.049***
	(0.01)
Teaches in Secondary Non-Vocational	0.002
	(0.01)
Teaches in Secondary Vocational	-0.035**
	(0.01)
Teaches in Metropolitan Region	-0.002
	(0.01)
Teaches in Public School	0.008
	(0.01)
Teaches in Private Voucher School	0.002
	(0.01)
Observations	11,420

Notes: (i) Standard errors in parentheses; (ii) Standard Errors clustered by Year, Year dummies are included; (iii)* p<0.10, ** p<0.05, *** p<0.01

Chapter 7 - Methodology

In this section, summarizing and following the recommendations made by Skovron and Titiunik (2015) in "*A Practical Guide to Regression Discontinuity Designs in Political Science*", Calonico et al. (2017) in "*A Practical Introduction to Regression Discontinuity Designs: Part I*", and Calonico (2017) in "*RDrobust: Software for regression discontinuity designs*", I formally show the parameter of interest for my research questions, its interpretation, how to estimate it, and how to make statistical inferences. I also take into consideration the seminal work presented by Imbens and Wooldridge (2009). The graphical analysis and presentation of the results follow Calonico et al. (2015b) in the paper *"Optimal Data-Driven Regression Discontinuity Plots"*.

7.1. The Parameter of interest

To answer the research questions, I ideally need to identify the causal effect of the certification of the Teaching Excellence Program[19] (AEP for its initials in Spanish) on different outcomes of interest after the teachers received it. The identification strategy that allows me to estimate a credibly causal ex-post effect of the program relies in exploiting the discontinuity in its rewarding process. The evaluation process gave every applicant a score that went from 1 to 4, only those with scores equal or higher to 2.75 were rewarded with the certification.[20] Using this assignment rule, I use a Sharp Regression Discontinuity Design (SRD) to identify the effect of getting the certification given by AEP.

[19] Asignación de Excelencia Pedagógica.

[20] In 2002, teachers with scores equal or higher to 2.5 were rewarded with the certification.

The rationale of the SRD design can help us to understand how it can be an effective identification strategy. The basic idea behind the SRD Design in this case is that assignment to the certification (treatment), is determined completely by the value of the teacher *j*'s final score (the running variable s_j) being on either side of the threshold ($\tilde{S} = 2.75$). This generates a discontinuity of size one in the conditional probability of receiving the certification as a function of the score given by the applicant's evaluation made by AEP. This running variable is itself associated with potential outcomes, and this association is assumed smooth. As a result, any discontinuity of the conditional distribution of the outcome as a function of the score at the cutoff can be interpreted as evidence of a causal effect of AEP (Imbens and Wooldridge, 2009).

Formally, I estimate the local average treatment effect (LATE) of AEP, which is the parameter of interest and can be defined as:

$$\tau_{SRD} = E[Y_j(1) - Y_j(0) | S_j = \tilde{S}] \quad (7.1)$$

$$\tau_{SRD} = E[Y_j(1) | S_j = \tilde{S}] - E[Y_j(0) | S_j = \tilde{S}] \quad (7.2)$$

Here I assume that each teacher j has two potential outcomes, $Y_j(1)$ and $Y_j(0)$, which correspond, respectively, to the outcomes that would be observed following the assignment rule $r_j = 1\{s_j - \tilde{s} \geq 0\}$, where $1\{\ \}$ is the indicator function, equal to one if the applicant *j* has a score value of at least $\tilde{S}$; and zero otherwise. This means that all applicants with a score value less than $\tilde{S}$ are in the control group. This characteristic of the assignment rule r_j is tested by analyzing the discontinuity in the probability of treatment around the cutoff as explained in Section 8.1 The result of this analysis is presented in Section 10.1.

Nevertheless, in Equation 7.2, the two conditional expectations or regression functions $E[Y_j(1)|\, S_j = \tilde{S}]$ and $E[Y_j(0)|\, S_j = \tilde{S}]$ cannot be observed simultaneously. By design there are no teachers with a score that equals the minimum score to obtain the certification ($S_j = \tilde{S}$) for whom we observe its performance (Y_j) with and without treatment—this is the fundamental problem of causal inference. To solve this problem and estimate the regression functions $E[Y_j(r)|\, S_j = \tilde{S}]$, for $r = 0,1$, without making functional form assumptions, I exploit the SRD design that gives the possibility of observing teachers with scores close to $\tilde{S}$. In order for this to work, I make a smoothness assumption that the regression functions are continuous at the cutoff $\tilde{S}$. Under this assumption,

$$E\big[Y_1(0)|S_1 = \tilde{S}\,\big] = \lim_{(S \to \tilde{S})} E[Y_1(0)|S_1 = a\,] \;=\; \lim_{(S \to \tilde{S})} E[Y_1|S_1 = a\,] \tag{7.3}$$

implying that:

$$\tau_{SRD} = \lim_{(\tilde{s}_4 \to \tilde{S})} E[Y_1|S_1 = a\,] - \lim_{(S - \tilde{S})} E[Y_1|S_1 = a\,] \tag{7.4}$$

Equation 7.4 says that the difference between the limits of the average observed outcomes of teachers who received the certification and those who didn't, as the scores converge to $\tilde{S}$ is equal to the LATE at S̃. I call this the SRD treatment effect, defined as Equation 7.1.

7.2. Parameter's interpretation

The SRD treatment effect is defined as local. It might not represent the Average Treatment Effect (ATE) for every teacher and applicant, in consequence, the interpretation of the SRD point estimate needs further explanation. As suggested by Skovron and Titiunik [2015), in the general case where the ATE varies as a function of S_j, the SRD treatment effect may not be informative

about the ATE at values of S_j different than $\tilde{S}$. This may be the case considering that the relationship between the score obtained by the applicant and her potential outcome are arguably non-constant functions of the score. Since the score is related to teachers' ability, resources or other characteristics, those with much higher scores are often systematically different from teachers whose scores are much lower. For this reason, in the absence of specific assumptions about the global shape of the regression functions, from a conservative perspective, the effect recovered by the SRD design in this case is assumed as the local average treatment effect (LATE) at *S*, where the SRD point estimate is identified. This means that results found are representative of the sample of teachers whose score is around the cutoff, which restricts the external validity of the results to this group. This interpretation is also supported by the significant differences between the average covariates for certified and non-certified teachers as shown in Table 6.8.

7.3. Approximating the functional form

Now that I have defined the parameter of interest as the LATE at the threshold, given by $\tau_{SRD} = E[Y_j(1) - Y_j(0)|S_j = \tilde{S}]$, I determine the functional form of $E[Y_j(1) \mid S_j = s]$ and $E[Y_j(0) \mid S_j = s]$. However, the exact functional form of these regression functions is unknown, so they need to be approximated. One way of doing this is by using local polynomial methods (Porter, 2003, Imbens and Kalyanaraman, 2011, Calonico et al., 2014]. This means that the unknown regression function $E[Y_j|S_j = s]$ is approximated locally in a neighborhood of $\tilde{S}$ by a polynomial on the normalized score—i.e., on $s - \tilde{S}$.

As stated by Skovron and Titiunik (2015), the local-polynomial estimation and inference of the SRD point estimate ($\hat{\tau}_{SRD} = \hat{\tau}_1 - \hat{\tau}_0$) consists of the following steps:

1. Choose bandwidth *h*

Local polynomial methods estimate a polynomial using only observations whose scores are between $\tilde{S} - h$ and $\tilde{S} + h$, where *h* is some chosen bandwidth. As suggested by Gelman and Imbens (2014), this local approach is preferable to global methods because, in the latter, observations far from the cutoff can distort the approximation near the cutoff and with that misleading results.

Given the great importance of choosing the size of the bandwidth, the plan is to automatically choose the bandwidth that minimizes an approximation to the asymptotic mean squared error (MSE) of the SRD point estimator, $\hat{\tau}_{SRD}$. In order to avoid the search for specifications and ad-hoc decisions, the *h* is selected in a data-driven fashion, looking to reduce the bias-variance trade-off associated with choosing a small or large bandwidth. This trade-off comes from the fact that choosing a very small *h* reduces the bias of the local polynomial approximation but increases the variance of the estimated coefficients because few observations are used for estimation. Similarly, a large *h* may result in a large bias if the regression function differs from the polynomial approximation, but results in lower variance due to the larger number of observations. This bias-variance trade-off can be minimized by optimizing the MSE of the estimator, which is the sum of its bias squared plus its variance. This procedure involves deriving the asymptotic MSE approximation, optimizing it with respect to *h*, and estimating the unknown quantities in the resulting formula (Calonico et al., 2014, Imbens and Kalyanaraman, 2011). This

approach effectively chooses the *h* that optimizes the bias-variance trade-off, making this bandwidth optimal for point estimation.

2. For each observation *j,* calculate weight $w_j = K((S_j - \tilde{S})/h)$

After having determined which observations are used to locally approximate the regression functions, I assign a weight to each one of these observations. Within the bandwidth, observations closer to $\tilde{S}$ receive more weight than observations further away, where the weights are determined by a kernel function $K[\cdot]$. This estimation approach is nonparametric because it does not assume a particular parametric form of the regression functions.

More precisely, I use a triangular kernel function $K(((S_j - \tilde{S})/h)) = (1 - |(S_j - \tilde{S})/h|)(|(S_j - \tilde{S})/h| \leq 1)$. This is because, when using an optimal MSE bandwidth, triangular kernel at a boundary point leads to a point estimator with optimal variance and bias properties (Skovron and Titiunik, 2015).

3. Choose additional covariates X_j

Before beginning the local-polynomial estimation itself, I decide whether to use additional covariates to improve the fit of the local-polynomial approximation as suggested by Calonico et al. (2016b). They show that a covariate-adjusted RD estimator remains consistent for the standard SRD treatment effect and characterizes precisely the potential point estimation and inference improvements.

Following Calonico et al. (2016b), I choose additional covariates that are "balanced" at the cutoff, which ensures that $\hat{\tau}_{SRD} \rightarrow_p \tau$, where *p* is the chosen polynomial order. This condition is tested empirically while performing the balance checks described in Chapter 8. If the balance

checks show no significant difference of the teachers' characteristics between treated and controls in the application year, t, those that certainly cannot be affected by the treatment in $t+l$ are included as additional covariates in a vector X_j.

The way of illustrating a successful covariate-adjustment in the local polynomial RD estimation is to show a decrease in the length of the confidence interval while simultaneously leaving the point estimate roughly unchanged. This exercise is presented as robustness check in Chapter 10.

Based on implementation history of the program, fixed effects of subject, grade level and application year are included. In terms of grade and subject, it is important to account for the fact that the different grade levels in which both teachers teach, and students are tested have specific characteristics that may influence intermediate and final outcomes. In addition, the evolution of the program over the years needs to be accounted for as well. For example, the tendency of increasing the number of certification categories at the time that the pool of potential new applicants was reduced became a trend in terms of the number and kind of teachers applying every year. The program started certifying only primary teachers, the larger group of teachers in Chile. Years passed and teachers from other less massive grades and subjects were progressively invited to apply, increasing the number of potential applicants at a decreasing rate. At the same time, the self-selection process of a semi-fixed stock of applicants resulted in a reduction of potential applicants. These two processes might have shaped a trend for the characteristics and number of teachers applying to the program, which need to be taken into account.

4. Choose the polynomial order p.

The polynomial order needs to be defined by choosing between the flexibility given by a higher polynomial order and its over-fitting risk. The high polynomial orders bring flexibility and improves the accuracy of the approximation, but it can lead to severe approximation errors due to over-fitting or biases at boundary points (Gelman and Imbens, 2014). In this context, Skovron and Titiunik 2015 explains that in the case of the RD point estimate, since the object of interest is a conditional expectation, the recommended choice is a polynomial of order one, that is, a local—i.e., inside the bandwidth—linear regression.

5. Fit the regression on each hand side and calculate the SRD point estimate.

For observations above the cutoff and within the chosen bandwidth ($\tilde{S} \leq S_j \leq \tilde{S} + h$), I fit a weighted least squares regression of the outcome Y_j on a constant, a vector of covariates X_j, $(\tilde{S} \leq S_j), (\tilde{S} \leq S_j)^2, \ldots, (\tilde{S} \leq S_j)^p)$ where p is the chosen polynomial order one, with the weight ω_j for each observation. The estimated intercept is an estimate of τ_1 is the SRD point estimate ($\hat{\tau}_{SRD} = \hat{\tau}_1 - \hat{\tau}_0$). For the left-hand side, the same procedure is followed, but using observations below the cutoff and within the chosen bandwidth ($\tilde{S} - h \leq S_j \leq \tilde{S}$). In this case, the estimated intercept is an estimate of τ_0. Now, the SRD point estimate can be calculated as ($\hat{\tau}_{SRD} = \hat{\tau}_1 - \hat{\tau}_0$), which is an optimal and consistent estimator (Skovron and Titiunik, 2015). This follows from assuming that the bandwidth sequence shrinks appropriately as the sample size increases and using MSE-optimal bandwidths. However, Skovron and Titiunik (2015) also explains that the rate of convergence of the MSE-optimal bandwidth leads to a bias in the

distributional approximation of the estimator that is used to create confidence intervals. Ignoring the bias term leads to invalid inferences. This is why an additional step needs to be taken in order to make valid hypothesis tests.

6. Local Polynomial Inference.

As stated by Skovron and Titiunik (2015), the MSE-optimal bandwidth is designed to be optimal for point estimation, not for inference. To address this issue, I use the robust inference method developed by Calonico et al. (2014) (CCT method henceforth) to have smaller coverage errors when using the MSE-optimal bandwidth. This CCT method, besides being a data-driven method that reduces the ad-hoc decisions, results in robust confidence intervals that lead to valid local polynomial inference when the MSE-optimal is used (Calonico et al., 2015a).

These six steps are followed by using the *robust* command in STATA that was developed by Calonico et al. (2016a). This is the CCT method that allows me to fit a weighted linear least squares regression of the outcome within MSE-optimal bandwidths that optimize the bias-variance trade-off. In addition, by using the *rdrobust* command, robust inference methods are used. In consequence, the proposed estimation procedure leads to a consistent and optimal (in an asymptotic MSE sense) covariate-adjusted SRD point estimator of the LATE, enabling robust confidence intervals, valid inferences and hypothesis tests for the SRD point estimate, $\hat{\tau}_{SRD}$.

The next step is related to clustering the standard errors. In this case the units of analysis are teachers, who are clustered into schools. This may result in errors correlated within but not across schools. In such settings default standard errors can greatly overstate estimator precision. Thus, clustering the standard errors at school level is needed. As suggested by Calonico et al.

(2017), it may be appropriate to employ variance estimators that are robust to the clustered nature of the data. The *robust* command allows for employing cluster-robust variance estimators, which results in different estimated standard errors and bandwidths relative to the unclustered case. In addition, cluster-robust variance estimators can be smaller or larger than variance estimators that do not account for clustering. As explained by Calonico et al. (2017), all this means that when cluster-robust variance estimators are employed, cluster-robust standard errors can lead to re-centered confidence intervals that can be either shorter or longer in length.

In practice the main estimations include standard errors at school level and fixed effect for grade and subject, and application year.

7.4. Graphical analyses

The SRD design can be illustrated graphically by a scatter plot of the aggregated observed outcomes against the score values. In this case, the outcome of interest can be student test scores, teachers' effort or self-esteem, is plotted against the distance between the applicant's score and the cutoff. The graphical analysis with the RD plots is very informative by showing more global polynomial fit, and local sample means. On one hand, the global polynomial fit allows us to see if there is a discontinuity in the outcome around the cutoff. On the other hand, the RD plot shows the local sample means of the outcome computed over an evenly-spaced partition of the support of the running variable with binned sample means mimicking the underlying variability of the data, for certified and non-certified teachers separately. In order to get the most out of the RD plots, I use those recommended by Calonico et al. (2015b) and implemented using STATA's command (*rdplot*): RD plots.

Chapter 8 - Empirical strategy

Linked to the previous chapter, this chapter describes the analysis that is carried out. Therefore, this chapter starts with the Sharp Regression Discontinuity (SRD) design validation. This includes analysis of the discontinuity in the probability of treatment around the cutoff and running specification checks. The empirical strategy also describes some potential methodological problems and how they are tested in order to provide further evidence on the robustness of the results. Lastly, the analysis of the outcomes of interest is explained.

8.1. Analysis of the discontinuity in the probability of treatment

First of all, I test if a SRD design should be used. I look for basic evidence of the existence of a valid SRD design, which is the discontinuity of size one in the probability of being assigned to the treatment group around the cutoff. To check this, the relationship between the distance to the cutoff and the frequency of receiving the Teaching Excellence Award[21] (AEP for its initials in Spanish) is plotted. This graph simultaneously includes the first application of teachers in the sample for all years of the program (2002-2011) excluding those who were caught cheating. The exclusion of this group of applicants is unlikely to affect the global interpretation of the plot because they represent less than 0.54% of the total number of applications (89).

8.2. Specification checks

Having tested the discontinuity in the probability of treatment around the cutoff, further evidence

[21] Asignación de Excelencia Pedagógica.

about the design's validity is presented. I indirectly test the smoothness assumption[22] that provides support of the SRD design. The smoothness assumption can be indirectly tested using various implications of the identification argument underlying the SRD design. Specifically, I test if there are: (i) no other changes of other covariates at the same threshold, and (ii) any manipulations of the score that underlies the assignment mechanism. Although the proposed tests do not directly test the null hypotheses required for the SRD approach to be valid, they allow me to argue for the approach's validity if these null hypotheses hold.

Following Imbens and Wooldridge (2009)'s recommendations, I carry out three sets of specification checks. The first one is the density test of the running variable to test for manipulation of the assignment rule. The second and third specification sets test for a non-treatment effect on predetermined covariates and placebo outcomes respectively. Both null hypotheses are expected to be non-rejected in order to provide further support to the SRD design's validity. Complementarily to these three specification checks, a final falsification test for the non-treatment effect with alternative cutoffs is performed following Skovron and Titiunik (2015)'s recommendations.

Testing manipulation of the running variable

A discontinuity in the density of the running variable (the score (S_j) obtained for the applicant) at the particular point where the discontinuity in the conditional expectation occurs is suggestive of violations in the assumption of non-manipulation of the assignment rule. To check for this, I test the density of the running variable by looking whether, in a local neighborhood near the cutoff

[22] See Section 7.1 for further details.

($\tilde{S}$), the number of observations below the cutoff is considerably different from the number of observations above it. To do so, I show a histogram of the running variable and visually check whether the number of observations above and below the cutoff is similar. Additionally, I formally test the null hypothesis of continuity in the density of the covariate that underlies the assignment at the threshold, against the alternative of a jump in the density function at that point. Here the focus is on the difference $\tau_{f(s)} = lim_{\tilde{s}\to s}(E[f_{\tilde{s}}(s)] - lim_{s\to\tilde{s}}(E[f_{\tilde{s}}(s)]$. In practice, I perform McCrary (2008) density test in STATA by using the command *DCdensity*. A substantially and statistically significant difference in the left and right limits would suggest that there may be problems with using the SRD approach.

Testing treatment effect on predetermined covariates

A discontinuity in other covariates around the cutoff casts doubt on the RD underlying assumptions. In fact, such a discontinuity goes against the assumption that teachers around the cutoff do not systematically differ in their unobservable characteristics, thereby offering valid counterfactual comparisons between control and treatment groups. To test this assumption, I look for discontinuities in the average value of pre-treatment covariates around the threshold. Specifically, for covariates Z_j, the test would look at the difference $\tau_Z = lim_{\tilde{s}\to s}(E\big[Z_j\big(S_j = s\big)\big]) - lim_{s\to\tilde{s}}(E\big[Z_j\big(S_j = s\big)\big])$.

In practice, I perform a balance check on predetermined covariates of the treatment and control groups. These estimates the treatment effect of AEP on those covariates. The null hypothesis of no treatment effect on the covariates around the threshold is expected to hold.

The balance checks are calculated for the application period t on teacher characteristics (experience, gender, age, having a specialized degree, tenure[23], teaching in the metropolitan region, teaching in a rural school, the number of schools and classrooms where the teacher works, if the pool of schools where the teacher taught in $t+1$ is different from the pool of school the teacher taught in t (mover), and if the teacher was not in teaching in a public-funded school in $t+1$, but he was teaching in t (leaver), student characteristics (student gender, having a mother with more than 12 years of education, having more than 50 books at home, and being in the wealthiest quintile), and school characteristics (if the school publicly funded and administrated, and if the school's enrollment has on average low or medium socioeconomic status (SES)). Importantly, the balance check includes data for teachers assigned to classrooms that take the national standardized test[24] (SIMCE for its initials in Spanish) in the same application year (*t*).

The linear polynomial specification of the regressions to run the balance check on teacher characteristics (Equation 8.1), student characteristics (Equation 8.2) and school characteristics (Equation 8.3) correspondingly are:

$$B_{jt} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + a_2(S_{jt} - \tilde{S}_t) + \alpha_3(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \omega_t + \upsilon_{jt} \quad (8.1)$$

$$B_{ijt} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + a_2(S_{jt} - \tilde{S}_t) + \alpha_3(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \omega_t + \upsilon_{jt} \quad (8.2)$$

$$B_{jkt} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + a_2(S_{jt} - \tilde{S}_t) + \alpha_3(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \omega_t + \upsilon_{jt} \quad (8.3)$$

[23] Tenure means that there is an open-ended employment contract. Teachers may be fired only under few and very specific conditions.

[24] Sistema Nacional de Medición de la Calidad de la Educación.

In Equation 8.1 B_{jt} is the characteristic of teacher *j* in the application year *t*, while in Equation 8.2 B_{ijt} is the characteristic of the student *i* of teacher *j* in the application year *t*, and in Equation 8.3 B_{jkt} is the characteristic of the school *k* where teacher *j* teaches in application year *t*. S_{jt} is the score of teacher *j* in year *t*, $\tilde{S}_t$ is the cutoff in year *t*. Application year, subject and grade dummies (ω_t) are included.

As a way to test the robustness of the balance checks, I also show the results of estimating Equations 8.2 and 8.3 including a set of balanced pre-treatment covariates (X_{jt}). This set of covariates is defined after testing the balance of teacher *j*'s characteristics by the estimation of Equation 8.1.

This linear specification with interactions does not constrain the slope of the outcome/score relationship to be identical on both sides of the cut-point. In fact, Equations 8.1, 8.2 and 8.3 specify a different polynomial function of score on either side of the threshold.

Testing treatment effect on placebo outcomes

Complementarily, Imbens and Wooldridge (2009) and Skovron and Titiunik (2015) recommend studying the treatment effect on placebo outcomes by the implementation of **falsification tests**. Specifically, the exercise is to estimate the program effect on students' test scores and teachers' behavior during the application period. These are outcomes that should not be affected by the treatment in the application year since the teachers have not yet received the program's results. As in the balance checks, the falsification tests use data linked to teachers assigned to classrooms that take the SIMCE test in *t*.

The falsification test is performed with the same nonparametric local linear method used to estimate the balance check. In this case the specification is the following:

$$F_{ijt} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + \alpha_2(S_{jt} - \tilde{S}_t) + \alpha_3(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \omega_t + \upsilon_{jt} \quad (8.4)$$

This specification includes three different placebo outcomes (F_{ijt}): (i) the math or language test scores of student *i* teacher *j* in the application year *t*; (ii) the effort of teacher *j* in year *t*; and (iii) the self-confidence of teacher *j* in year *t* (F_{jt}). The standard errors are clustered at school level. Lastly, as done with the balance checks, the robustness of the results of the falsification tests to the inclusion of covariates is also included.

As explained in Chapter 7, Equations 8.1, 8.2, 8.3 and 8.4 are approximated by using the nonparametric local polynomial approach within a mean-squared-error (MSE) optimal bandwidth (optimal for point estimation) (Calonico et al., 2014) and with robust inference methods (Calonico et al., 2015a). In practice, to do so I use the STATA command *rdrobust*. As suggested by Gelman and Imbens (2014), I use estimators based on linear polynomials ($p = 1$). Given that nested nature of the data, I cluster standard errors at school level and include fixed effects for application year and grade.

Testing treatment effect for alternative cutoffs

Additionally, Skovron and Titiunik (2015) recommends performing a falsification test that involves replacing the true cutoff value with another value and performing an estimation and inference of the treatment effect on the outcome of interest. This means that Equation 8.6 is estimated with two fake cutoffs: 0.25 points above the real cutoff and 0.2 below. It is expected to find not statistically significant effect of the "false" program on student achievement.

Only one adaptation of the method is done in order to avoid "contamination" due to real treatment effects. As suggested by Calonico et al. (2017), the data is restricted to the appropriate group: only treated observations for cutoffs above the actual cutoff, and only control observations for cutoffs below the actual cutoff should be considered.

The results of this last falsification test are presented after the full and unbundled program effect on final outcomes is estimated in Chapter 10, as a way of showing evidence to support the validity of the results obtained.

8.3. Analysis of two potential methodological problems

I address two methodological problems that, at least in theory, might threaten my identification strategy. The first problem is the potential sorting of teachers to students previous to the award. The estimates for $t + l$ may be biased by previous students' assignment patterns that (dis)favored certified teachers. For example, the principal, knowing the skills distribution among her teachers, may have decided to assign low-performing students to future AEP teachers as a targeted strategy to help those students. Following Taylor and Tyler (2012) where they study the effect of evaluation on a sample of mid-career math teachers in the Cincinnati Public Schools, I investigate this potential issue by looking for evidence of bias when comparing observable teacher and student characteristics across treatment groups before the treatment was given. Specifically, I test whether the average observable variables of certified and non-certified teachers are significantly different in the application year (t). In addition, I estimate the effect of the program on students' test scores for the period of application *t* in order to test if there was a trend in the assignment of students to teachers, which may bias the program effect estimator. In other words, the balance checks and falsification tests performed for math and language teachers

as part of the specification checks help examine the sorting issue. If sorting is not a problem, there would be non-significant effects of the AEP on covariates and placebo outcomes in period *t*. Importantly, finding sorting in $t + l$ is considered as a part of the program's intermediate effect and not necessarily as a methodological problem. These effects are evidence for testing the Motivational Model's assumption—the *ceteris paribus* assumption.

The second potential methodological problem is attrition, whereby the treatment affects observation of the outcome of interest (Lee and Lemieux, 2010). Importantly, this could drive the estimates' results. This problem is highly unlikely for $t + 1$ because the reward process takes place after the school year has started. However, for $t + 2$ and $t + 3$ it is more likely. Thus, non-random attrition of the sample in $t + l$ might exist if high-performing teachers—who were more likely to obtain the certification—tended to be more frequently assigned to the students who took the SIMCE test in $t + l$. To investigate this possible threat, I test the existence of attrition by regressing the probability of attrition in $t + l$ on the result of the AEP, as described by,

$$A_{jt+\iota} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + \eta(S_{jt} - \tilde{S}_t) + \psi(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \omega_t + \upsilon_{jt} \qquad (8.5)$$

Where A_{jt+l} is a dummy variable that takes the value 1 if the teacher *j* is observed in the dataset $l = 1, 2, 3$ years after the application and 0 otherwise. The fact that the teacher *j* is observed means that she was assigned in $t + l$ to a classroom that took the SIMCE test in that year. S_{jt} represents teacher *j*'s final score, and S_t is the minimum score required to be awarded. $1[S_{jt} - \tilde{S}_j]$ indicates whether a teacher's final score is greater or equal to the cutoff. Finally, w_t are application year, subject and grade dummies. This specification for the attrition test is performed using the *rdrobust* STATA command to estimate the same nonparametric local linear

method as in the estimation of the balance checks and falsification tests. In practice, data for applicants from year 2003 until 2011 is used for these estimates. The standard errors are clustered at school level.

If the probability of attrition does not depend on the probability of being rewarded, we can assume that the sample attrition was random, positing no threat of bias to the estimates.

8.4. Outcome variable analysis

Having tested the validity of the SRD design, the estimation of the SRD needs to define the outcome of interest, and the model specification. These definitions are critical for answering the research questions previously defined, see Table 8.1. Then the proposed robustness checks are described.

Outcome of interest

In general, $y_{(ijt+z)}$ is the outcome at year $t + l$, where l is the number of years after the application year, i.e., $t + 1$ is both one year after the application and the year of receiving the treatment. As explained in Chapter 5, in 2007 and 2008, AEP only included the financial incentive; thus, estimating the program's effect for those years is equivalent to isolating the effect of the financial component, while using the data for the rest of the years provides information for the effects of the Full program. The outcome definition varies with the research question being explored in the following way:

- To answer ***Question #1 Does AEP affect teacher's performance? (Full program effect),***

y_{ijt+l} is student i's SIMCE test score for applicant teacher j in $t + l$ for application years t 2003 to 2006, and 2009 to 2011; $l \in (1, 2, 3)$.

- To answer ***Question #5, Does the effect of AEP fade-out? (Fade-out of the full program effect)***, y_{ijt+l} is student *i*'s SIMCE test score for applicant teacher *j* in $t + l$ where $l \in (1, 2 \text{ or } 3)$ for application years 2003 to 2006, and 2009 to 2011.

- To answer ***Question #2 Does each component of AEP affect teachers' performance? (Unbundled program effect),*** y_{ijt+l} is student i's SIMCE test score for the applicant teacher j in $t + l$ for application years t 2007 to 2008; $l \in (1, 2, 3)$.

- To answer ***Question #3 Does AEP affect teachers' behavior? (Underlying mechanism of the full program),*** y_{jt+l} is teacher *j*'s self-confidence (self-report on how prepared the teacher feels to teach the curriculum evaluated by SIMCE) in $t + l$ for application years *t* 2003 to 2006, and 2009 to 2011. Additionally, y_{jt+1} is teacher *j*'s effort level (proportion of the curriculum that teacher covers during the academic year and number of hours spent in class preparation) in $t + l$ for application years *t* 2003 to 2006, and 2009 to 2011; $l \in (1, 2, 3)$.

- To answer ***Question #4, Does each component of AEP affect teachers' behavior? (Underlying mechanism of each program component),*** y_{jt+l} is teacher j's self-confidence or the effort level in $t + l$ for application years t 2008 to 2009; $l \in (1, 2, 3)$.

Regression discontinuity estimates

The SRD estimates are obtained using two supplemental methods: one parametric and another nonparametric. Using these two methods allows us, on the one hand, to estimate the effect of the full and the unbundled program, and, on the other hand, to make a specification robustness check of the results found. For a better understanding of the estimation methods, they are explained

below; complementarily, Table 8.1. shows the relationship between the outcome and parameter of interest in Equations 8.6 and 8.7, with the corresponding research question for each method.

A nonparametric local linear polynomial estimation method: the robust specification

In practice, after defining the outcome of interest, I determine the regression to estimate the SRD treatment effect $_{SRD}$. With that, the **robust bias-corrected local linear regression** that I use is:

$$Y_{ijt+1} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + a_2(S_{jt} - \tilde{S}_t) + \alpha_3(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \omega_t + \upsilon_{jt} \quad (8.6)$$

Y_{ijt+l} is outcome of interest within three years after application year–i,e, $l = 1, 2$ *or* 3. The outcome depends on the research question being addressed, S_{jt} is the score of teacher j in year t, $\tilde{S}_j$ is the cutoff in year t, and $1[S_{jt} - \tilde{S}_j]$ is 1 if the teacher j was certified in $t + 1$. Application year, subject and grade dummies (ω_j).

This specification has important features and implications. First, it is estimated twice. Once to estimate the full program effect with data for years t 2003 to 2006 and 2009 to 2011, and another to estimate the financial component effect with data for years t 2007 and 2008. In both cases, α_1 is the parameter of interest that represents the local average treatment effect. Second, the specification aims to locally approximate the regression functions by using a polynomial of grade 1, as recommended by Gelman and Imbens (2014), to improve the estimation of causal effects. Third, it allows the approximation of the two regression functions by differing on either side of the threshold. Fourth, this specification allows to use the *rdrobust* STATA command to perform the local linear polynomial method to approximate the regression functions within the MSE-optimally bandwidth as suggested by Skovron and Titiunik (2015). This nonparametric

approximation uses a triangular kernel to weight observations and to fit a weighted least squares regression of Equation 8.6. The standard errors are clustered at school level.

A parametric estimation method: the alternative specification

In order to show the unbundled program effect on teacher's behavior and performance, I estimate the following alternative parametric specification:

$$Y_{ijt+1} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + a_2(S_{jt} - \tilde{S}_t) + \alpha_3(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} +$$
$$\beta_1\{S_{jt} - \tilde{S}_t \geq 0\}Pin_t + \beta_2(S_{jt} - \tilde{S}_t)Pin_t + \beta_3 Pin_t(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \omega_t + \upsilon_{jt} \quad (8.7)$$

Y_{ijt+1} is the outcome of interest depending on the research question being addressed, S_{jt} is the score of teacher j in year t, $\tilde{S}_j$ is the cutoff in year t and $Pin_t = 1$ if in $t+1$ there was a ceremony to give a pin to each teacher who applied in t and was certified in $t+l$. Year, subject and grade dummies (ω_j) are included.

In this alternative specification, $_1$ is the effect of the program when only financial incentives are provided; and $\alpha_1 + \beta_1$ provides the effect of the program when financial incentives and non-financial incentives (ceremony and pin) were given. Thus β_1 is the additional effect of the non-financial reward.

It is important to notice that this alternative specification cannot be estimated using STATA command *rdrobust* because it includes interaction terms needed to estimate the unbundled effect of the program. This implies that the polynomial estimation within the MSE-optimally defined bandwidth cannot be calculated. Instead, a least-squares regression is used with a triangular kernel to weight observations within a certain data-driven neighborhood. In practical terms, this means running Equation 8.7 by using the observations within the same two bandwidths that were

automatically calculated when estimating Equation 8.6 by using the CCT method. Therefore, I can recover the bandwidths estimated by Equation 8.6 for the full program, to estimate the full program's and the pin's effect by running Equation 8.7. In the same way, I use the bandwidth estimated for the financial component. The standard errors are clustered at school level.

Parametric versus nonparametric method

The alternative specification has important features that make it both similar and different to the robust specification defined by Equation 8.6. They are similar because both aim to locally approximate the two regression functions independently by using a polynomial of grade 1. Despite these similarities, the alternative specification uses pooled data for years 2003 to 2011 in order to test for differences in the program components simultaneously. Another difference has to do with the fact that the alternative specification imposes a parametric form on the unknown regression functions, while the robust specification leaves these functions unspecified and employs nonparametric local polynomial methods for estimation and inference.

Parametric and nonparametric approaches are different but complementary. The nonparametric local linear polynomial approach has three distinctive features:

(i) the bandwidth is chosen in a data-driven way based on nonparametric approximations, (ii) the RD point estimator is asymptotically MSE-optimal, and (iii) inference procedures explicitly incorporate the effects of local parametric misspecification (i.e., nonparametric smoothing bias) (Cattaneo et al., 2017, p. 654).

While the alternative specification—Equation 8.7—represents a parametric method that does not account for misspecification bias in estimation and inference procedures. Nevertheless, and

despite the positive features, the robust specification does not allow the interaction needed to estimate the marginal effect of the pin over the financial component of the program. This is why the parametric estimation method is helpful as a complement to identify the unbundled effect.

Robustness Checks

In order to show that the results in the next chapter are not driven by the choices made to estimate the program effect, two strategies are followed. The first one is related to the specification of the model, while, the second one is related to the sensitivity of the results to the bandwidth that is automatically calculated by the CCT Method.

First, I estimate the robust and alternative specifications with covariates following Equations 8.8 and 8.9 and clustering the standard errors at school level. These are similar to Equations 8.6 and 8.7 but include a vector of covariates (X_{jt}), which might bring efficiency gains (Calonico et al., 2016b). Due to this, the set of covariates are defined by taking into consideration the balance check on teachers' characteristics as described in Section 8.2. This set of covariates might comprise teacher *j*'s characteristics such as experience, and gender. In doing so, I apply the same estimation method but with covariates in all estimations except to those balance checks. The results of this exercise are presented in Appendix I.

$$Y_{ijt+l} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + a_2(S_{jt} - \tilde{S}_t) + \alpha_3(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \alpha_4 X_{jt+l} + \omega_t + \upsilon_{jt+l} \quad (8.8)$$

$$Y_{ijt+l} = \alpha_0 + \alpha_1 1\{S_{jt} - \tilde{S}_{jt} \geq 0\} + a_2(S_{jt} - \tilde{S}_t) + \alpha_3(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \beta_1\{S_{jt} - \tilde{S}_t \geq 0\}Pin_t + \beta_2(S_{jt} - \tilde{S}_t)Pin_t + \beta_3 Pin_t(S_{jt} - \tilde{S}_t) \times 1\{S_{jt} - \tilde{S}_t \geq 0\} + \beta_4 X_{jt+l} + \beta_4 X_{jt+l} Pin_t + \omega_t + \upsilon_{jt} \quad (8.9)$$

The second strategy is a sensitivity analysis of the results to the window length. Given that the bandwidth for each estimation varies, the sample of teachers used to estimate the effect of the program on final and intermediate outcomes also varies. This is because the CCT Method automatically calculates the data-driven MSE-optimal bandwidth for each estimation done with the different outcomes of interest. Thus, using the CCT Method does not follow the more traditional way of presenting the RD results, which is using a fixed bandwidth for all estimations.

In this context, the sensitivity analysis is done in order to see if the results vary as a function of the bandwidth size in a way that affects its statistical significance. The analysis shows the p-values calculated when testing the null hypothesis on the parameters of interest for a range of pre-determined bandwidths.

More precisely, this sensitivity exercise is constructed in the following way. First, I define a list of values for the bandwidth. Then, for each window I estimate the treatment effect using the robust specification with Equation 8.6. Lastly, I recover the robust p-value of testing the hypothesis of null treatment effect. Each p-value is plotted against the bandwidth size and presented at the end of the results' section in Chapter 10. The corresponding plots also highlight the p-values shown in main results helping make a comparison between this p-value found using a data-driven bandwidth and those found when the program effect is estimated using fixed windows for the running variable.

Table 8.1. Research questions, parameters, and outcomes.

Research Question	Outcome of Interest in Equation 8.6	Parameter of Interest in Eq. 8.6	Outcome of Interest in Eq. 8.7	Parameter of Interest in Eq. 8.7
Question #1, Does AEP affect teacher's performance? (Full program effect)	y_{ijt+l} is student *i*'s SIMCE Test Score of AEP teacher *j* in $t+l$ for years *t* 2003 to 2006 and 2009 to 2011, with $l = 1,2,3$.	α_1	y_{ijt+l} is student *i*'s SIMCE Test Score of AEP teacher *j* in $t+l$ for years *t* 2003 to 2011, with $l = 1,2,3$.	$\alpha_1 + \beta_1$
Question #3, Does each component of AEP affect teachers' performance? (Unbundled program effect)	y_{ijt+l} is student *i*'s SIMCE Test Score of AEP teacher j in $t+l$ for years t 2007 and 2008, with $l = 1,2,3$.	α'_1 for the financial component		α_1 for the financial component; β_1 for the non-financial component.
Question #4.1, Does AEP affect teachers' self-confidence? (Underlying mechanism of the full program)	y_{jt+l} is teacher *j*'s self-confidence (self-report on how much teacher feels prepared to teach the curriculum that were evaluated by SIMCE) in $t+l$ for years t 2003 to 2006 and 2009 to 2011, with $l = 1,2,3$.	α_1	y_{jt+l} is teacher *j*'s self-confidence (self-report on how much teacher feels prepared to teach the curriculum that were evaluated by SIMCE) in $t+l$ for years *t* 2003 to 2011, with $l = 1,2,3$.	$\alpha_1 + \beta_1$
Question #5.1, Does each component of AEP affect teachers' self-confidence? (Underlying mechanism of each component of the program)	y_{jt+l} is teacher *j*'s self-confidence (self-report on how much teacher feels prepared to teach the curriculum that were evaluated by SIMCE) in $t+l$ for years t 2007 to 2008, with $l = 1,2,3$.	α'_1 for the financial component		α_1 for the financial component; β_1 for the non-financial component.
Question #4.2, Does AEP affect teachers' effort (Underlying mechanism of each component of the program)	y_{jt+l} is the teacher *j*'s effort level (proportion of the curriculum that teacher covers during the academic year, or class preparation time) in $t+l$ for years t 2002 to 2006 and 2009 to 2011, with $l = 1,2,3$.	α_1	y_{jt+l} is the teacher *j*'s effort level (proportion of the curriculum that teacher covers during the academic year and class preparation) in $t+l$ for years *t* 2004 to 2011.	$\alpha_1 + \beta_1$
Question #5.2, Does each component affect teachers' effort (Underlying mechanism of each component of the program)	y_{jt+l} is the teacher j 's effort level (proportion of the curriculum that teacher covers during the academic year, or class preparation time) in $t+l$ for years t 2007 and 2008, with $l = 1,2,3$.	α'_1 for the financial component		α_1 for the financial component; β_1 for the non-financial component.
Question #2, Does the effect of AEP fade-out? (Fade-out of the full program effect)	y_{ijt+l} is student *i*'s SIMCE test score for applicant teacher *j* in $t+l$ where l can take the value 1, 2 or three. This is for application years *t* 2002 to 2006, and 2009 to 2011.	α_1	-	-

Chapter 9 - Variables and Measures

In order to estimate Equations 8.6 and 8.7 as explained in Chapter 8, the outcome of interest and the covariates need to be defined according what factors are to be tested (Motivational Model), the research questions, and the methodological considerations previously described. As presented in Table 8.1, there are three outcomes of interest to be measured. Specifically, a measure of teachers' performance, a measure of teachers' self-esteem, and a measure of teachers' effort are needed. In this regard, the following section presents a discussion on how these variables and measures can be interpreted and their limitations.

9.1. Teachers' Performance

The main and expected effect of the program was to foster educational quality. As stated in Chapter 2, this objective of the program was defined by Law 19715 as "*to strengthen the quality of education and to recognize and highlight the merit of teachers, fostering their retention at teaching and helping identify those that show knowledge, skills and competencies of excellence*." Thus, at this point, the question is how to measure the program effect on the quality of the educational system. One—at least partial—answer to this question is to measure the effect of the program on student learning, which is assessed by the Measurement System of the Quality of Education[25] (SIMCE for its initials in Spanish) in math and reading. In this context, the following question turns out to be an explanation under which conditions a change in these SIMCE test scores can be potentially understood as a change in teachers' performance. Having answered it allows to address the first research question.

[25] Sistema Nacional de Medición de la Calidad de la Educación.

Bringing a valid measure, that can be credibly attributed to teacher's performance, has been a practical, theoretical and empirical challenge for academics and policy makers. Test scores as a measure of student achievement have often been used to approximate teachers' performance or effectiveness. Loeb (2013) explains that the use of student test scores to measure teaching practice is both an advantage and a disadvantage. She notes that the clear benefit is that, however imperfect, test scores are a direct measure of student learning, which is the key outcome. Students who learn more in school tend to complete more schooling, have greater earning potential, and lead healthier lives. In addition, she mentions that basing teacher assessments on student learning also recognizes the complexity of the teaching process; many different teaching styles can benefit students. This is why school districts in the US have explored complementing test scores with multiple measures of teacher evaluation, as is the case in New York City and Washington, DC (Grossman et al., 2013). This might be also important considering that it's not fair to judge teachers' effectiveness solely on the basis of end-of-year test scores, without regard to where the teachers' students started at the beginning of the year. End-of-year test scores do not show how much students learned that year in that class, so measures that take into account where students started are surely an improvement. In this context, value added measures (VAM) are developed to compute the teacher's unique contribution in promoting student achievement gains from grade to grade, net of student background and prior ability. Because VAM adjust for student characteristics in a given classroom, they are less biased measures of teacher performance than unadjusted test score measures.

Despite the advantages of using a VAM, the available data does not allow for it. In Chile, the national government developed the SIMCE; a yearly, national and standardized student test. All

students in 4th grade yearly and 8th and 10th grade every other year take the test, thus follow-up of student achievement is not possible. This kind of cross-sectional data does not allow for calculating a value added in order to provide a more valid/credible measure for teacher performance whenever estimating the program effect.

Nevertheless, SIMCE test scores as a measure of student learning can be interpreted as an approximation of teachers' performance when estimating the effect of the Teaching Excellence Award[26] (AEP for its initials in Spanish) under a Sharp Regression Discontinuity (SDR). There are at least two reasons suggesting this. First, the theoretical and empirical modeling allows me to set and test the conditions under which the program effect on test scores can be attributed to a change in teachers' behavior. Second, the identification strategy allows me to make a causal interpretation of the program effect on teachers' performance measured by student outcomes.

Specifically, below is a discussion on two points regarding using test scores and how they capture the teacher effect: (i) SRD design as a good identification strategy that provides consistent results even though I use cross sectional instead of longitudinal data; (ii) when changes in test scores can be interpreted as a measure of teacher' performance in the context of the Motivational Model (MM).

The Identification Strategy and VAM

The problem with using cross sectional test scores to measure teachers' performance is the lack of adjustment for student background and prior ability (Loeb, 2013). However, the need of

[26] Asignación de Excelencia Pedagógica.

adjustment becomes less crucial when exploiting the discontinuity in the probability of reward as the strategy to identify the effect of AEP on teachers' performance.

Comparing the average test scores of different teachers, may not result in an accurate rank of their effectiveness, since the raw scores do not take into consideration differences in the starting point of the students. However, in theory, SRD design assumes that units around the cutoff do not systematically differ in their unobservable characteristics, thereby offering valid counterfactual comparisons between control and treatment groups (Skovron and Titiunik, 2015). This means that, around the threshold that defines who obtains the reward, the students are on average expected to be equal, therefore, they have the same starting point. Even though this SRD design underlying assumption is by definition untestable, I perform several specification checks (balance checks and falsification tests) that help support the design's assumption. As described in Chapter 8, I test for discontinuities on the average observables of students and school characteristics around the cut-off. The results of the specification checks are shown in Chapter 10. There, the results show, as expected, that there were not statistically significant differences in teacher's behavior, prior test scores, student background and school characteristics between treated and controls, previous to receiving the reward. This adds evidence to the argument that finding a discontinuity in student outcomes after the reward was given can be interpreted as a causal average treatment effect (ATE) of AEP on teachers' performance without any need of a VAM. Despite this, I make a further conservative interpretation of the recovered effect, as explained below.

As detailed in Section 7.2, the SRD treatment effect found might not represent the ATE for every teacher and applicant (Skovron and Titiunik, 2015). The specification checks allow me to

say that the SRD design's assumption holds for those teachers whose scores are close to the cutoff point. However, it might be the case that the relationship between the running variable and potential outcomes are arguably non-constant functions. Since the score is related to teachers' ability, resources, or other teacher characteristics, teachers with much higher scores were often systematically different from those whose scores are much lower. Moreover, it is possible their students were also different, perhaps violating the SRD design's assumption that reduces the need for a VAM. For this reason, in the absence of further assumptions, the effect recovered by the SRD design is conservatively interpreted as the local average treatment effect (LATE) of the program on test scores for those teachers whose AEP scores are close to the cutoff point—making VAM dispensable in representing teacher's performance. In consequence, interpreting the SRD point estimate as the LATE of AEP, makes it more likely that the SRD design's assumptions hold, bringing support for using cross sectional test scores as a valid proxy of teacher's performance, without need for a VAM.

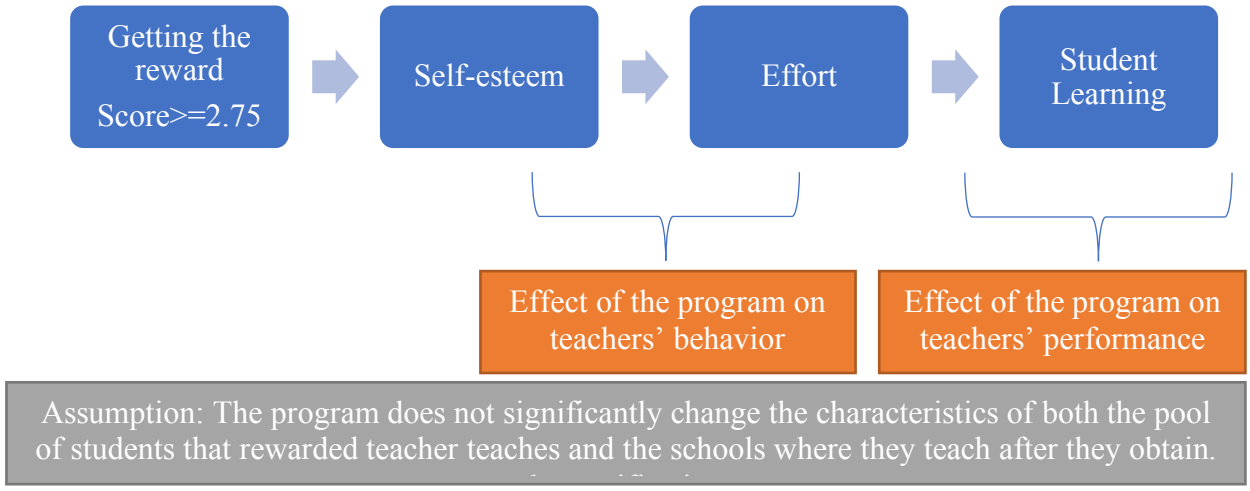
The MM and Interpretation of Teacher's Performance

As explained in Chapter 4, the MM helps me understand the underlying mechanism by which the program may have affected teachers' behavior, thereby affecting students' test scores. Student outcomes represent a direct measure of student learning, which in turn might be a credible proxy of teachers' performance. Nonetheless to argue that student outcomes are in fact a measure of teacher's performance, one important assumption of MM must hold: everything else remains unaffected by the program.

Figure 9.1 shows that, according to the MM, the program could have affected self-esteem and thus effort and test scores. However, one important assumption is implicit. If the program

changed something else that also affects test scores, interpreting a discontinuity in test scores around the cutoff as a change in teacher performance might be misleading. For instance, rewarded teachers might have been assigned high-achievement students, resulting in an upward bias for the effect of the program on student outcomes. Attributing this effect to an enhanced teachers' performance is a misinterpretation of the results. Then, it is necessary to test the MM's implicit assumption. In practice, this assumption is empirically tested by checking if the pool of students assigned to the teachers changed due to the reward. This hypothesis is tested by studying if there are differences in the variables: the income quantile of students assigned to teachers, the schooling level of students' mothers, the number of books students had at home, the school socioeconomic status, and the public/private status of the school where teachers taught.

Figure 9.1. The Motivational Model and its assumptions.



Summing up, the test scores can be interpreted as students learning, which, in turn, might be interpreted as a measure of teacher performance. Additionally, in order to facilitate its

interpretation, scores are re-standardized using this information: SIMCE Scores have been standardized on a scale that has a mean of 250 points and standard deviation of 50 points, allowing comparison of student performance across years.

9.2. Teachers' Behavior

To test the MM theory, as shown in Figure 9.1, the outcomes of interest are related to teachers' behavior, which can be called intermediate outcomes. One of them is the self-declared level of self-esteem. In practice, teacher's self-confidence is approximated by an index that measures teacher self-reported preparedness to teach several topics. There is a question that addresses this measure: *"Considering your preparation and experience in curriculum and teaching practices, how prepared do you feel to teach the following contents of subject in grade?"* The potential answers are: not prepared, prepared, or very prepared. The self-confidence index constructed is equal to the percentage of topics teacher j feels "very well prepared to teach" by the time of SIMCE in $(t + l)$.

Another intermediate outcome of interest related to teachers' behavior is effort, which is measured as an index of the number of contents covered during the school year. This measure can be interpreted as the learning opportunities given to students (Rowan et al., 2002), which depends on the effort made by the teacher. Specifically, it is measured using a question related to the coverage level of a content list that students should have learned during the year according the mandatory Chilean curriculum. Each teacher whose students would take the SIMCE test were asked about this. Specifically, the question is: *"Given that class time is limited, and it is likely you could not address all curricular content, to what extent could you teach the following contents of subject in grade?"* The possible answers are: totally taught, considerably taught,

some taught, and not taught at all. Effort then is represented by an index going from 0 to 1, calculated as the percentage of topics teacher *j* has "completely'" taught by the time of SIMCE.

It is expected that a teacher who exerts a higher level of effort would cover a higher amount of content. This is likely for at least four reasons. First, the school community with its owner, manager, principal and teachers have the incentive to effectively cover the math and reading curriculum as much as possible, since their progress is assessed by high-stakes standardized testing, SIMCE. These results obtained by each school are used by the Chilean Ministry of Education (MINEDUC) in order to inform parents' decisions in the context of a full choice school system, and to allocate public resources. This means that SIMCE test scores have the potential to affect the enrollment and school resources. Schools' income can change as a result of the effect of SIMCE test scores on parents' enrollment decisions. In addition, the schools' financial status is affected by SIMCE test scores because they are used to distribute resources. For example, the National System of Performance Evaluation of Public-Funded School[27] (SNED for its initials in Spanish) is public program that rewards schools' outstanding performance giving them additional financing. Second, it is likely that teachers were also focusing their effort on "teaching to the test"—i.e., focusing on the material tested—suggested by many research papers (David 2011, Jones et al. 2003, Russell et al. 2009, Polesel et al. 2014, Reay and Wiliam 1999, Rentner et al. 2006 cited in Ashadi and Rice (2016)). Third, following the Curricular Basis[28] is mandatory in Chile (Ministry of Education, 2009). This official document was

[27] Sistema Nacional de Evaluación del Desempeño.

[28] Bases Curriculares.

produced by the MINEDUC, as stated by the 2009 General Education Law. The Curricular Basis indicates what all the students must learn during their school trajectory (Ministry of Education, 2012b). In consequence, teachers must put their effort into covering the contents defined in the survey question. Fourth, evidence shows that content coverage can also be an important predictor of student achievement (Rowan et al., 2002). This is why teacher observation tools, such as the Protocol for Language Arts Teaching Observation (PLATO) developed by Pam Grossman and her colleagues, used in the context of teacher's performance assessment, include checklists for the major content domains required (Grossman et al., 2013).

Nevertheless, less content covered might not be interpreted unequivocally as less effort. One teacher could be superficially going through all of the material, while another teacher may focus on ensuring that their students are learning the material, even if that leads to fewer topics being covered. Taking this point into consideration, another measure of effort is used to estimate the effect of the program on teachers' behavior: class preparation. Data about this can be obtained because math and reading teachers are asked: *"In this school, in case you spend time preparing your classes, how many hours do you allocate in to it?"*. The answer is a continuous variable, which is truncated at eleven hours for the 2010 survey. To address this truncation issue, the variable "Class Preparation>5" was calculated, which takes the value of 1 if the teacher stated spending five or more hours every week to prepare for her classes. This indicator could be a proxy of effort, assuming spending more time in preparation reduces the time that teaches can spend in other kind of activities that may bring a higher level of satisfaction. Note that in Chile teachers are paid for very few hours outside the classroom. OECD (2017) states that: *in spite of large class sizes and student-teacher ratios, teachers in Chile work more hours in other OECD*

countries. Their statutory working time is 2,015 hours per year at the pre-primary to upper secondary levels, the highest among OECD countries with available data. The time spent on teaching is also high, 1,157 hours per year from pre-primary to upper secondary levels. (p. 5). In fact, until 2015, in Chile by law the ratio of paid teaching to non-teaching hours was 75/25 (Ministry of Education, 1997).

Two additional measures of class preparation help to illuminate if there is any difference in the program effect as a function of the level of hours spent on it. The number of hours spent in class preparation is expected to have an important role in terms of defining the likelihood of the marginal allocation of extra hours into class preparation. One may suggest that the class preparation is an input in the production function of learning outcomes, in which at some point the law of diminishing returns starts to operate. Increasing class preparation, while holding all others constant (i.e., ceteris paribus), would some point yield lower incremental per-unit return. An optimal increase in the time allocated on this task becomes less likely for teachers who already spend a large number of hours a week preparing classes than for teachers spending fewer. For this reason, two additional variables were calculated: "Class Preparation>2" and "Class Preparation>7."

9.3. Covariates

Students', teachers' and schools' characteristics are used as described in Chapter 8 to run specification checks. The selection of variables is based on their correlation with student achievement.

Based on the parents' survey given along with the SIMCE, three students' characteristics are considered. The variables associated take the value of one if:

- the monthly family income belongs to the highest income quintile in the student sample. Parents are asked to report where in fifteen categories their family income can be classified,
- the mother has more than 12 years of education, which in Chile means that she has more than high school education, and
- there are more than 50 books at home. Parents are asked to classify the number of books at home based on five categories.

Based on both administrative data and the teacher's survey given with the SIMCE, eight teachers' characteristics are considered when running specification checks. Six variables are dichotomic and take the value of one if teacher is female, has a graduate degree, is tenured, works in the Santiago, Chile's capital city, works in a rural area, and works in a public funded and administrated school. From this group of variables, covariates are chosen to estimate the SRD models. Doing so has the potential of increasing the efficiency of the SRD point estimator.

Based on administrative data provided by SIMCE dataset, there are two school characteristics variables. They take the value of one if:

- The school's socioeconomic status (SES) is either low or medium. This SES index classifies schools into five groups and is calculated by the MINEDUC every year. It is used to report SIMCE test scores in a more equitable way, comparing schools of similar SES (Ministry of Education, 2013). Four variables are used to calculate this SES index by a cluster analysis: mother's schooling, father's schooling, monthly family income, and a vulnerability indicator. All except for the last are parental self-report. The vulnerability indicator is an index calculated by the MINEDUC to target need-based school programs. The following Table shows the average value of each variable for the five SES groups for SIMCE 2012.

Table 9.1. Description of school SES indicator calculated by MINEDUC in 2012.

School SES	Average Number of School Years (Mother)	Average Number of School Years (Father)	Monthly Family Income (2012 US Dollars)	% of vulnerable children in the School
Low	8 years (1)	8 years (1)	315 US$ (126)	88% (9)
Medium-Low	10 years (1)	10 years (1)	464 US$ (122)	71% (8)
Medium	12 years (1)	12 years (1)	709 US$ (170)	49% (9)
Medium-High	14 years (1)	14 years (1)	1,454 US$ (433)	24% (9)
High	16 years (1)	17 years (1)	3,730 US$ (673)	1% (4)

Note: Standard Deviations in parentheses. Source is Ministry of Education (2013).

- The school is public funded and administrated. These schools are administrated by the municipalities and get funding from the central government. They receive a fixed amount of money for each enrolled student in the school. In 2012, according to official statistics, 39% of all students enrolled in the education system attended a public funded and administrated school, while 53% attended a public-funded and private-administrated school (Ministry of Education, 2012a).

Chapter 10 - Results

In this chapter I answer the research questions, following the empirical strategy discussed in Chapter 8. First, I confirm that the data shows a discontinuity in the probability of treatment around the cutoff. Then I show the specification checks. Next, is an analysis of the potential methodological problems described in Section 8.3, which do not seem to be threatening the causal interpretation of the results, although they affect the interpretation of the parameters. Lastly, I present the estimates for the local average treatment effect (LATE) of the program on the final and intermediate outcomes.

10.1. Discontinuity in the treatment probability

As the very first step in the analysis, I check if the data shows a discrete and deterministic change in the probability of receiving the AEP certification at the cutoff point. Figure 10.1 shows that all applicants below the cutoff point are not awarded, while individuals above the cutoff receive the certification—with either only the financial component or both parts. This result indicates that a Sharp Regression Discontinuity (SRD) design can be used. Next, I carry out three sets of specification checks as suggested by Imbens and Wooldridge (2009) and described in Section 8.2.

10.2. Specification checks

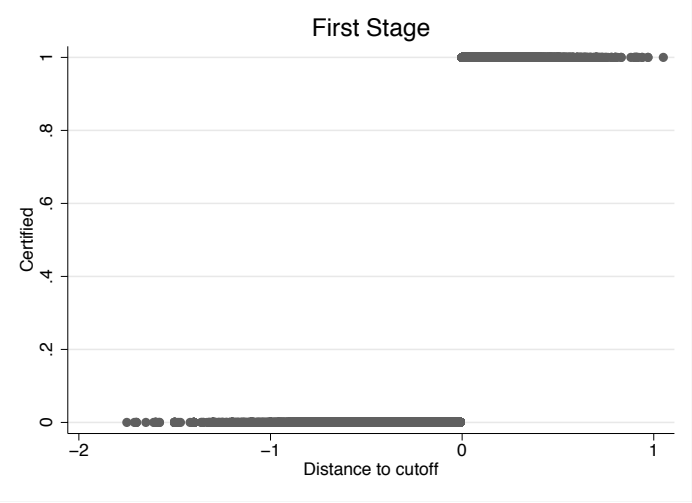
Testing manipulation of the running variable

An underlying assumption that gives validity to the regression discontinuity (RD) design is that individuals do not have the ability to precisely manipulate the score that they receive, so the number of treated observations just above the cutoff should be approximately similar to the

number of control observations below it. Although this assumption is neither necessary nor sufficient for the validity of an RD design, as explained by Skovron and Titiunik (2015), RD designs where there is an unexplained abrupt change in the number of observations right at the cutoff tend to be less credible.

There is qualitative information regarding teachers' limited possibility to manipulate their scores. Crucially, while teachers knew about the evaluation process from the program's website and the law,[29] they do not know about the evaluation rubric nor the cutoff, or how the scores are processed. In addition, there were no institutionalized mechanisms for appeal and teachers were never told their final scores.

Figure 10.1. Evidence of a SRD.



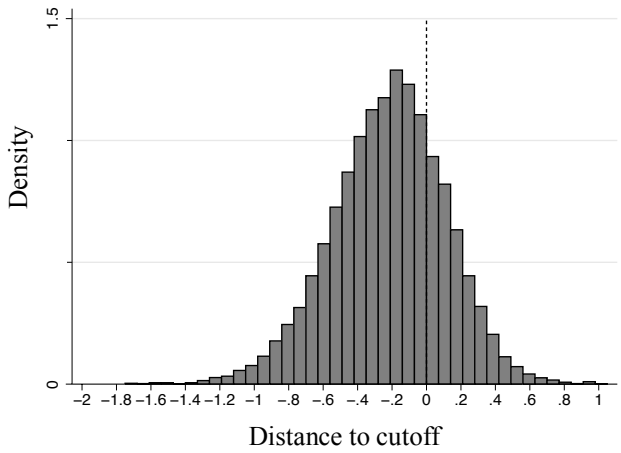
Note: this figure shows a scatter plot of the probability of obtaining the certification at each score distance to the cutoff for teachers' first application. This plot does not include re-takers or cheaters. Total observation number (14,626) equals the number of total applications in 2002-2011 (16,468) minus cheaters (89) and re-takers (1,753).

[29] http://www.aep.mineduc.cl/?numeroPag=3 and https://www.leychile.cl/Navegar?idNorma=1039324

There is also empirical evidence to rule out the possibility of teachers having manipulated their scores. First, I look at raw data with the number of observations above and below the cutoff.

Figure 10.2 shows histogram of the running variable for teachers who were applying for first time. Visually, the running variable does not seem to have a discontinuity around the cutoff.

Figure 10.2. Histogram of running variable.

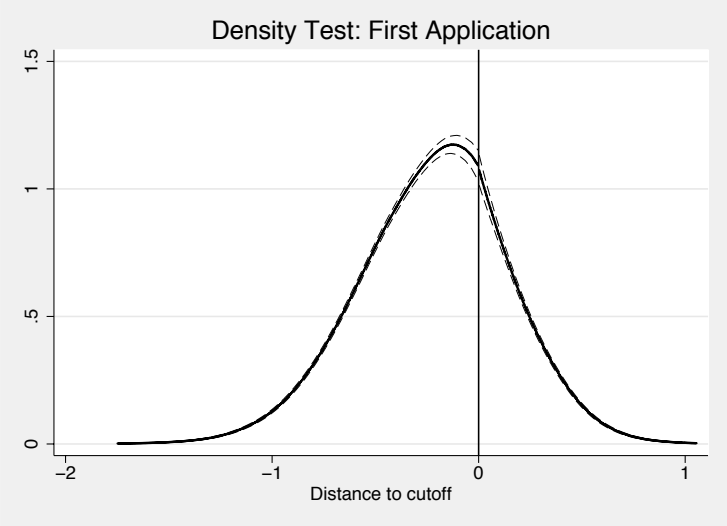


Note: the histogram shows the density of the score distance to the program's cutoff. Forty bins are plotted. Only data for first-takers and non-cheaters is included. The dashed line indicates the cutoff or zero distance to it. Data for 2003 to 2011 is included.

In addition to a graphical illustration of the density of the running variable, I perform a density test of the running variable by calculating the McCrary test for discontinuity in the AEP score. The null hypothesis of the test is that the density of the running variable is continuous at the cutoff, and its implementation requires the estimation of the density of observations near the cutoff separately for observations above and below the cutoff. The coefficient estimated by the McCrary test is -0.007 with a standard deviation of 0.04, which implies a t-statistic of -0.175. Therefore, there is no evidence of discontinuity in the running variable. The Figure on the

following page shows the density pooling of the data for all available years. The results of the McCrary test suggest that was no manipulation of AEP scores.

Figure 10.3. McCrary test.



***Note*:** the figure shows the plot calculated by the DCdensity STATA command for the score distance to the program's cutoff. Only data for first-takers and non-cheaters for 2003-2011 is included.

Testing treatment effect on predetermined covariates and placebo outcomes

In order to further check for the internal validity of the identification strategy, the treatment effect on predetermined covariates (balance check) and placebo outcomes (falsification test) are calculated. These tests are important because finding a discontinuity in these predetermined covariates might cast doubt on the assumptions underlying the RD design. As explained by Skovron and Titiunik (2015), if teachers lacked the ability to manipulate their scores, there should not be any systematic differences among teachers with similar scores. Thus, except for their treatment status, teachers just above and just below the cutoff should be similar in all those characteristics that could not have been affected by the treatment.

To approximate the regression functions, a local linear polynomial, without covariates within a Mean-Square-Error-(MSE)-optimal bandwidth for all years is used. To do so and for inference purposes, the CCT method (*rdrobust* STATA command) is used. When applying this nonparametric method, three-point estimators are obtained: the conventional RD estimator, the bias-corrected RD estimator that is optimal for point estimation since optimizes the bias-variance trade-off by a MSE-optimal bandwidth, and the robust RD estimator that complementarily presents robust-bias corrected confidence intervals for valid inference. When estimating the RD effect, tables show the conventional RD estimator, the robust-bias corrected confidence intervals, robust standard errors, and robust p-values.

As a robustness exercise, the results of balance checks and falsification tests using robust specifications with covariates are presented in Appendix G. In general, these tables show that the results of the specification checks are robust to the inclusion of covariates. In addition, Appendix H shows the graphical analysis for the specification checks.

Balance checks

A balance check is performed on predetermined covariates jointly for language and math teachers whose students took the national standardized SIMCE test[30] (SIMCE for its initials in Spanish) and have SIMCE test score data. This is done without distinguishing if the program had either only the financial component or both components, since at the time of the application teachers did not know whether the program included a pin or not. The covariates chosen to

[30] Sistema Nacional de Medición de la Calidad de la Educación.

perform the balance check are related to teacher, student, and school characteristics. In general, balance checks show no imbalance. One of seventeen characteristics are significantly different at the 5% confidence level between certified and non-certified teachers in the application year. The imbalanced covariate is a teacher characteristic related to having graduate education.

Panel A.1 in Table 10.1 shows the balance check on teachers' characteristics for the application period (t). The variables examined are: experience, gender, age, having a specialized degree, having tenure, teaching in the metropolitan region, teaching in a rural school, and the average number of schools and classrooms where the teacher teaches, whether the teacher moved to a new school or left a school where she was teaching before the application to the program, and whether the teacher was not teaching in the year previous to the application. Overall, the table shows that teachers below and above the threshold are similar in almost all of the variables examined, except for gender. Female teachers, in the application year, were 20 percentage points less likely to obtain the certification in t + 1. When covariates are included, gender differences are still significant, as shown in Appendix G.

As suggested in Section 8.4 of Chapter 8, the results of the balance check on covariates also help test the assumptions needed to include pre-treatment covariates, which are aimed to increase the efficiency of the SRD estimator. As shown in Appendix G, using this variable does result in smaller confidence intervals for most of the specifications, however the gains are small. In fact, using age as covariate to estimate the program's effect on a set of placebo test scores in year *t* does not affect the standard error and hence there is no change in the length of the confidence interval. In consequence, the results including age as a covariate are presented as a complement in the appendixes.

Panels A.2 of Tables 10.1 shows the balance checks on the following student characteristics: being female, having a mother with more than 12 years of education, having more than 50 books at home, and being in the wealthiest quintile. The results indicate that there are not statistically significant differences in students with teachers below and above the cutoff, then there is no evidence of previous sorting on teachers above and below the threshold.

Differences in school characteristics where the teachers taught are also analyzed. The variables tested are the school socio-economic status and the probability of teaching in a public school. As shown in Panels A.3 of Table 10.1, there are not statistically significant differences in the schools where teachers below and above the cutoff taught. Overall, the results indicate there is a balance all but one of the observable characteristics for teachers below and above the cutoff.

Falsification tests

Panel B of Table 10.1 shows the estimates of treatment effect on placebo outcomes. The first exercise is to estimate the effect of the program on students' test scores during the application period. Table 10.1 shows not statistically significant differences in test scores for students of teachers who were just below and above the threshold. The second falsification exercise tests if there is some imbalance in teacher behavior. The behavior measures are: an index of the number of topics covered during the academic year, an index that measures teacher's self-perceived preparedness to teach several topics in class, and an indicator if teachers spend more than 5 hours a week preparing for classes. The result in Panel B of Table 10.1 shows that there are no significant differences in the number of topics covered or how well-prepared teachers felt, but on average, the probability of preparing classes for more than 5 hours a week was 16 percentage points larger for certified teachers. Since, only 2 out of 21 variables are statistically significant,

the result for preparing classes maybe due to chances. These findings are robust to including covariates, as shown in Table H.1.

10.3. Two potential methodological issues

As explained in Section 8.3, two methodological problems might threaten the identification strategy—sorting and attrition. Here, I test for the presence of both problems. The balance checks and falsification tests show no evidence of teacher sorting based on their characteristics or school characteristics across treatment and control groups. Complementarily, the results of the falsification tests can be interpreted as evidence of the random sorting of math and language teachers to students and schools during the year of application.

Regarding attrition, Table 10.2 shows that being awarded with the certification does not affect the probability of teaching a class that takes the SIMCE test in a window of one, two or three years after the reward was given. Similarly, from Figure 10.4 we do not observe a discrete change in the probability of taking the test after being rewarded. Finally, in the Appendix H and I present the figures for the balance check and the falsification test, showing again that samples below and above the threshold do not differ statistically. Taken altogether, one can assume that any difference in ex-post outcomes during the three years after the result of the AEP application process is known can be attributed to the reward. In other words, in the absence of specific assumptions about the global shape of the regression functions, the effect recovered by the SRD design would be LATE at the cutoff.

Table 10.1. Specification checks for teacher's, student's and school's characteristics in *t*.

SRD estimates using robust bias-corrected local linear polynomial regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	RD treatment effect	**Robust SE**	**Robust p-value**	h_{MSE}	**Robust 95% CI**	N_h	**N**
Panel A. Balance Variables							
Panel A.1 Teachers' characteristics							
Experience	-0.327	2.387	0.838	0.161	(-5.168, 4.19)	713	1,756
Female	-0.208	0.105	*0.019*	0.093	(-0.452, -0.04)	411	1,758
Age	0.573	2.374	0.801	0.15	(-4.054, 5.25)	624	1,756
Spec. Degree	0.056	0.048	0.248	0.174	(-0.039, 0.15)	750	1,758
Tenure	0.063	0.117	0.511	0.128	(-0.152, 0.306)	531	1,758
Metro. Region	0.076	0.097	0.387	0.189	(-0.106, 0.274)	783	1,758
Rural	-0.001	0.08	0.925	0.185	(-0.15, 0.165)	783	1,758
# Schools by Teacher	0.046	0.061	0.397	0.169	(-0.068, 0.171)	714	1,758
# Classrooms by Teacher	0.149	0.433	0.959	0.16	(-0.825, 0.87)	714	1,758
Teacher is a mover	-0.045	0.076	0.454	0.168	(-0.206, 0.092)	638	1,523
Teacher is a leaver	-0.001	0.028	0.884	0.202	(-0.05, 0.058)	770	1,523
Panel A.2 Student's Characteristics							
Female	0.046	0.056	0.384	0.144	(-0.061, 0.158)	23,360	64,548
Mother's Schooling>12	0.041	0.043	0.226	0.142	(-0.032, 0.137)	20,491	56,813
Books at home>50	0.034	0.03	0.205	0.181	(-0.021, 0.096)	26,435	58,221
Family Income Top Quintile	0.055	0.047	0.154	0.147	(-0.025, 0.159)	21,026	58,221
Panel A.3 School's Characteristics							
Public School	-0.056	0.112	0.555	0.18	(-0.285, 0.153)	1012	2,246
Low-Medium SES	-0.024	0.121	0.937	0.153	(-0.246, 0.227)	869	2,246
Panel B. Falsification Variables							
Topics Covered	-0.02	0.066	0.573	0.148	(-0.166, 0.092)	662	1,808
Well Prepared	0.104	0.084	0.229	0.151	(-0.064, 0.266)	713	1,800
Class Preparation>5	0.165	0.101	*0.043*	0.115	(0.006, 0.403)	470	1,601
Test Score	0.049	0.122	0.595	0.171	(-0.174, 0.303)	27,795	64,548

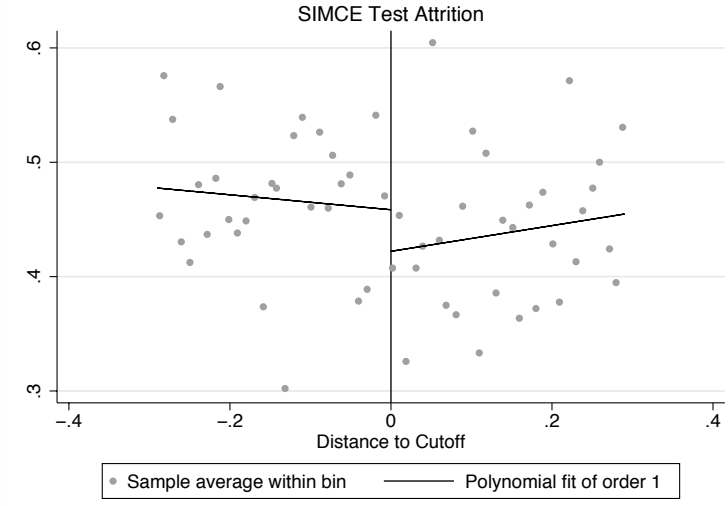
Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014); (iii) h corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (iv) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$; (v) standard errors (SE) are clustered at school level; (vi) no covariates are included; (vii) application year, subject and grade dummies are included; and (viii) the observation number (N) varies for several reasons. First, the variables in the table come from different datasets, which have different observations. For instance, test scores come from the SIMCE dataset, while teacher's characteristics come from administrative dataset. Second, variables coming from the same source have a different observation number because they were collected with different timings. For instance, the question regarding the number of hours spent preparing classes was part of the SIMCE teacher's survey in 2004, 2006 and 2010, while the question regarding topics covered was asked every year in the same survey. (ix) Robust SE are estimated.

Table 10.2. Attrition test in ($t + l$).

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}	(5) Robust 95% CI	(6) N_h	(7) N
Dependent Variables							
Probability of being assigned to a classroom that takes Math SIMCE in ($t + l$) with l=1,2,3	-0.042	0.042	0.287	0.209	(-0.128, 0.038)	3,315	7,489

Notes: (i) the reported RD treatment effects are estimated from a local linear regression between the probability of finding a SIMCE test scores in years t + 1, t + 2 or t + 3 for a teacher who teaches math or language in those years and apply for math or language certification in t and the distance to the cutoff; (ii) the reported coefficient is estimated with a single regression between the AEP score ($\tilde{S}_t$) and dummy variable that takes the value 1 if the teacher was assigned to a classroom that takes the SIMCE test in years t + 1, t + 2 or t + 3, if $\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h_{MSE}$(certified teachers within the MSE bandwidth estimated above) and $\tilde{S}_t - h_{MSE} \leq S_{jt} < \tilde{S}_t$ (non-certified teachers within the MSE bandwidth); (iii) no covariates are included; (iv) application year, subject and grade dummies are included; (v) standard errors (SE) are clustered at school level; (vi) robust SE are estimated; (vii) point estimators are constructed using local polynomial estimators with triangular kernel; (viii) robust p-values are constructed using bias correction with robust standard errors as derived in Calonico et al. (2014); (ix) h_{MSE} corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (x) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$.

Figure 10.4. RD plots for specification checks for SIMCE test attrition in $t + l$.



Notes: (i) data-driven RD plots using evenly spaced 25 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; (iv) all panels are based on administrative data for 2003-2011 applicant cohorts and restrict observations to individuals with scores within 0.4 points of the cutoff; and (v) calculations made at the teacher level.

10.4. Program effect on the outcomes of interest

Having tested if the identification strategy is valid, this section aims to answer the research questions described in Chapter 5. I start by testing the effect of the program, after that the certification was given, on final ex-post outcomes, including any potential fade-out process. Subsequently, I study the unbundled effects. The Motivational Model (MM) that might underlie the program's effect on test scores is also explored. Specifically, as suggested by Bénabou and Tirole (2005), the basic idea is to test if AEP and its components affect teachers' self-confidence and effort after the program has given the reward. In addition, the MM's assumption, as explained in Section 9.1, is tested in order to see if a change in test scores can be interpreted as a change in teachers' performance.

Before showing the estimated results, there are six issues that need to be made explicit. First, the results presented are estimated by pooling the data of math and language teachers. Second, the results presented used the robust specification described by Equation 8.6. This specification does not include covariates, and the standard errors are clustered at school level. Third, to show further evidence regarding the unbundled program effect, the results obtained by the alternative specification are presented. In this alternative specification, to test the differences by reward component, Equation 8.7 was used. Further, each table indicates the effect of the program when only the financial component is provided (α_1), the additional effect of the non-financial component, the pin component (β_1), and the full program effect when the financial incentives and a ceremony were given ($\alpha_1 + \beta_1$). For the full program effect, also presented is the p-value for the test ($\alpha_1 + \beta_1 = 0$). Fourth, notice that, as mentioned in Chapter 8, the bandwidth is chosen in a data-driven manner, taking into consideration the MSE-optimal bandwidths automatically

calculated by the robust specification developed by Calonico et al. (2016a). Fifth, the graphical analysis of the outcomes around the cutoff are presented. Sixth, and lastly, as a way of testing the results' robustness, Appendix I show the robustness exercise as explained in Section 8.4. Specifically, this appendix shows the estimates that are obtained by using the robust specification with covariates (as in Equation 8.8) and the alternative specification with covariates described by Equation 8.9.

The full and unbundled program effect on final outcomes

Did AEP and its components affect ex-post test scores? The answer is found by using the SIMCE test scores of each student *i* of the AEP teacher *j* in $t + l$ for years t 2003-2011. The test scores are normalized for each year/grade using information for all students. The data has been pooled for all grades where test scores are available—2nd, 4th, 8th, and 10th grades.

Question #1, does AEP affect ex-post teachers' performance? Panel A.1 of Tables 10.3 and 10.4 and plots in Figure 10.5 show that the program effect is positive, with a coefficient of 0.086. This suggests that giving a reward could increase subsequent student achievement by 0.09 standard deviations, but it is not statistically different from 0, which is robust to the alternative specifications (see Tables I.1 and I.4). These results are consistent with the evidence on the impact of NBPTS on student learning. The literature has found not statistically significant or small ex-post effects, that vary by subject and grade (Cowan and Goldhaber, 2016, What Works Clearinghouse, 2018).

Question #2, does the ex-post effect of AEP fade-out? Tables 10.3, 10.4, I.1, and I.4 and Figure 10.5 show that the estimate of the full program effect increases over time, from 0.07 to 0.10. However, the effect is not statistically significant. Then the program does not affect ex-post

test scores either in the following three years or in one, two or three years after that certification was given.

Question #3, does each component of AEP affect ex-post teachers' performance? In terms of the final outcomes, Panel A.2 of Tables 10.3, and I.1, Table I.4 and plots of Figure 10.6 show the effect of each component of the program on test scores. In contrast to the full program results, the estimate of the financial component is negative and equal to -0.021 standard deviations, though it becomes large and positive three years after the certification. However, the effects are not statistically significant. In Table10.4, Panel C.3, I present the effects for the non-financial component of the program, finding in general a positive but not significant effect, which becomes negative and non-significant three years after the reward was given.

These results are robust to alternative specifications and the inclusion of covariates. I tried two different cutoffs, the first one being 0.25 points below the current cutoff and the second one being 0.2 above the current cutoff. Since there are more teachers bellow the actual cutoff, the placebo cutoff is larger for that case. To further support the validity of the results found for final ex-post outcomes, an additional falsification test is run where the cutoff score was changed. As expected, a not statistically significant effect of the false program on student achievements is found (see Panels B and C in Table 10.3).

Table 10.3. Final program effect on standardized test scores in $t + l$. SRD estimates using robust bias-corrected local linear regression.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}	(5) Robust 95% CI	(6) N_h	(7) N
Panel A. Final Outcomes							
Panel A.1 FULL PROGRAM							
Test Score if l=1,2,3	0.084	0.086	0.216	0.218	(-0.062, 0.275)	66,779	145,247
Test Score if l=1	0.067	0.126	0.48	0.192	(-0.158, 0.337)	23,784	54,713
Test Score if l=2	0.097	0.138	0.395	0.2	(-0.153, 0.388)	23,057	52,139
Test Score if l=3	0.106	0.15	0.383	0.227	(-0.163, 0.424)	17,367	38,395
Panel A.2 FINANCIAL COMPONENT							
Test Score if l=1,2,3	-0.021	0.182	0.945	0.137	(-0.37, 0.344)	17,296	51,061
Test Score if l=1	-0.068	0.214	0.688	0.121	(-0.504, 0.333)	5,475	17,021
Test Score if l=2	-0.108	0.233	0.712	0.115	(-0.542, 0.37)	5,242	18,958
Test Score if l=3	0.379	0.254	0.088	0.122	(-0.064, 0.93)	4,524	15,082
Panel B. Placebo cutoff value: -0.2							
Panel B.1 FULL PROGRAM							
Test Scores if l=1,2,3	0.002	0.166	0.913	0.062	(-0.307, 0.343)	13,696	50,452
Panel B.2 FINANCIAL COMPONENT							
Test Scores if l=1,2,3	0.243	0.256	0.235	0.030	(-0.198, 0.804)	1,495	14,335
Panel C. Placebo cutoff value: +0.25							
Panel C.1 FULL PROGRAM							
Test Scores if l=1,2,3	0.011	0.159	0.918	0.062	(-0.295, 0.328)	16,720	80,404
FINANCIAL COMPONENT							
Test Scores if l=1,2,3	0.088	0.229	0.546	0.055	(-0.311, 0.588)	6,838	31,695

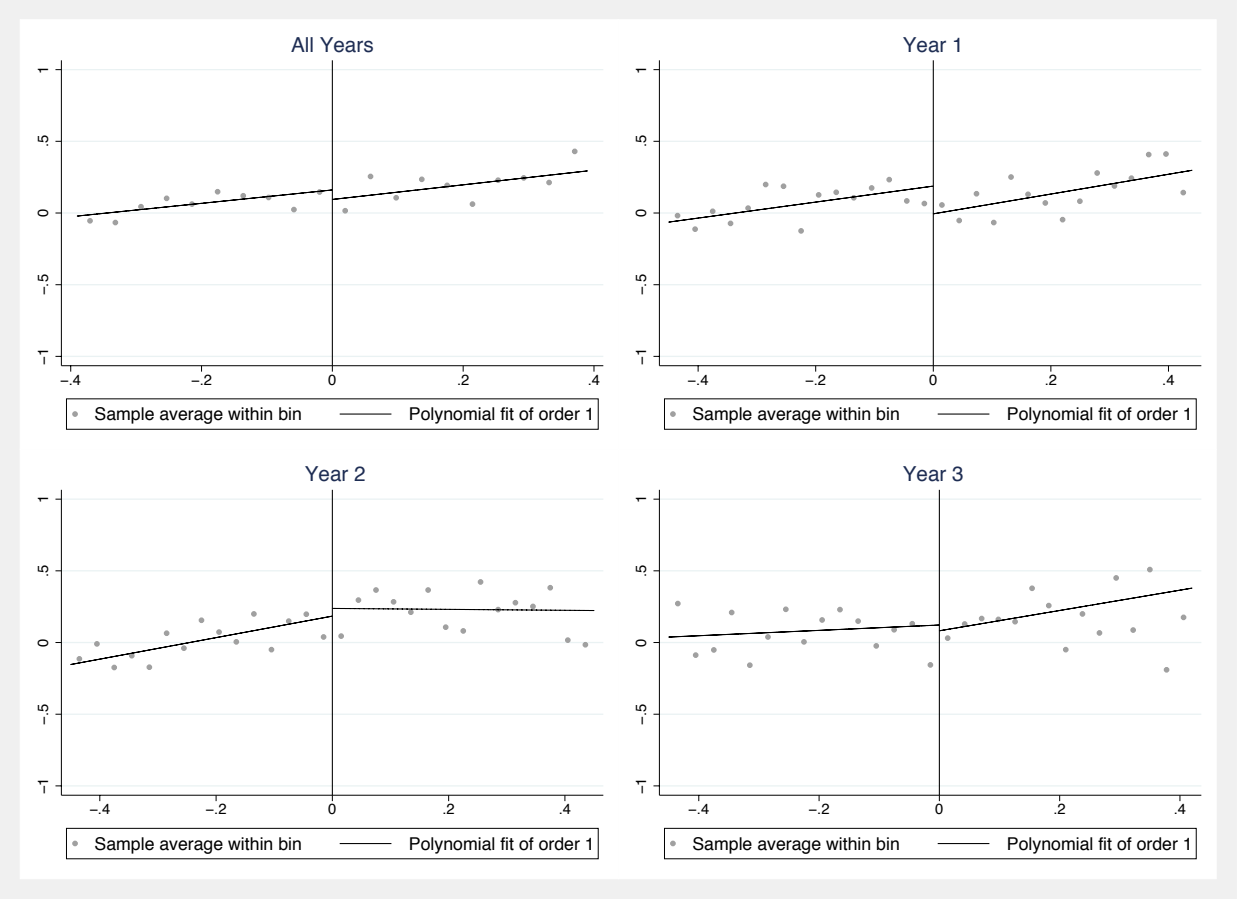
Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014) ; (iii) h_{MSE} corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (iv) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$; (v) standard errors (SE) are clustered at school level; (vi) teachers' age is included as covariate; (vii) application year, subject and grade dummies are included; (viii) the observation number (N) varies for reasons such as data availability and changes in surveys; (ix) robust SE are estimated; (x) in Panel A each row reports coefficients from a single regression restricting the sample by the value of l =1, 2, 3. Then rows for test scores if l =1, 2 or 3 stand for the RD program effect on next-year, two-year-later, and three-year-later outcomes, respectively; (xi) test scores are normalized within each grade and year to have a mean of 0 and a standard deviation of 1; (xii) the estimates for placebo cutoffs restrict the data to only certified teachers when the fake cutoff is lower than the actual cutoff (-0.2), and only non-certified teachers in the other case (+0.25). This aims to avoid the contamination coming from the potential significant effect of the program on test scores; and (xiii) column eight shows the increase rate of the confidence interval length if the program effect is estimated with covariates versus no covariates.

Table 10.4. Final program effect on standardized math test scores in $t + l$. SRD estimates using alternative local linear regression.

	(1) **RD treatment effect**	(2) **Robust SE**	(3) **Robust p-value**	(4) h_{MSE}
Panel C. Final Outcomes				
Panel C.1 FULL PROGRAM				
Test Score if l=1,2,3	0.082	0.076	0.28	0.218
Test Score if l=1	0.069	0.111	0.534	0.192
Test Score if l=2	0.082	0.118	0.486	0.2
Test Score if l=3	0.105	0.13	0.42	0.227
Panel C.2 FINANCIAL COMPONENT				
Test Score if l=1,2,3	-0.021	0.157	0.896	0.137
Test Score if l=1	-0.077	0.192	0.689	0.121
Test Score if l=2	-0.111	0.203	0.585	0.115
Test Score if l=3	0.363	0.224	0.105	0.122
Panel C.3 NON-FINANCIAL COMPONENT				
Test Score if l=1,2,3	0.12	0.161	0.455	0.218
Test Score if l=1	0.124	0.204	0.542	0.192
Test Score if l=2	0.241	0.21	0.25	0.2
Test Score if l=3	-0.067	0.224	0.764	0.227

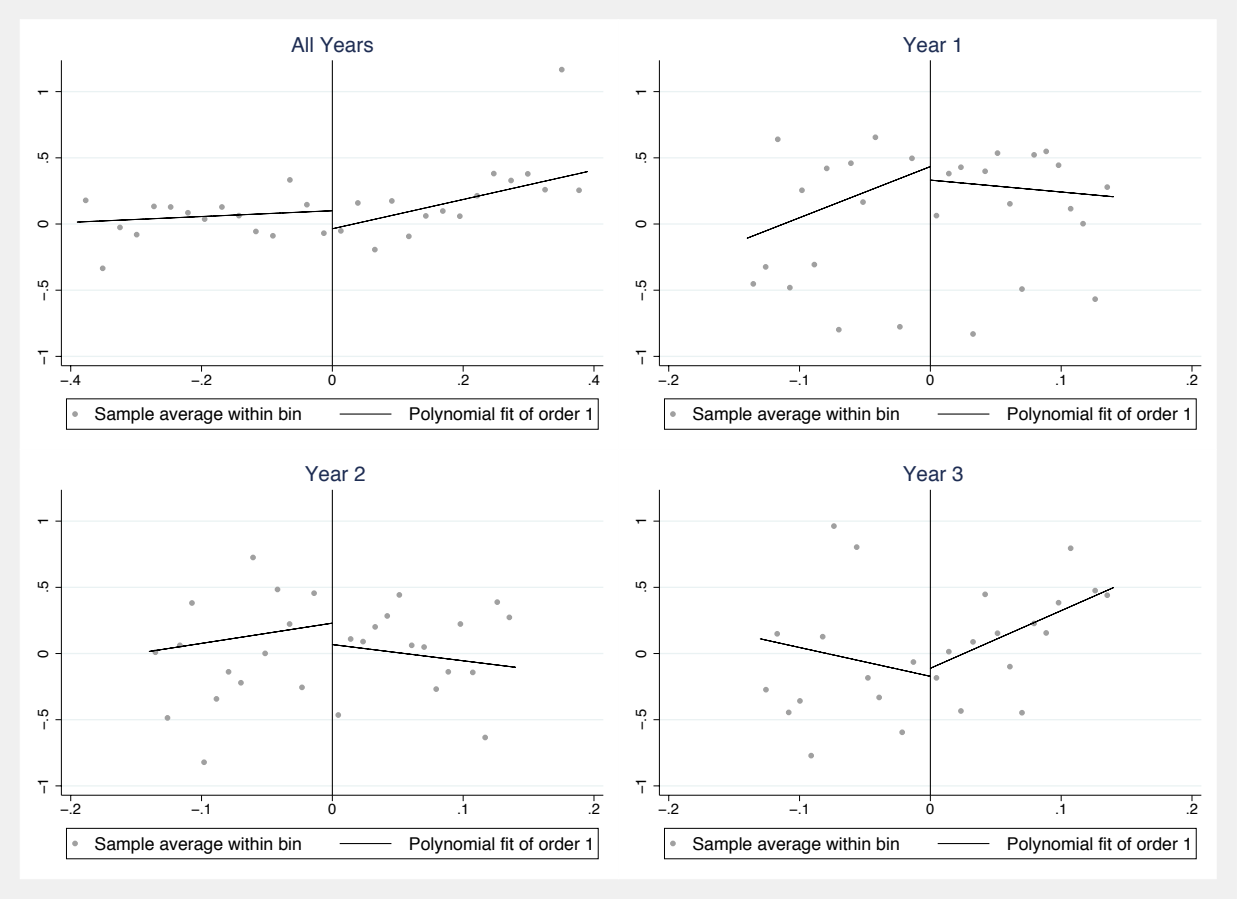
Notes: (i) all estimators are constructed using linear ordinary least-squares with robust standard errors (SE); (ii) h corresponds to the second generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b) that is recovered from the estimation of the program effect on the corresponding outcome using a robust bias-corrected local linear regression with triangular kernel; (iii) Robust SE clustered at school level are estimated; (vi) no covariates are included; and (iv) application year, subject and grade dummies are included.

Figure 10.5. RD plots for full program effect on test scores in $t + l$.



Notes: (i) data-driven RD plots using evenly spaced 15 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.4 points of a cutoff.

Figure 10.6. RD plots for financial component effect on test scores in $t + l$.



Notes: (i) data-driven RD plots using evenly spaced 15 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.2 points of a cutoff.

Underlying mechanism and intermediate outcomes

Tables 10.5 and 10.6 show the ex-post effect of receiving AEP on two intermediate outcomes (self-confidence and effort). In the case of effort, the results for two measures are presented: time spent for class preparation and rate of completely taught topics.

Question #4, does AEP affect ex-post teachers' behavior? There is no evidence of a significant treatment effect of AEP on teachers' practices and self-confidence after the reward is given. The average self-esteem for teachers above the threshold is positive but not significantly different than for teachers below the threshold regardless the specification used (see Figure 10.7, Panel A of Tables 10.5, and I.2 and Panel A.1 of Table I.5). In addition, certified teachers do not significantly change the rate of topics completely taught, or the time spent preparing classes.

Question #5, does each component of AEP affect ex-post teachers' behavior? Panel B of Table 10.5 and plots in Figure 10.8 show no effect on self-confidence for the financial component of the program. In addition, Panel A.3 in Table 10.6 also shows no change in self-confidence for the non-financial component. For the first measure of effort, I find a reduction of topics covered for the financial component of the program, and the estimate indicates that teachers that received the financial reward decreased the rate of topics covered by 14 percentage points after the reward was given, though significant at a 10% of significance level. On the other hand, the non-financial component tends to increase the rate of topics covered by 15 percentage points, also significant at a 10% of significance level. Since each component goes in a different direction, the final result of the program on the variable is closer to zero and not statistically significant.

When the effort is measured by class preparation, I find an increment in the probability of preparing classes for more than 7 hours a week, which increases almost 23 percentage points for teachers that received only the financial reward. This occurs only for the financial component of the program. No such changes are observed for preparing classes for more than 2 or 5 hours a week. The non-financial component of the program does not have a significant effect on effort. The estimate is large and negative, though not significant, for preparing classes using more than 7 hours a week. This negative effect explains the small and not significant effect of the program.

Table 10.5. Intermediate program effect on teachers' behavior in $t + l$. SRD estimates using robust bias-corrected local linear regression.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}	(5) Robust 95% CI	(6) N_h	(7) N
Behavioral Outcomes							
Panel A. FULL PROGRAM							
Well Prepared	0.052	0.05	0.347	0.234	(-0.051, 0.144)	1,717	3,477
Topics Covered	0.004	0.056	0.89	0.169	(-0.117, 0.102)	1,233	3,485
Class Preparation>2	0.021	0.084	0.656	0.18	(-0.128, 0.203)	807	2,147
Class Preparation>5	0.047	0.111	0.503	0.172	(-0.143, 0.291)	774	2,147
Class Preparation>7	0.01	0.096	0.738	0.201	(-0.156, 0.22)	888	2,147
Panel B. FINANCIAL COMPONENT							
Well Prepared	-0.071	0.103	0.375	0.126	(-0.293, 0.11)	369	1,229
Topics Covered	-0.145	0.092	0.08	0.138	(-0.342, 0.02)	420	1,230
Class Preparation>2	-0.038	0.064	0.552	0.156	(-0.163, 0.087)	558	1,440
Class Preparation>5	0.119	0.136	0.249	0.136	(-0.11, 0.425)	494	1,440
Class Preparation>7	0.225	0.126	0.039	0.162	(0.013, 0.505)	581	1,440

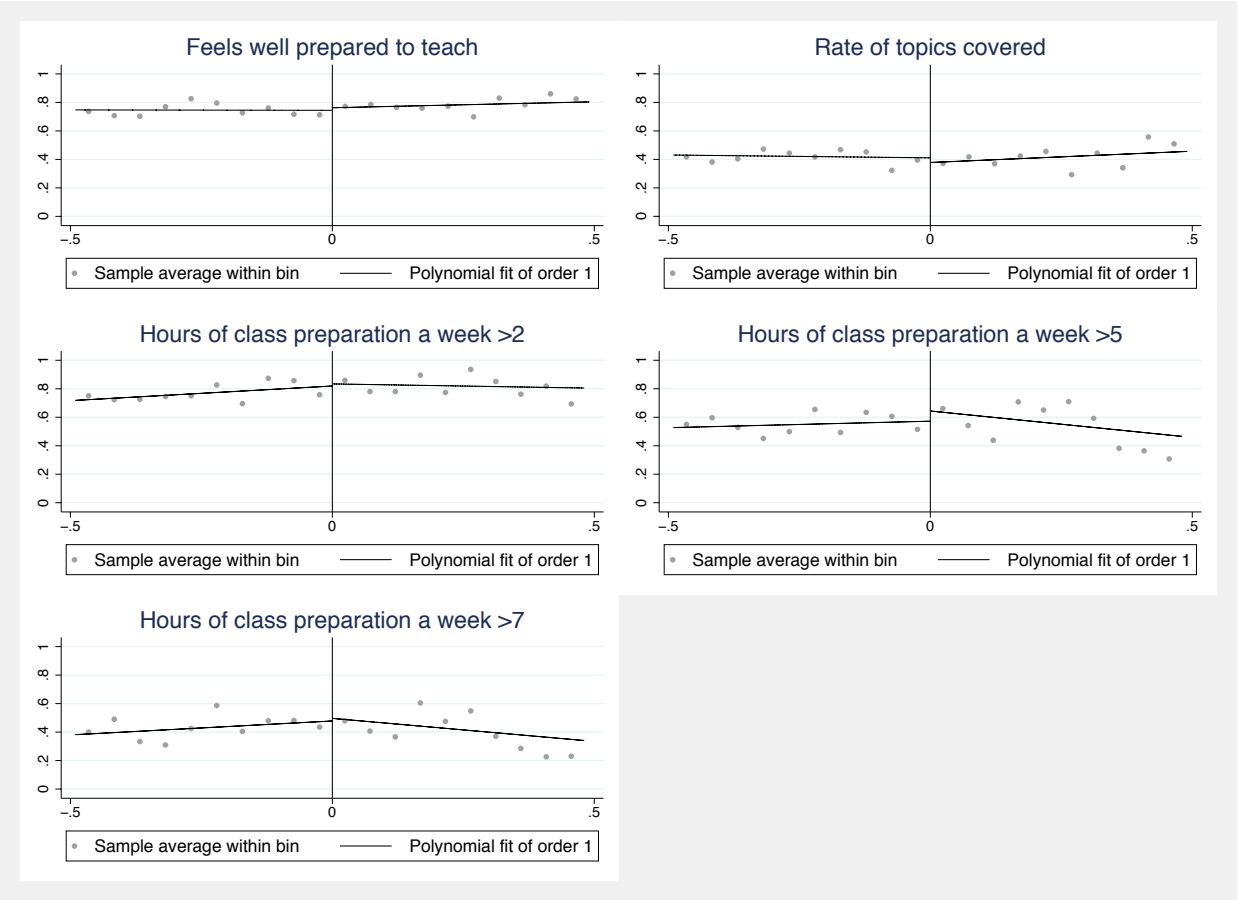
Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) robust p-values are constructed using bias-correction with robust standard errors (SE) as derived in Calonico et al. (2014); (iii) h corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (iv) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$; (v) SE are clustered at school level; (vi) no covariates are included; (vii) application year, subject and grade dummies are included; and (viii) the observation number (N) varies for several reasons. First, the variables in the table come from different datasets, which have different observations. Second, variables coming from the same source have a different observation number because they were collected with different timings. For instance, the question regarding the hours spent preparing classes was part of the SIMCE teacher's survey was asked in 2004, 2006 and 2010, while the question regarding topics covered was asked every year in the same survey. Finally (ix) robust SE are estimated.

Table 10.6. Intermediate program effect on teachers' behavior in $t + l$. SRD estimates using alternative local linear regression.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}
Panel A. Behavioral Outcomes				
Panel A.1 FULL PROGRAM				
Well Prepared	0.053	0.042	0.21	0.234
Topics Covered	0	0.047	0.997	0.169
Class Preparation>2	0.018	0.072	0.797	0.18
Class Preparation>5	0.049	0.094	0.603	0.172
Class Preparation>7	0.011	0.082	0.89	0.201
Panel A.2 FINANCIAL COMPONENT				
Well Prepared	-0.064	0.088	0.468	0.126
Topics Covered	-0.139	0.078	0.074	0.138
Class Preparation>2	-0.037	0.055	0.5	0.156
Class Preparation>5	0.119	0.123	0.335	0.136
Class Preparation>7	0.225	0.111	0.044	0.162
Panel A.3 NON-FINANCIAL COMPONENT				
Well Prepared	0.067	0.083	0.416	0.234
Topics Covered	0.151	0.084	0.073	0.169
Class Preparation>2	0.056	0.088	0.528	0.18
Class Preparation>5	0.011	0.144	0.939	0.172
Class Preparation>7	-0.168	0.13	0.196	0.201

Notes: (i) all estimators are constructed using linear ordinary least-squares with robust standard errors (SE); (ii) h_{MSE} corresponds to the second generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b) that is recovered from the estimation of the program effect on the corresponding outcome using a robust bias-corrected local linear regression with triangular kernel; (iii) Robust SE clustered at school level are estimated; (vi) no covariates are included; and (iv) application year, subject and grade dummies are included.

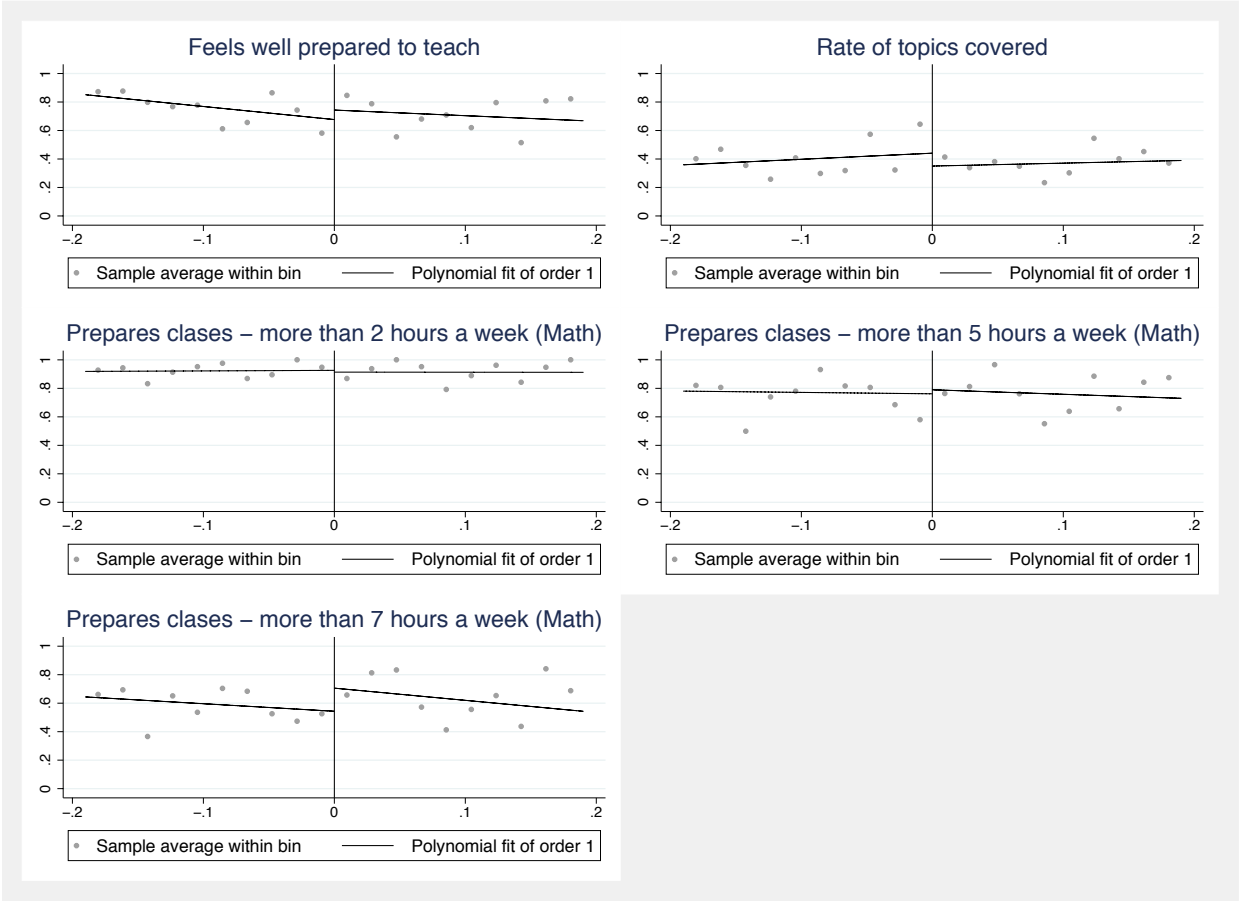
Figure 10.7. RD plots for intermediate full program effect on teachers' behavior in $t + l$.



Notes: (i) data-driven RD plots using evenly spaced 10 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.5 points of a cutoff.

Figure 10.8. RD plots for intermediate program effect on teachers' behavior in $t + l$.

Financial component.



Notes: (i) data-driven RD plots using evenly spaced 10 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.2 points of a cutoff.

Testing the assumption of the Motivational Model.

As explained in Section 9.1, it is also important to test if the assigned student pool changed due to the reward because it would affect the interpretation of the program's effect on test scores. This hypothesis is tested by studying if there are differences in the following: the income quantile of their students, their maternal schooling, the number of books they have at home, the school's SES, and the public/private status of their school. The results shown in Tables 10.7 and 10.8 and Figures 10.9, 10.10, 10.11, and 10.12 indicate that there is no sorting after the reward was given.

I also test if the pool of students assigned to the teachers changed for each component of the program. The results shown in Table 10.7, Panel B and Table 10.8, Panel B.3 indicate that teachers receiving either the full program or only the financial component are not assigned to a different set of students or schools compared to their non-rewarded counterparts. This evidence is robust the inclusion of covariates (see Tables I.3 and I.6).

Lastly, I test if teachers moved to a different school after receiving the award. Panels A.3 and B.3 in Table 10.7 and Panels B.1.3 and B.3.3 in Table 10.8 indicate that teachers do not move to or from a different school after the reward was given to them.

Table 10.7. Intermediate program effect on student, school, and teacher characteristics in $t + l$. SRD estimates using robust bias-corrected local linear regression.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}	(5) Robust 95% CI	(6) N_h	(7) N
Alternative Mechanism							
Panel A. FULL PROGRAM							
Female Student	0.002	0.031	0.872	0.283	(-0.056, 0.066)	83,966	145,247
Mother's Schooling>12	0.031	0.036	0.297	0.205	(-0.033, 0.108)	57,853	129,945
Books at home>50	0.014	0.024	0.482	0.211	(-0.03, 0.064)	60,792	130,856
Family Income Top Quintile	0.038	0.042	0.26	0.197	(-0.035, 0.131)	55,947	130,856
Panel A.2 School's Characteristics							
Public School	-0.059	0.085	0.526	0.221	(-0.22, 0.112)	2,059	4,298
Low-Medium SES	-0.084	0.085	0.236	0.244	(-0.268, 0.066)	2,230	4,299
Panel A.3 Teacher's Characteristics							
Teacher is a school-mover	0.009	0.06	0.836	0.227	(-0.106, 0.131)	1,476	2,951
Panel B. FINANCIAL COMPONENT							
Panel B.1 Student's Characteristics							
Female	0.032	0.039	0.3	0.114	(-0.036, 0.116)	14,588	51,061
Mother's Schooling>12	-0.057	0.073	0.354	0.13	(-0.211, 0.076)	13,457	43,760
Books at home>50	0.004	0.034	0.998	0.127	(-0.066, 0.066)	14,145	46,030
Family Income Top Quintile	-0.036	0.066	0.488	0.133	(-0.175, 0.084)	15,741	46,030
Panel B.2 School's Characteristics							
Public School	-0.117	0.151	0.439	0.181	(-0.412, 0.179)	814	1,772
Low-Medium SES	0.165	0.189	0.295	0.113	(-0.173, 0.569)	524	1,772
Panel B.3 Teacher's Characteristics							
Teacher is a mover	-0.054	0.103	0.471	0.156	(-0.276, 0.128)	540	1,360

Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014); (iii) h_{MSE} corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (iv) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$; (v) standard errors (SE) are clustered at school level; (vi) no covariates are included; (vii) application year, subject and grade dummies are included; and (viii) the observation number (N) varies for several reasons. First, the variables in the table come from different datasets, which have different observations. For instance, math test scores come from the SIMCE dataset, while teacher's characteristics come from administrative dataset. Second, variables coming from the same source have a different observation number because they were collected with different timings. For instance, the question regarding the number of hours spent preparing classes was part of the SIMCE teacher's survey was asked in 2004, 2006 and 2010, while the question regarding topics covered was asked every year in the same survey. Finally (ix) robust SE are estimated.

Table 10.8. Intermediate program effect on student, school, and teacher characteristics in $t + l$. SRD estimates using alternative local linear regression.

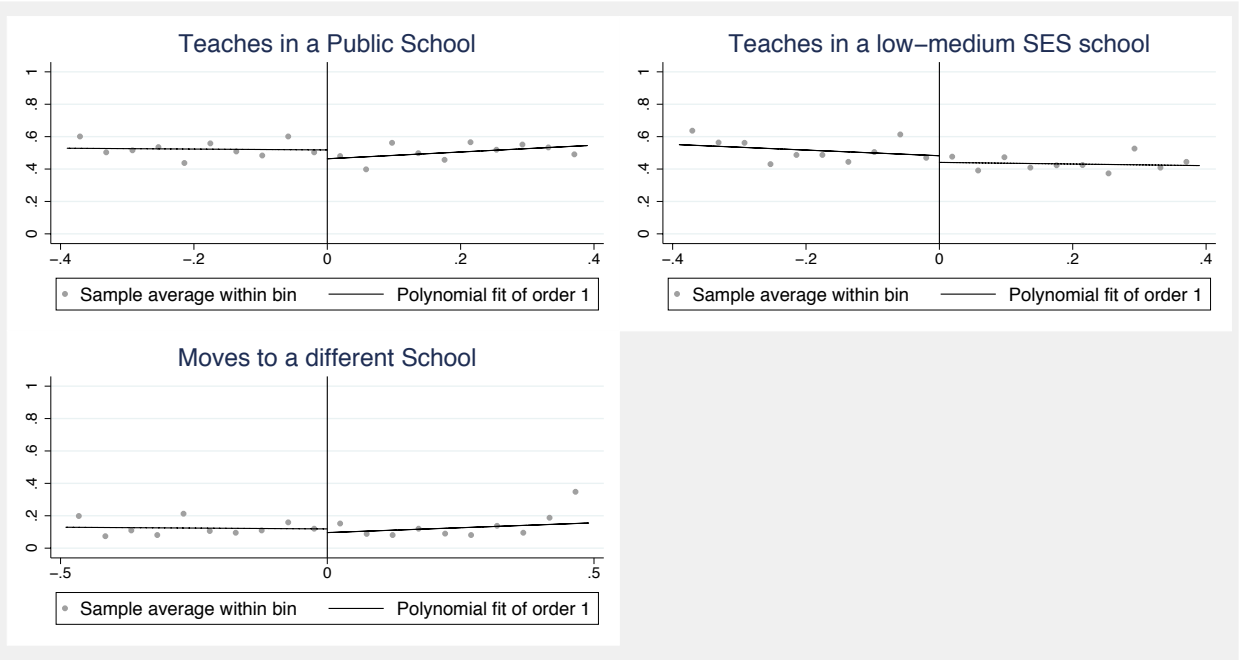
	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}
Panel B. Alternative Mechanism				
Panel B.1 FULL PROGRAM				
Panel B.1.1 Student's Characteristics				
Female Student	0.002	0.026	0.95	0.283
Mother's Schooling>12	0.028	0.029	0.337	0.205
Books at home>50	0.013	0.021	0.535	0.211
Family Income Top Quintile	0.035	0.035	0.323	0.197
Panel B.1.2 School's Characteristics				
Public School	-0.061	0.071	0.394	0.221
Low-Medium SES	-0.084	0.072	0.242	0.244
Panel B.1.3 Teacher's Characteristics				
Teacher is a school-mover	0.015	0.051	0.771	0.227
Panel B.2 FINANCIAL COMPONENT				
Panel B.2.1 Student's Characteristics				
Female	0.03	0.034	0.37	0.114
Mother's Schooling>12	-0.051	0.064	0.422	0.13
Books at home>50	-0.006	0.032	0.846	0.127
Family Income Top Quintile	-0.035	0.059	0.554	0.133
Panel B.2.2 School's Characteristics				
Public School	-0.109	0.131	0.404	0.181
Low-Medium SES	0.176	0.163	0.28	0.113
Panel B.2.3 Teacher's Characteristics				
Teacher is a school-mover	-0.051	0.088	0.563	0.156
Panel B.3 NON-FINANCIAL COMPONENT				
Panel B.3.1 Student's Characteristics				
Female	-0.022	0.037	0.547	0.283
Mother's Schooling>12	0.07	0.061	0.248	0.205
Books at home>50	0.008	0.033	0.802	0.211
Family Income Top Quintile	0.063	0.066	0.334	0.197
Panel B.3.2 School's Characteristics				
Public School	0.064	0.139	0.647	0.221
Low-Medium SES	-0.125	0.135	0.354	0.244
Panel B.3.3 Teacher's Characteristics				
Teacher is a school-mover	0.06	0.09	0.508	0.227

Notes: (i) all estimators are constructed using linear ordinary least-squares with robust standard errors (SE); (ii) h_{MSE} corresponds to the second generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b) that is recovered from the estimation of the program effect on the corresponding outcome using a robust bias-corrected local linear regression with triangular kernel; (iii) Robust SE clustered at school level are estimated; (vi) no covariates are included; and (iv) application year, subject and grade dummies are included.

Figure 10.9. RD plots for full program intermediate effect on student characteristics in $t + l$.

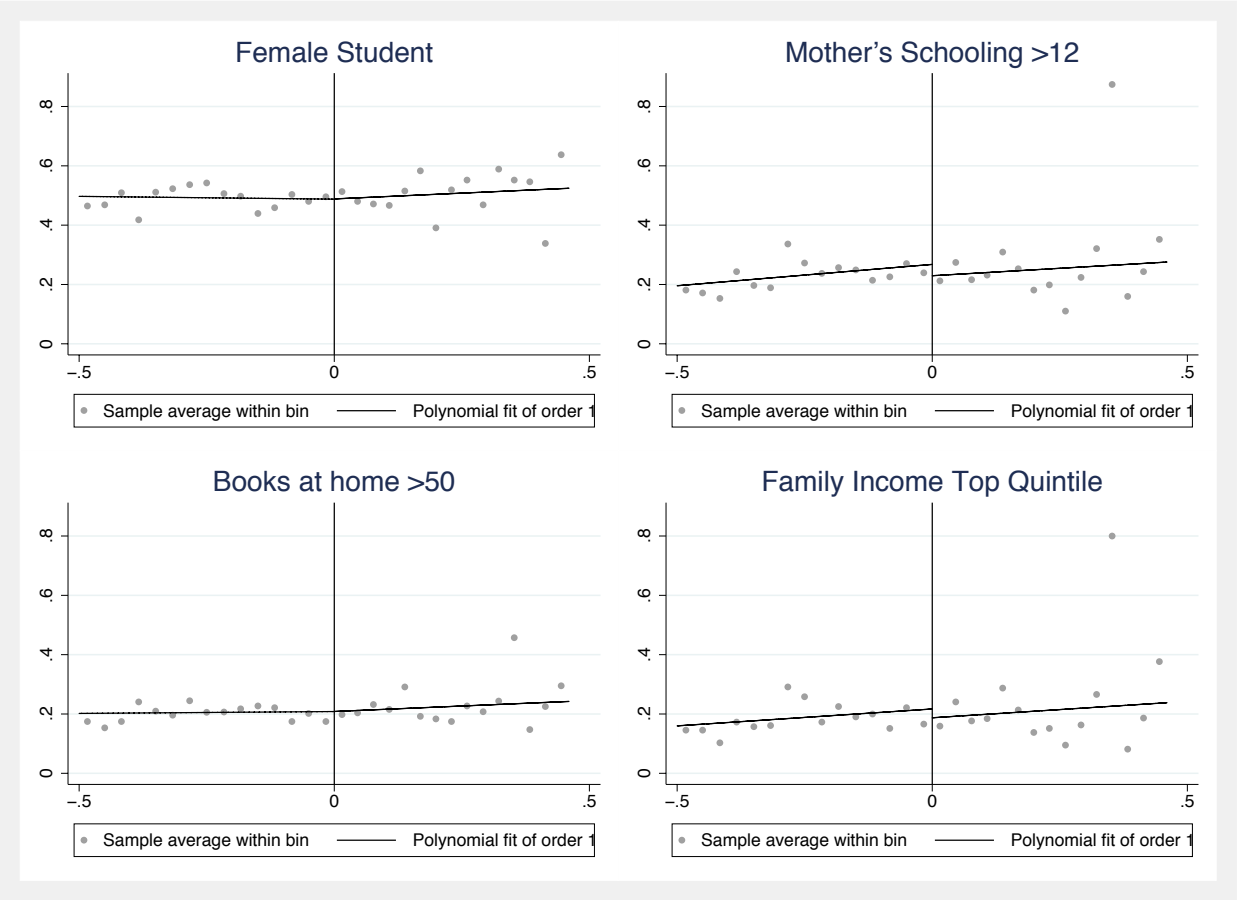
Notes: (i) data-driven RD plots using evenly spaced 10 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.4 points of a cutoff.

Figure 10.10. RD plots for full program intermediate effect on school characteristics in $t + l$.



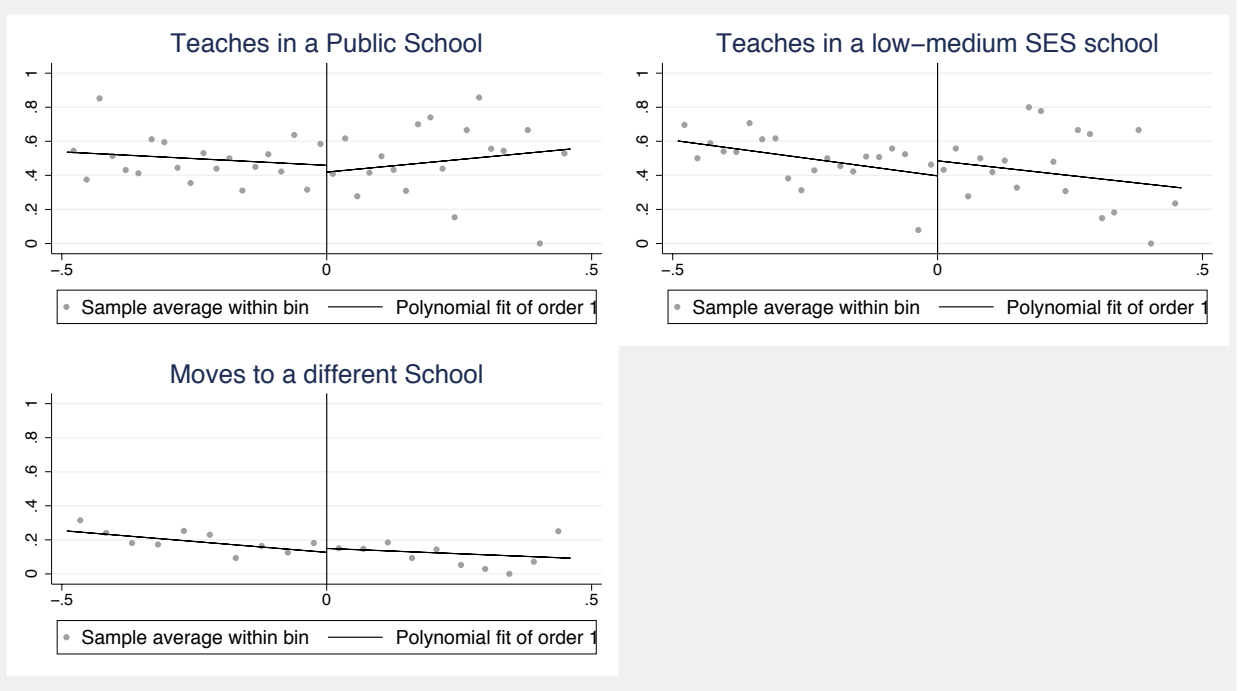
Notes: (i) data-driven RD plots using evenly spaced 10 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.4 points of a cutoff.

Figure 10.11. RD plots for intermediate effect on student characteristics in $t + l$. Financial component.



Notes: (i) data-driven RD plots using evenly spaced 15 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.5 points of a cutoff.

Figure 10.12. RD plots for intermediate effect on school characteristics in $t + l$. Financial component.



Notes: (i) data-driven RD plots using evenly spaced 20 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.5 points of a cutoff.

Sensitivity to window length

A sensitivity analysis of the results is performed and presented in Figures 10.13 and 10.14. I change the bandwidth and re-estimate the effects of the program; the figures show the p-value of the effect of the program of each bandwidth. Figure 10.13 suggests that the results for the full program are robust to the window length. There is no significant effect on final and intermediate behavioral ex-post outcomes, and school and student characteristics at any value of the windows under analysis. Only for topics covered at very small windows do. I find a significant effect of the program. For the financial component of the program (Figure 10.14), self-confidence and

effort measured by class preparation, are not statistically significant for any window analyzed. However, for a window between 0.15 and 0.19, there is a significant effect on the rate of topics completely covered by the teacher. For the remaining variables (class preparation, student gender, number of books at home, family income, school administration and socio-economic level of the school), the bandwidth choice neither affects the results nor their interpretation.

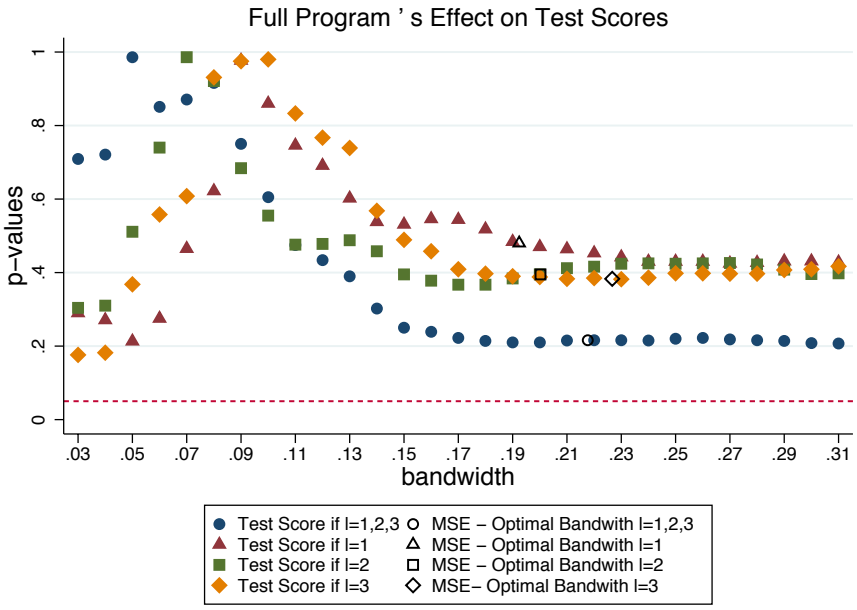
In this research, choosing a data-driven bandwidth has provided credible support for the program effect for at least two reasons. First, the automatically calculated bandwidth avoids arbitrary selection of the window length. Second, for this research, the program effect obtained by choosing the data-driven bandwidth resulted in a low sensitivity to the window selection.

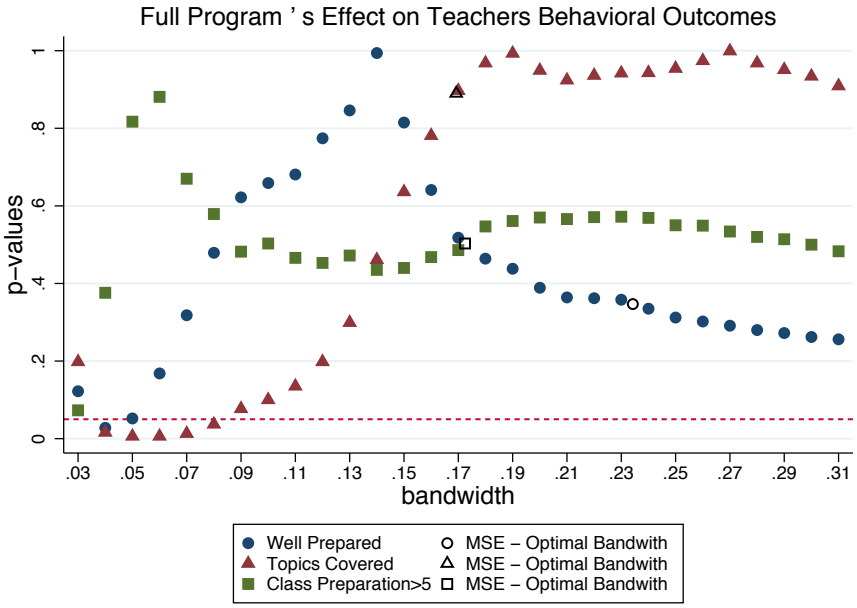
Results summary

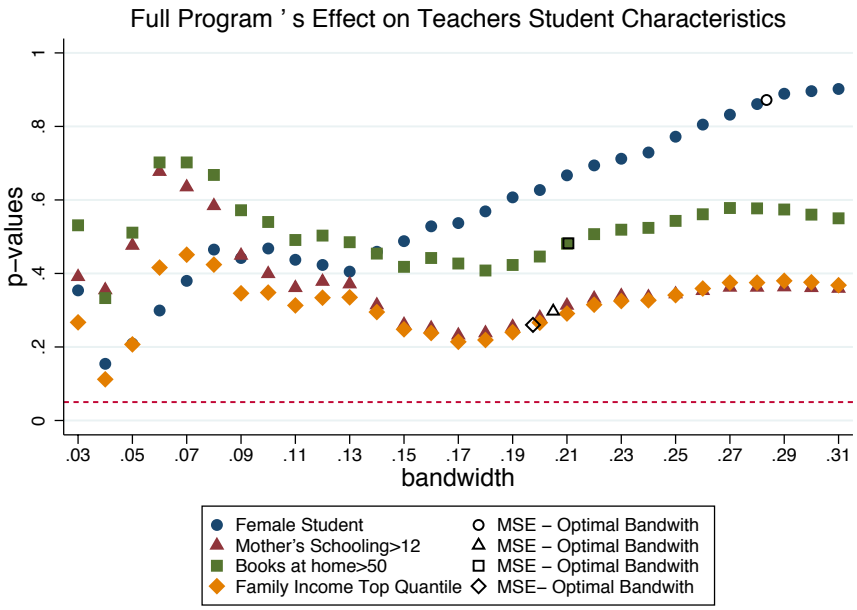
Overall, I find that behavior and performance were not affected by the program after the reward was given to teachers in the first year or the following two years after the certification. The financial component of the program was able to affect the time spent for class preparation for those teachers who were already spending extensive time for class preparation. However, these changes did not result in higher final ex-post outcomes. This lack of statistically-significant results cannot be explained by changes in either the student pool assigned to certified teachers or the schools where they taught. Overall, certified teachers did not modify their teaching practices, self-confidence, or effectiveness. This is consistent with the conclusions of the qualitative study performed by Araya (2015), in which the teachers report that the program, on one hand, came to reaffirm what they already knew in terms of their capacity; and on the other hand, it did not make them elicit any more effort after they received the reward. They reported no increased effectiveness in their practices. All this did not trigger a change in their behavior or performance.

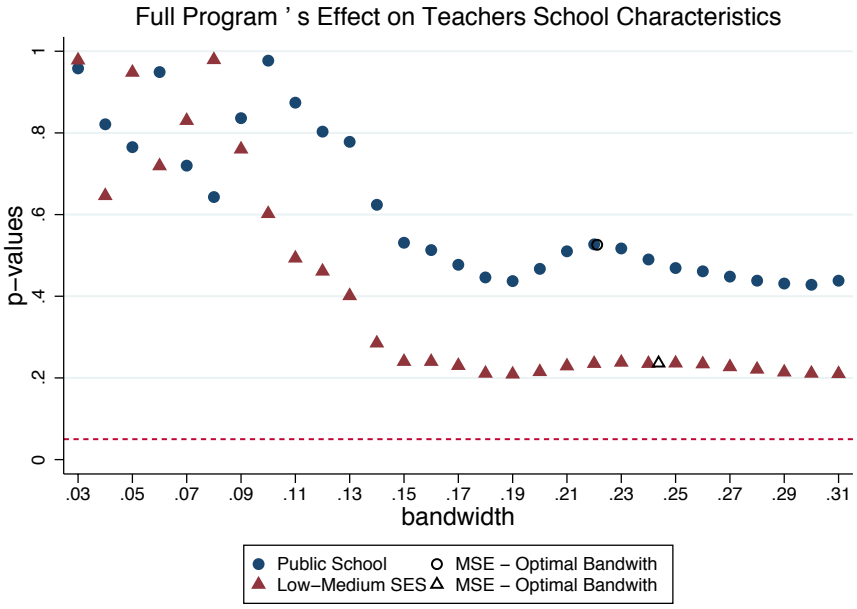
Taken all together, the program may have been effective in acknowledging and rewarding high-quality math teachers, but that does not translate into increased education quality.

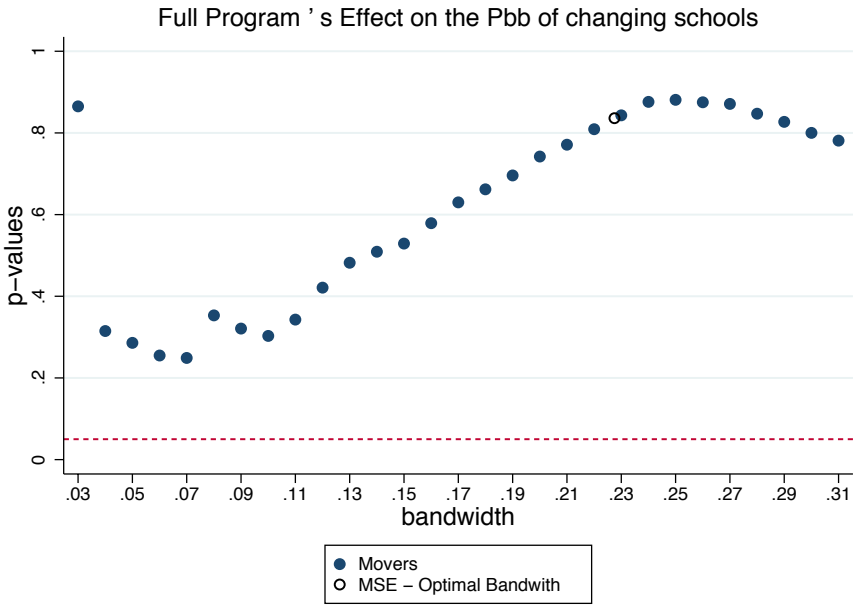
Figure 10.13. Robust p-value and window length. Full program.





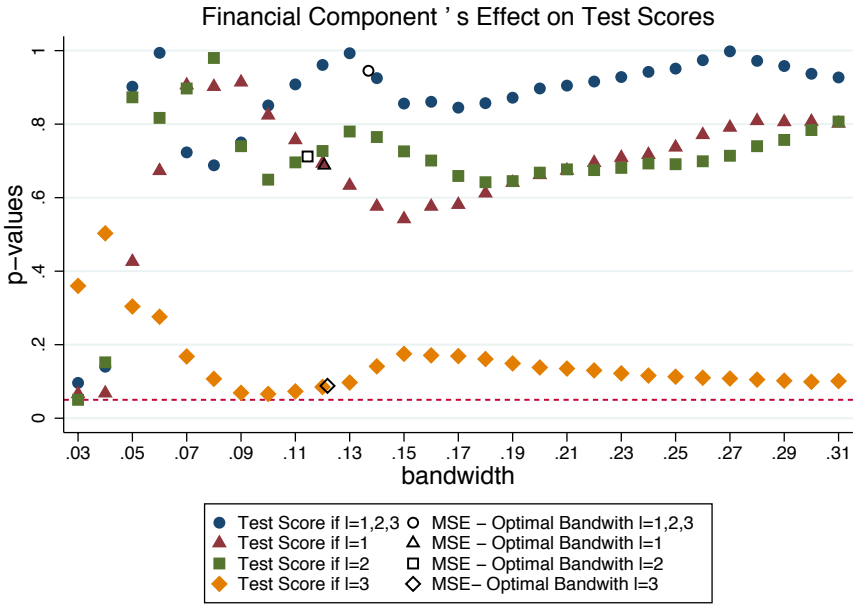

Full Program ' s Effect on Teachers Student Characteristics
p−values
0
.2
.4
.6
.8
1
.03
.05
.07
.09
.11
.13
.15
.17
.19
.21
.23
.25
.27
.29
.31
bandwidth
Female Student
Mother's Schooling>12
Books at home>50
Family Income Top Quantile
MSE – Optimal Bandwith
MSE – Optimal Bandwith
MSE – Optimal Bandwith
MSE– Optimal Bandwith

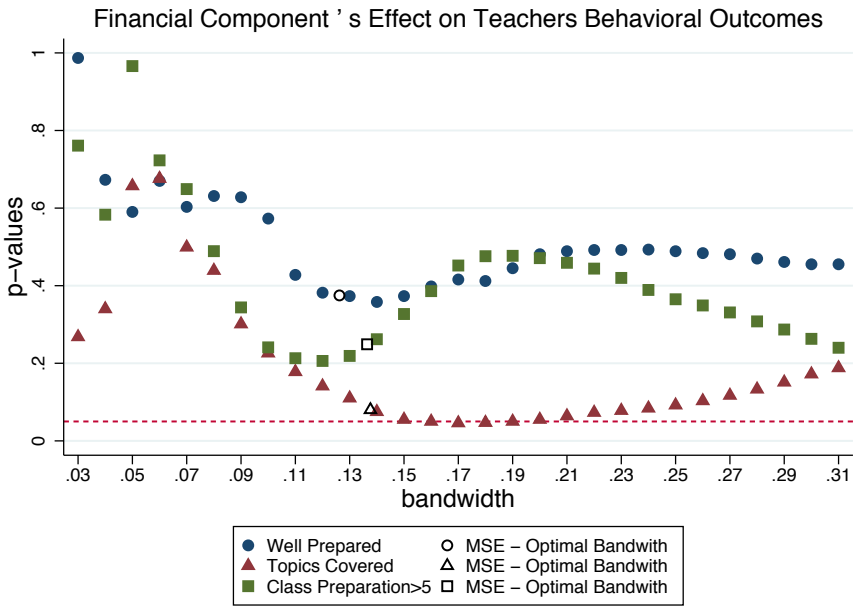

Full Program ' s Effect on Teachers School Characteristics
p−values
0
.2
.4
.6
.8
1
.03
.05
.07
.09
.11
.13
.15
.17
.19
.21
.23
.25
.27
.29
.31
bandwidth
Public School
Low−Medium SES
MSE – Optimal Bandwith
MSE – Optimal Bandwith

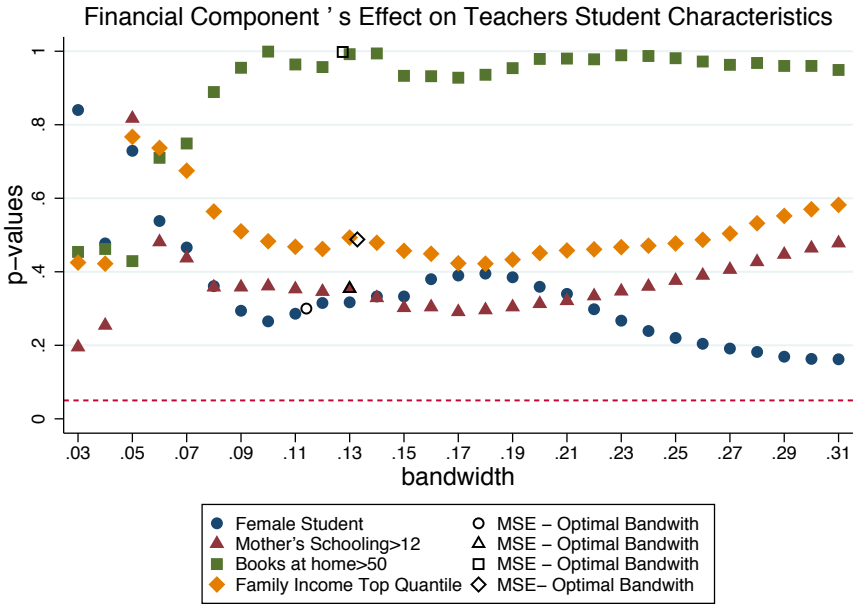


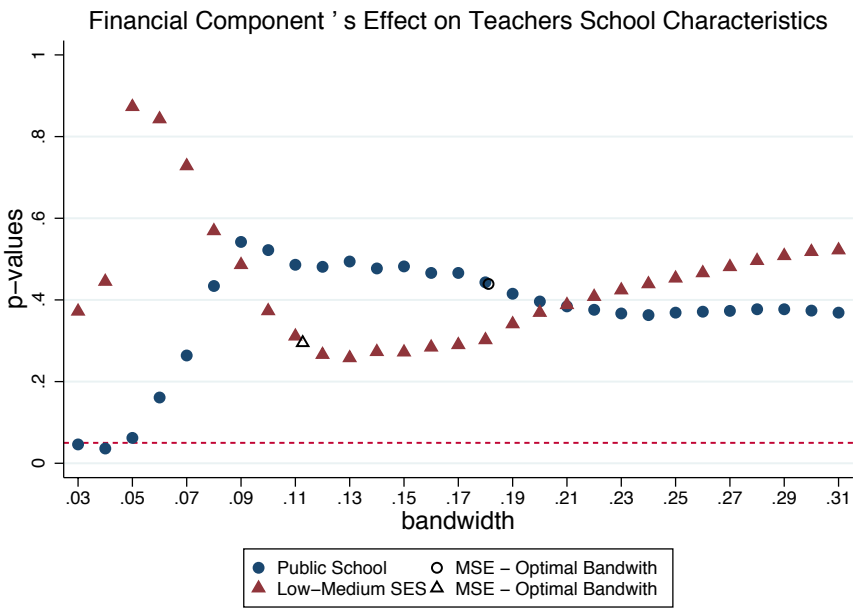
Notes: for each window, treatment effects using the robust specification with Equation 8.6 are estimated. The robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014). The plots graph the robust p-value recovered when testing the hypothesis of null treatment effect against the bandwidth.

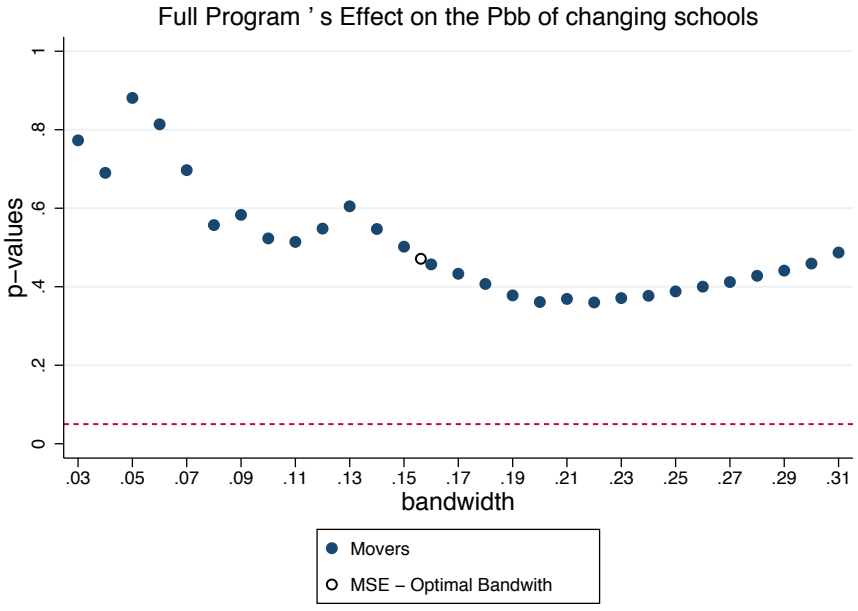
Figure 10.14. Robust p-value and window length. Financial component of the program.




Financial Component ' s Effect on Teachers Behavioral Outcomes
p–values
0
.2
.4
.6
.8
1
.03
.05
.07
.09
.11
.13
.15
.17
.19
.21
.23
.25
.27
.29
.31
bandwidth
Well Prepared
Topics Covered
Class Preparation>5
MSE – Optimal Bandwith
MSE – Optimal Bandwith
MSE – Optimal Bandwith


Financial Component ' s Effect on Teachers Student Characteristics
p–values
0
.2
.4
.6
.8
1
.03
.05
.07
.09
.11
.13
.15
.17
.19
.21
.23
.25
.27
.29
.31
bandwidth
Female Student
Mother's Schooling>12
Books at home>50
Family Income Top Quantile
MSE – Optimal Bandwith
MSE – Optimal Bandwith
MSE – Optimal Bandwith
MSE– Optimal Bandwith





Notes: for each window, treatment effects using the robust specification with Equation 8.6 are estimated. The robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014). The plots graph the robust p-value recovered when testing the hypothesis of null treatment effect against the bandwidth.

Chapter 11 - Discussion and Conclusion

As mentioned in previous sections, this research has been developed and conducted to bring a deeper analysis and understanding of the effects of a teacher's certification program that operated in Chile from 2002 to 2011. I start with an explanation of the Teaching Excellence Award[31] (AEP for its initials in Spanish), then I show mixed evidence regarding the overall effect of incentives to increase teacher efforts in developing and developed countries. Based on this literature review, I showed how the Motivational Model (MM), developed by Bénabou and Tirole (2005), could help us to understand the effects of a program like AEP under the specific assumption of ceteris paribus in terms of student and school sorting. This leads to research questions about the full program effect, the unbundled program effect, the underlying mechanism, and the fade-out process of the effects. In order to address those questions, I presented the methodology and empirical strategy showing my identification strategy in theory.

After this theoretical work, I moved toward the empirical implementation of my research, starting with showing that the identification strategy is valid for identifying the effects of AEP on educational outcomes. I ran specification checks showing that a Sharp Regression Discontinuity (SRD) design could be used. That is, I find a discrete and deterministic change in the probability of receiving the incentive at a predefined cutoff point. Additionally, there is no evidence of manipulation of the running variable and no effect on placebo outcomes during the application year. I also found no evidence of sorting or attrition.

[31] Asignación de Excelencia Pedagógica

The first important finding involves identifying the program's results, which were estimated jointly for math and language subjects. Even though the program does not work as expected for math and language, the reasons for this are different based on the subject. I found that rewarding teacher performance did not systematically change teaching practices or self-confidence or result in higher test scores due to the expected design of the program. Overall, when measured by test scores, the program was ineffective in increasing the quality of the education system.

A second important finding of this research is related to testing the MM in the context of teacher incentives. Originally the MM was expected to be the main explanation for the program's effect. However, I found not statistically significant support for MM. This does not reduce the importance of testing MM in the context of AEP. On the contrary, it is key studying how a program aimed at acknowledging high-performance to improve educational quality was not able to significantly change teachers' behavior. Consequently, we learn that when studying financial and non-financial incentives, complementary theories need to be used in order to better understand their effects and consequences.

A third important finding is that the non-financial and financial components of the reward do not consistently affect behavior, working conditions, or test scores. Though in principle I expected each component of the program to provide some insight on how the program might work, the evidence does not provide relevant information about it.

Fourth, this research has been able to test if the effect of AEP faded out, bringing evidence relative to a potential fading-out of the incentive program's effects. I found that the program had no significant effect after the first, second, and third year after the reward was given.

Fifth, I found that teachers who apply to the program are on average different than teachers who do not apply on several observable characteristics, this imply that the results found do not necessarily extrapolate to the whole teacher population.

In terms of further research, there are several questions that need to be addressed. First, what is the effect of the program as measured by a teacher's value added? It would be interesting to investigate the results' robustness by using a teacher's value added as an outcome of interest. To do so, another data set would be required. In this case, the value-added measure is not possible to estimate because there is only one observation of student achievement while the literature suggests having a student's prior test scores as a control variable. Second, does the effect of AEP vary with teacher and school characteristics? It might be the case that the effects were heterogeneous by teachers' and schools' characteristics, such as credentials, experience, and school size. For example, public recognition might have a larger effect in smaller schools. In schools with many students and colleagues, the reward that one individual teacher receives could go unnoticed. In addition, another potential question is if the school's internal culture affected the program's results. For instance, principals who acknowledged the teaching merit and supported collaborative behavior among teachers may leverage the positive effects of the program in terms of improving students' outcomes. Lastly, is there any spillover effect from AEP-certified teachers on their colleagues within schools? Their colleagues might benefit from the informational component of the certification. In this case, the signaling power of the program may be effective in increasing the quality of the school system. If we assume that individual teachers do not know how to improve their teaching performance or how much effort they should be using, then they might benefit from the interaction with AEP teachers in learning

better practices, following their model, and obtaining their advice. Importantly, this would be evidence of complementarities in the education production function, which brings support to collective incentives rather than individual ones.

Summing up, this research makes several contributions. In more general terms, studying AEP has contributed to understanding how a reward might affect the educational system, with empirical evidence on the financial and non-financial effect; the latter incentive is rarely analyzed despite its potential to be a cost-effective program. Also, the duration of the effects was studied. Ultimately, the objective of this research was met by identifying and testing the underlying mechanisms of the program and then showing whether a program that rewards teacher merit is effective at increasing educational quality. In this context, this paper also contributes to public policies aimed at improving learning outcomes. Here the effect of AEP and its underlying mechanisms are found to have not worked as expected. In fact, the certification was not systematically capable of increasing teacher self-esteem or fostering educational quality. It also did not result in the matching of the most effective teachers to the neediest students. This evidence, as I said above, is informative to researchers and policymakers, opening both new avenues of research and improvements to public policy design.

Bibliography

Daniel Aaronson, Lisa Barrow, and William Sander. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, 25(1):95–135, January 2007. ISSN 0734-306X. doi: 10.1086/508733. URL http://www.journals.uchicago.edu/doi/10.1086/508733.

Rosemarie Anderson, Sam T. Manoogian, and J. Steven Reznick. The undermining and enhancing of intrinsic motivation in preschool children. *Journal of personality and social psychology*, 34 (5):915, 1976.

Maria Caridad Araujo, Pedro Manuel Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady. Teacher Quality and Learning Outcomes in Kindergarten. SSRN Scholarly Paper ID 2750279, *Social Science Research Network*, Rochester, NY, March 2016. URL http://papers.ssrn.com/ abstract=2750279.

Carolina Araya. Estudios de Validez del Programa AEP. In Beatriz Rodríguez, Jorge Manzi, Claudia Peirano, Roberto González, and David Bravo, editors, Reconociendo el Mérito Docente. Programa de Asignación de Excelencia Pedagógica 2002-2014, Chapter 7, pages 307–320. Centro UC. Medición. MIDE, Santiago, 2015.

Carolina Araya, Sandy Taut, Verónica Santelices, and Jorge Manzi. Validez consecuencial del programa de asignación de excelencia pedagógica en chile. Estudios pedagógicos (Valdivia), 37 (2): 25–42, 2011.

Dan Ariely, Anat Bracha, and Stephan Meier. Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially. *American Economic Review*, 99(1):544–555, March 2009. ISSN 0002-8282. doi: 10.1257/aer.99.1.544. URL https://www.aeaweb.org/ articles?id=10.1257/aer.99.1.544.

Ashadi Ashadi and Suzanne Rice. High Stakes Testing and Teacher Access to Professional Opportunities: Lessons from Indonesia. *Journal of Education Policy*, 31(6):727–741, 2016.

Ghazala Azmat and Nagore Iriberri. The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7–8):435–452, August 2010. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2010.04.001. URL http://www.sciencedirect.com/science/article/pii/S0047272710000411.

Gadi Barlevy and Derek Neal. Pay for Percentile. *American Economic Review*, 102(5):1805–1831, May 2012. ISSN 0002-8282. doi: 10.1257/aer.102.5.1805. URL https://www.aeaweb.org/ articles?id=10.1257/aer.102.5.1805.

Jere R. Behrman, Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Journal of Political Economy*, 123(2):325–364, April 2015. ISSN 0022-3808. doi: 10.1086/ 675910. URL http://www.journals.uchicago.edu/doi/10.1086/675910.

Timothy Besley and Maitreesh Ghatak. Status Incentives. *The American Economic Review*, 98(2): 206–211, 2008. ISSN 0002-8282. URL http://www.jstor.org/stable/29730021.

Eric P. Bettinger. Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. *Review of Economics and Statistics*, 94(3):686–698, July 2011. ISSN 0034-6535.

Jordi Blanes i Vidal and Mareike Nossol. Tournaments Without Prizes: Evidence from Personnel Records. *Management Science*, 57(10):1721–1736, August 2011. ISSN 0025-1909. doi: 10. 1287/mnsc.1110.1383. URL http://pubsonline.informs.org/doi/abs/10.1287/mnsc.1110.1383.

Jennifer Booher-Jennings. Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal*, 42(2):231–268, June 2005. ISSN 0002-8312, 1935-1011. doi: 10.3102/00028312042002231. URL http://aer.sagepub.com/content/42/2/231.

George J. Borjas. Labor economics, volume 2. McGraw-Hill New York, 2000.

Roland Bénabou and Jean Tirole. Self-confidence and Personal Motivation. *The Quarterly Journal of Economics*, 117(3):871–915, 2002. doi: 10.1162/003355302760193913. URL +http://dx.doi. org/10.1162/003355302760193913.

Roland Bénabou and Jean Tirole. Intrinsic and Extrinsic Motivation. *The Review of Economic Studies*, 70(3):489–520, January 2003. ISSN 0034-6527, 1467-937X. doi: 10.1111/1467-937X. 00253. URL http://restud.oxfordjournals.org/content/70/3/489.

Roland Bénabou and Jean Tirole. Self-Confidence and Personal Motivation. In Bina Agarwal and Alessandro Vercelli, editors, Psychology, Rationality and Economic Behaviour, *International Economic Association Series*, pages 19–57. Palgrave Macmillan UK, 2005. ISBN 978-1-349-52144-9 978-0-230-52234-3. URL http://link.springer.com/chapter/10.1057/9780230522343_ 2. DOI: 10.1057/9780230522343_2.

David Bravo, Denise Falck, Roberto González, Jorge Manzi, and Claudia Peirano. La relación entre la evaluación docente y el rendimiento de los alumnos: evidencia para el caso de Chile. *Manuscrito Centro de Microdatos*, Universidad de Chile, 2008.

Sebastián Calonico. Rdrobust: Software for Regression-Discontinuity Designs. *Stata Journal*, 17 (2):372–404(33), 2017. URL http://www.stata-journal.com/article.html?article=st0366_1.

Sebastian Calonico, Matias D. Cattaneo, and Rocio Titiunik. Robust Nonparametric Confidence Intervals for Regression Discontinuity Designs. *Econometrica*, 82(6):2295–2326, November 2014. ISSN 1468-0262. doi: 10.3982/ECTA11757. URL http://onlinelibrary.wiley.com/doi/10. 3982/ECTA11757/abstract.

Sebastian Calonico, Matias D. Cattaneo, and Max H. Farrell. On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. arXiv:1508.02973 (math, stat), August 2015a. URL http://arxiv.org/abs/1508.02973. arXiv: 1508.02973.

Sebastian Calonico, Matias D. Cattaneo, and Rocio Titiunik. Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*, 110(512):1753–1769, 2015b. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2015.1017578.

Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell, and Rocıo Titiunik. rdrobust: Software for Regression Discontinuity Designs. Technical report, *Technical report Working paper*, University of Michigan. URL: http://www-personal. umich. edu/cattaneo/software/rddensity/R/rddensitymanual. pdf, 2016a. URL http://faculty.chicagobooth.edu/ max.farrell/research/Calonico-Cattaneo-Farrell-Titiunik_2016_Stata.pdf.

Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell, and Rocıo Titiunik. Regression Discontinuity Designs Using Covariates. *Technical report, working paper*, University of Michigan, 2016b. URL http://sticerd.lse.ac.uk/seminarpapers/em12052016.pdf.

Sebastian Calonico, Nicolas Idrobo, and Rocío Titiunik. A Practical Introduction to Regression Discontinuity Designs: Part I. *Technical report*, Monograph prepared for Cambridge Elements: Quantitative and Computational Methods for Social Science Cambridge University Press, 2017. URL http://www-personal.umich.edu/~cattaneo/books/Cattaneo-Idrobo-Titiunik_2017_ Cambridge-Part1.pdf.

A. Colin Cameron and Douglas L. Miller. A Practitioner's Guide to Cluster-robust Inference. *Journal of Human Resources*, 50(2):317–372, 2015.

Steven Cantrell and Thomas J. Kane. Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study. *Policy and Practice Brief*. MET Project. Bill & Melinda Gates Foundation, January 2013. URL http://eric.ed.gov/?id= ED540958.

Scott E. Carrell and James E. West. Does professor quality matter? evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3):409–432, 2010.

Matias D Cattaneo, Rocio Titiunik, and Gonzalo Vazquez-Bare. Comparing inference approaches for rd designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management*, 36(3):643–681, 2017.

Christopher P. Cerasoli, Jessica M. Nicklin, and Michael T. Ford. Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. Psychological bulletin, 140(4): 980, 2014.

Raj Chetty, John N. Friedman, and Jonah E. Rockoff. The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. *NBER Working Paper No. 17699*. National Bureau of Economic Research, December 2011.

Arnaud Chevalier and Peter Dolton. The labour market for teachers. University College Dublin, Department of Economics, 2004.

What Works Clearinghouse. February 2018. National board for professional teaching standards certification. URL https://files.eric.ed.gov/fulltext/ED580902.pdf.

Olivier Compte and Andrew Postlewaite. Confidence-Enhanced Performance. *American Economic Review*, 94(5):1536–1557, December 2004. ISSN 0002-8282. doi: 10.1257/0002828043052204. URL https://www.aeaweb.org/articles?id=10.1257/0002828043052204.

James Cowan and Dan Goldhaber. National board certification and teacher effectiveness: Evidence from Washington state. *Journal of Research on Educational Effectiveness*, 9(3):233–258, 2016. doi: 10.1080/19345747.2015.1099768. URL https://doi.org/10.1080/19345747.2015.1099768.

Charlotte Danielson. Enhancing Professional Practice: A Framework for Teaching. Association for Supervision and Curriculum Development, 1250 N. Pitt St., Alexandria, VA 22314-1453 (ASCD Stock No. 196074, $16.95 members; $19.95 nonmembers)., 2nd edition, 1996. ISBN 978-0-87120-269-7. URL http://eric.ed.gov/?id=ed403245.

Jane L. David. What Students Need to Learn: High-stakes Testing Narrows the Curriculum. *Educational Leadership*, 2011

Edward L. Deci. The effects of contingent and noncontingent rewards and controls on intrinsic motivation. *Organizational behavior and human performance*, 8(2):217–229, 1972.

Edward L. Deci and Richard M. Ryan. Intrinsic motivation. *Wiley Online Library*, 1975. ISBN 0-470-47921-3.

Edward L. Deci and Richard M. Ryan. The empirical exploration of intrinsic motivational processes. *In Advances in experimental social psychology*, volume 13, pages 39–80. Elsevier, 1980.

Edward L. Deci, Richard Koestner, and Richard M. Ryan. Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of educational research*, 71(1):1–27, 2001.

Thomas S. Dee and James Wyckoff. Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, 34(2):267–297, March 2015. ISSN 1520-6688. doi: 10.1002/pam.21818. URL http://onlinelibrary.wiley.com/doi/10.1002/pam. 21818/abstract.

Esther Duflo, Rema Hanna, and Stephen P. Ryan. Incentives Work: Getting Teachers to Come to School. *American Economic Review*, 102(4):1241–78, June 2012. ISSN 0002-8282. doi: 10.1257/aer.102.4.1241. URL https://www.aeaweb.org/articles?id=10.1257/aer.102.4.1241.

Eric Eide, Dan Goldhaber, and Dominic Brewer. The teacher labour market and teacher quality. *Oxford Review of Economic Policy*, 20(2):230–244, 2004.

Amitai Etzioni. Modern organizations. Technical report, 1964.

Denise Falck, Martha Kluttig, and Valentina Riberi. Resultados Generales 2002-2012. In Beatriz Rodríguez, Jorge Manzi, Claudia Peirano, Roberto González, and David Bravo, editors, Reconociendo el Mérito Docente. Programa de Asignación de Excelencia Pedagógica 2002-2014, Chapter 7, pages 54–77. Centro UC. Medición. MIDE, Santiago, 2015.

David N. Figlio and Joshua Winicki. Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics*, 89(2–3):381–394, February 2005. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2003.10.007. URL http://www.sciencedirect.com/science/article/ pii/S0047272704000209.

Antonio Filippin and Marco Paccagnella. Family background, self-confidence and economic outcomes. *Economics of Education Review*, 31(5):824–834, 2012.

William A. Firestone and James R. Pennell. Teacher Commitment, Working Conditions, and Differential Incentive Policies. *Review of Educational Research*, 63(4):489–525, December 1993. ISSN 0034-6543, 1935-1046. doi: 10.3102/00346543063004489. URL http://rer.sagepub.com/ content/63/4/489.

Bruno S. Frey. Awards as compensation. *European Management Review*, 4(1):6–14, March 2007. ISSN 1740-4762. doi: 10.1057/palgrave.emr.1500068. URL http://onlinelibrary.wiley.com/doi/ 10.1057/palgrave.emr.1500068/abstract.

Bruno S. Frey and Jana Gallus. Towards an economics of awards. Journal of Economic Surveys, 31(1):190–200, 2017.

Roland Fryer, Steven Levitt, John List, and Sally Sadoff. Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment. Technical Report w18237, *National Bureau of Economic Research*, Cambridge, MA, July 2012. URL http://www.nber.org/papers/w18237.pdf.

Roland G. Fryer. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, 31(2):373–407, April 2013. ISSN 0734-306X. doi: 10.1086/667757. URL http://www.journals.uchicago.edu/doi/full/10.1086/667757.

Roland G. Fryer Jr, Steven D. Levitt, John List, and Sally Sadoff. Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. *Technical report, National Bureau of Economic Research*, 2012.

Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. *Technical report, National Bureau of Economic Research*, 2014. URL http://www.nber.org/papers/w20405.pdf.

David Gillborn and Deborah Youdell. Rationing Education: Policy, Practice, Reform and Equity. *British Journal of Educational Studies*, 50(2):289–290, 2002.

Paul Glewwe and Michael Kremer. Schools, teachers, and education outcomes in developing countries. *Handbook of the Economics of Education*, 2:945–1017, 2006.

Paul Glewwe, Nauman Ilias, and Michael Kremer. Teacher Incentives. *American Economic Journal: Applied Economics*, 2(3):205–27, July 2010. ISSN 1945-7782. doi: 10.1257/app.2.3.205. URL https://www.aeaweb.org/articles?id=10.1257/app.2.3.205.

Sarena F. Goodman and Lesley J. Turner. The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. *Journal of Labor Economics*, 31 (2):409–420, April 2013. ISSN 0734306X. doi: http://www.jstor.org/action/showPublication? journalCode=jlaboreconomics.

Pam Grossman, Susanna Loeb, Julie Cohen, and James Wyckoff. Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers? value-added scores. *American Journal of Education*, 119(3):445–470, 2013.

Barton H. Hamilton, Jack A. Nickerson, and Hideo Owan. Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation. *Journal of Political Economy*, 111(3):465–497, 2003. ISSN 0022-3808. doi: 10.1086/374182. URL http://www.jstor.org/stable/10.1086/374182.

Herbert G. Heneman III, Anthony Milanowski, Steven Kimball, and Allan Odden. Standards-based teacher evaluation as a foundation for knowledge-and skill-based pay. 2006. URL http://repository.upenn.edu/cpre_policybriefs/33/?utm_source=repository.upenn.edu%2Fcpre_ policybriefs%2F33&utm_medium=PDF&utm_campaign=PDFCoverPages.

Bengt Holmstrom. Moral Hazard in Teams. *The Bell Journal of Economics*, 13(2):324–340, 1982. ISSN 0361-915X. doi: 10.2307/3003457. URL http://www.jstor.org/stable/3003457.

Bengt Holmstrom and Paul Milgrom. Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization*, 7(special issue):24–52, January 1991. ISSN 8756-6222, 1465-7341.

Bernardo A. Huberman, Christoph H. Loch, and Ayse ÖNçüler. Status as a Valued Resource. *Social Psychology Quarterly*, 67(1):103–114, March 2004. ISSN 0190-2725, 1939-8999. doi: 10.1177/019027250406700109. URL http://spq.sagepub.com/content/67/1/103.

Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, page rdr043, 2011. URL http://restud. oxfordjournals.org/content/early/2012/01/07/restud.rdr043.short.

Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86, 2009. URL http://www.ingentaconnect. com/content/aea/jel/2009/00000047/00000001/art00001.

Scott A. Imberman and Michael F. Lovenheim. Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System. *Review of Economics and Statistics*, 97(2):364–386, May 2015. ISSN 00346535. doi: http://www.mitpressjournals.org/loi/rest.

Hideshi Itoh. Incentives to Help in Multi-Agent Situations. *Econometrica*, 59(3):611, May 1991. ISSN 00129682. doi: 10.2307/2938221.

C. Kirabo Jackson, Jonah E. Rockoff, and Douglas O. Staiger. Teacher effects and teacher-related policies. *Annu. Rev. Econ*., 6(1):801–825, 2014. URL http://www.annualreviews.org/doi/abs/ 10.1146/annurev-economics-080213-040845.

Brian A. Jacob and Steven D. Levitt. Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics*, 118(3):843–877, 2003. ISSN 0033-5533. URL http://www.jstor.org/stable/25053925.

Nina Jalava, Juanna Schrøter Joensen, and Elin Pellas. Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115:161–196, July 2015. ISSN 0167-2681. doi: 10.1016/j.jebo.2014.12.004. URL http://www.sciencedirect. com/science/article/pii/S0167268114003163.

Gail M. Jones, Brett D. Jones, and Tracy Hargrove. The Unintended Consequences of High-stakes Testing. *Rowman & Littlefield Publishers*, 2003.

Eugene Kandel and Edward P. Lazear. Peer Pressure and Partnerships. *Journal of Political Economy*, 100(4):801–817, 1992. ISSN 0022-3808. URL http://www.jstor.org/stable/2138688.

Michihiro Kandori. Social Norms and Community Enforcement. *The Review of Economic Studies*, 59(1):63–80, 1992. ISSN 0034-6527. doi: 10.2307/2297925. URL http://www.jstor.org/stable/ 2297925.

Thomas J. Kane and Steven Cantrell. Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. *Research Paper. MET Project*. Bill & Melinda Gates Foundation, 2010. URL http://eric.ed.gov/?id=ED528382.

Thomas J. Kane and Douglas O. Staiger. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. *Working Paper 14607, National Bureau of Economic Research*, December 2008. URL http://www.nber.org/papers/w14607.

Thomas J. Kane and Douglas O. Staiger. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. *Research Paper. MET Project*. Bill & Melinda Gates Foundation, January 2012. URL http://eric.ed.gov/?id= ED540960.

Alexander Koch, Julia Nafziger, and Helena Skyt Nielsen. Behavioral economics of education. *Journal of Economic Behavior & Organization*, 115:3–17, July 2015. ISSN 0167-2681.

Daniel M. Koretz. Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. *Journal of Human Resources*, 37(4):752–777, 2002. ISSN 0022166X.

Michael Kosfeld and Susanne Neckermann. Getting More Work for Nothing? Symbolic Awards and Worker Performance. *American Economic Journal: Microeconomics*, 3(3):86–99, August 2011. ISSN 1945-7669. doi: 10.1257/mic.3.3.86.

David M. Kreps. Intrinsic Motivation and Extrinsic Incentives. *American Economic Review*, 87 (2):359–364, May 1997. ISSN 00028282.

Helen F. Ladd. The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes. *Economics of Education Review*, 18(1):1–16, February 1999. ISSN 0272-7757. doi: 10.1016/S0272-7757(97)00044-7.

Victor Lavy. Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement. *Journal of Political Economy*, 110(6):1286–1317, December 2002. ISSN 0022-3808. doi: 10.1086/342810. URL http://www.journals.uchicago.edu/doi/10.1086/342810.

Victor Lavy. Performance Pay and Teachers; Effort, Productivity, and Grading Ethics. *American Economic Review*, 99(5):1979–2011, December 2009. ISSN 0002-8282. doi: 10.1257/aer.99.5. 1979. URL https://www.aeaweb.org/articles?id=10.1257/aer.99.5.1979.

Edward P. Lazear. Salaries and Piece Rates. *The Journal of Business*, 59(3):405–431, 1986. ISSN 0021-9398. URL http://www.jstor.org/stable/2352711.

Edward P. Lazear. Performance Pay and Productivity. *American Economic Review*, 90(5): 1346–1361, 2000.

David S. Lee and Thomas Lemieuxa. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010. URL http://www.ingentaconnect.com/content/aea/jel/2010/00000048/00000002/art00001.

Rosalind Levačić. Teacher incentives and performance: An application of principal–agent theory. *Oxford Development Studies*, 37(1):33–46, 2009.

Steven D. Levitt, John A. List, Susanne Neckermann, and Sally Sadoff. The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. *Working Paper 18165, National Bureau of Economic Research*, June 2012. URL http://www.nber.org/papers/w18165.

Susanna Loeb. How can value-added measures be used for teacher improvement? what we know series: Value-added methods and applications. knowledge brief 13. *Carnegie Foundation for the Advancement of Teaching*, 2013.

Justin McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714, 2008. URL http://www.sciencedirect.com/science/article/pii/S0304407607001133.

Ministry of Education. Decreto con Fuerza de Ley 1. 1997.

Ministry of Education. El Marco para la Buena Enseñanza. Santiago, Chile, 2004.

Ministry of Education. Decreto 439. 2009.

Ministry of Education. Estadísticas de la Educación 2012. 2012a. URL /content/book/eag-2017-en.

Ministry of Education. Decreto con Fuerza de Ley 1. 2012b URL http://www.curriculumenlineamineduc.cl/605articles-22394_programa.pdf.

Ministry of Education. Metodología de Construcción de Grupos Socioeconómicos. 2013. URL http://archivos.agenciaeducacion.cl/Metodologia-de-Construccion-de-Grupos-Socioeconomicos-SIMCE-2012.pdf.

Susan Moore Johnson. Merit Pay for Teachers: A Poor Prescription for Reform. *Harvard Educational Review*, 54(2):175–186, July 1984. ISSN 0017-8055. doi: 10.17763/haer.54.2.36264448513xp4t5. URL http://hepgjournals.org/doi/10.17763/haer.54.2.36264448513xp4t5.

Karthik Muralidharan and Venkatesh Sundararaman. Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, 119(1):39–77, February 2011. ISSN 0022-3808. doi: 10.1086/659655.

National Research Council and others. Assessing Accomplished Teaching: Advanced-level Certification Programs. *National Academies Press*, 2008.

Derek Neal. The Design of Performance Pay in Education. *Working Paper 16710, National Bureau of Economic Research*, January 2011. URL http://www.nber.org/papers/w16710.

Derek Neal and Diane Whitmore Schanzenbach. Left Behind by Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics*, 92(2):263–283, February 2010. ISSN 0034-6535. doi: 10.1162/rest.2010.12318. URL http://dx.doi.org/10.1162/rest.2010.12318.

OECD. Education at a Glance 2014. Chile. 2014.

OECD. Indicator B6 on What Resources and Services is Education Funding Spent. 2015.

OECD. Education at a Glance 2017. 2017.

Michael J. Podgursky and Matthew G. Springer. Teacher performance pay: A review. *Journal of Policy Analysis & Management*, 26(4):909–950, October 2007. ISSN 02768739. doi: 10.1002/ pam.20292.

John Polesel, Suzanne Rice, and Nicole Dulfer. The Impact of High-stakes Testing on Curriculum and Pedagogy: A Teacher perspective from Australia. *Journal of Education Policy*, 29(5): 640–657, 2014.

Jack Porter. Estimation in the regression discontinuity model. *Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison*, pages 5–19, 2003. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.540&rep=rep1&type=pdf.

Diane Reay and Dylan Wiliam, "I'll Be a Nothing": Structure, Agency and the Construction of Identity through Assessment. *British Educational Research Journal*, 1999.

Randall Reback. Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics*, 92(5–6):1394–1415, June 2008. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2007.05.003.

Diane Stark Rentner, Caitlin Scott, Nancy Kober, Naomi Chudowsky, Victor Chudowsky, Scott Joftus, and Dalia Zabala. From the Capital to the Classroom: Year 4 of the No Child Left behind Act. 2006.

Steven G. Rivkin, Eric A. Hanushek, and John F. Kain. Teachers, Schools, and Academic Achievement. *Econometrica*, 73(2):417–458, March 2005. ISSN 1468-0262. doi: 10.1111/j.

1468-0262.2005.00584.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1468-0262.2005. 00584.x/abstract.

Jonah E. Rockoff. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review*, 94(2):247–252, May 2004. ISSN 0002-8282. doi: 10.1257/0002828041302244. URL https://www.aeaweb.org/articles?id=10.1257/0002828041302244.

Jesse Rothstein. Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, 125(1):175–214, 2010.

Brian Rowan, Richard Correnti, and Robert J. Miller. What Large-Scale, Survey Research Tells us about "teacher effects" on Student Achievement: Insights from the "Prospects" Study of Elementary Schools. *CPRE Research Report Series*. 2002.

Michael Russell, George Madaus, and Jennifer Higgins. The Paradoxes of High-stakes Testing: How They Affect Students, Their Teachers, Principals, Schools and Society. Charlotte, NC: Information Age Publishing, 2009.

Christopher Skovron and Rocío Titiunik. A Practical Guide to Regression Discontinuity Designs in Political Science. *Technical report, working paper, University of Michigan*, 2015. URL http://www-personal.umich.edu/~titiunik/papers/SkovronTitiunik2015.pdf.

Matthew G. Springer. The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, 27(5):556–563, October 2008. ISSN 0272-7757. doi: 10.1016/j.econedurev.2007.06.004. URL http://www.sciencedirect.com/science/ article/pii/S0272775707000933.

Matthew G. Springer. Performance incentives: Their growing impact on American K-12 education. *Brookings Institution Press*, 2009. ISBN 0-8157-0195-0.

Matthew G. Springer, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. Mc-Caffrey, Matthew Pepper, and Brian M. Stecher. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT). Society for Research on Educational Effectiveness, 2011. URL http://eric.ed.gov/?id=ED518378.

Eric S. Taylor and John H. Tyler. The effect of evaluation on teacher performance. The *American Economic Review*, 102(7):3628–3651, 2012. URL http://www.ingentaconnect.com/content/aea/ aer/2012/00000102/00000007/art00020.

Anh Tran and Richard Zeckhauser. Rank as an Incentive. 2009.

Sujata Visaria, Rajeev Dehejia, Melody M. Chao, and Anirban Mukhopadhyay. Unintended Con-sequences of Rewards for Student Attendance: Results from a Field Experiment in

Indian Class-rooms. *Working Paper 22528, National Bureau of Economic Research*, August 2016.

X. Henry Wang and Bill Yang. Why competition may discourage students from learning? a behavioral economic analysis. *Education economics*, 11(2):117–128, 2003.

World Bank. World Development Report 2018: Learning to Realize Education? Promise. Washington, DC: *World Bank*, 2018. doi: 10.1596/978-1-4648-1096-1.

Manuel Zymelman and Joseph DeStefano. Primary School Teachers' Salaries in Sub-Saharan Africa. *World Bank Discussion Papers* 45. ERIC, 1989. ISBN 0-8213-1173-5.

Appendix A - Program On-line Certificate

CPEIP

CERTIFICADO

Mediante el presente documento, certifico que Don(a) Larrys Redlich Hvalibota fue Acreditado(a) para percibir la Asignación de Excelencia Pedagógica en el proceso de postulación 2007, según consta la resolución .

Esta acreditación le da derecho a los siguientes beneficios:

- Certificación de Excelencia válida en todo el territorio nacional.
- Beneficio económico que consiste en, aproximadamente, 1 sueldo adicional por año. Este beneficio dura por un período de 10 años, es decir, hasta el año 2017, sólo mientras el postulante cumpla con los requisitos que se han establecido para mantener dicha asignación.
- Habilitación para postular a la Red Maestros de Maestros.

Para mayores detalles se puede visitar el sitio web www.aep.mineduc.cl.

CENTRO DE PERFEC. EXPER. E INVEST. PEDAGOGICAS
AREA ACREDITACION Y EVALUACION DOCENTE
MINEDUC

Rodolfo Bonifaz
Coordinador Nacional Área de Acreditación y Evaluación Docente
Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas

Appendix B - Results Letter

Santiago, abril 2011

Señor(a)
[redacted]
Establecimiento: [redacted]
Presente

Estimada Profesora:

Me dirijo a usted para informarle que, lamentablemente, luego de evaluar los antecedentes que presentó al Programa para la Asignación de Excelencia Pedagógica 2010, en esta oportunidad no ha obtenido la acreditación.

El año 2010 se presentaron voluntariamente a este programa 1.499 profesores de aula de todos los ciclos de enseñanza y provenientes de todas las regiones del país, de los cuales un 17% obtuvo la acreditación.

La decisión de acreditación se basó en el rendimiento que usted demostró tanto en la prueba escrita como en el portafolio. Estos instrumentos se construyeron orientados por el Marco para la Buena Enseñanza, que identifica cuatro dominios: (A) preparación de la enseñanza; (B) creación de un ambiente propicio para el aprendizaje; (C) enseñanza para el aprendizaje de todos los estudiantes, y (D) responsabilidades profesionales. La evaluación integra aspectos de los cuatro dominios, lo que se traduce en un puntaje final que determina la posibilidad de ser acreditado. Tal como se establece en las bases, dicho puntaje final se obtiene ponderando en un 30% el rendimiento en la prueba y en un 70% los resultados del portafolio.

Los instrumentos utilizados en este programa de acreditación han sido especialmente diseñados para el contexto de la educación chilena, atendiendo a las características específicas de la cultura nacional y la diversidad regional. Este proceso, así como la evaluación de todos los antecedentes presentados por usted como parte de su postulación, ha sido realizado por profesionales especialmente seleccionados y entrenados por el equipo técnico de este programa, integrado por académicos de la Pontificia Universidad Católica de Chile y de la Universidad de Chile.

Tal como consta en este informe, las variables que deben considerarse para evaluar un buen desempeño docente son diversas y comprenden los distintos momentos de la enseñanza. A partir de la evidencia entregada por usted sobre su ejercicio profesional en este proceso, es posible apreciar que, aunque se observan fortalezas, aún es necesario desarrollar y reforzar otros aspectos importantes.

A continuación, usted encontrará información detallada respecto del rendimiento que alcanzó en los cinco productos que componen el portafolio, y en las distintas dimensiones que fueron observadas en cada uno de ellos. Posteriormente, se exponen sus resultados en la prueba escrita, desagregados por los distintos ejes que la componen, y su desempeño en las preguntas abiertas pedagógicas y disciplinarias.

Espero que estos resultados le sean de utilidad para orientar y fortalecer su desarrollo profesional.

Sus resultados en el portafolio:

El nivel de logro alcanzado por usted en cada uno de los aspectos evaluados a través del Portafolio AEP, está organizado en función de:

- **Aspectos logrados:** ámbitos en los que usted muestra un desempeño que revela un adecuado dominio de las tareas solicitadas y/o habilidades evaluadas.

- **Aspectos medianamente logrados:** ámbitos en los que usted muestra un desempeño que revela un dominio parcial de las tareas solicitadas y/o habilidades evaluadas; dejando en evidencia la necesidad de trabajar en torno a ellas para reforzarlas.

- **Aspectos no logrados:** ámbitos en los que usted muestra un desempeño que revela un dominio insuficiente de las tareas solicitadas y/o habilidades evaluadas, y que le señalan un desafío ineludible respecto a la necesidad de desarrollarlas.

A continuación encontrará el detalle del desempeño por usted alcanzado en cada uno de los productos que componen el Portafolio AEP.

Producto 1: Planificación de una unidad de aprendizaje

El Producto 1 está orientado a evaluar aquellas competencias que deben ponerse en juego al momento de diseñar una unidad de aprendizaje. A través de este producto se evaluó específicamente su capacidad para describir las variables del contexto involucradas en el proceso de aprendizaje y justificar el modo en que deben ser consideradas al planificar.
Además, se evaluó la calidad del diseño de su planificación, la forma en que usted fundamentó las decisiones pedagógicas tomadas, y su capacidad para analizarla críticamente una vez implementada.

A continuación usted encontrará información acerca del rendimiento que alcanzó en los distintos aspectos evaluados en el Producto 1.

Dimensiones evaluadas en el Producto 1	no logrado	medianamente logrado	logrado
Justifica claramente de qué manera consideró las características de sus estudiantes en la planificación y justifica claramente de qué manera consideró las características del contexto sociocultural y de la institución en la planificación.	•		

En su planificación, formula correctamente los objetivos de aprendizaje para cada clase, manteniendo coherencia con el objetivo de la unidad.	●		
En su planificación, formula correctamente las actividades de aprendizaje para cada una de las clases y mantiene una secuencia lógica y consistente para la unidad.	●		
En su planificación, cada una de las clases presenta una estructura compuesta por los tres momentos de la clase: inicio, desarrollo y cierre.	●		
En la planificación de cada una de sus clases, muestra coherencia entre los objetivos propuestos, los contenidos señalados, las actividades diseñadas y los recursos seleccionados.	●		
En el análisis de una de las clases de la planificación, justifica la incorporación de los contenidos seleccionados en relación a los contenidos de la unidad.	●		
En el análisis de una de las clases de la planificación, justifica apropiadamente la elección de al menos uno de los recursos de aprendizaje en función de los objetivos planteados para ella.	●		
Reformula y adapta su planificación en función de la experiencia ganada en su implementación, con la finalidad de favorecer el logro de los aprendizajes en sus estudiantes.	●		

Producto 2: Estrategia de Evaluación

El producto 2 tiene como propósito evaluar sus competencias para diseñar una estrategia de evaluación y fundamentar las decisiones que la orientaron. A través de este producto se observó su capacidad para analizar críticamente la implementación de distintas actividades de evaluación, extraer información relevante de los resultados obtenidos por los estudiantes y las decisiones tomadas a partir de aquellos.

A continuación, usted encontrará información acerca del rendimiento que alcanzó en los distintos aspectos evaluados en el Producto 2.

Dimensiones evaluadas en el Producto 2	no logrado	medianamente logrado	logrado
Describe la estrategia de evaluación planificada y la explica en función de los criterios que utilizó en la selección y organización de los instrumentos de evaluación.		•	
Describe una actividad o situación de evaluación desarrollada durante la unidad y justifica claramente por qué esta corresponde al tipo de evaluación mencionado.	•		
Formula correctamente la totalidad de los objetivos de evaluación.			•
Formula correctamente la totalidad de los indicadores de evaluación.			•
Establece una relación coherente entre la totalidad de los objetivos de evaluación y sus respectivos indicadores de evaluación.			•
Explica la pertinencia del instrumento de evaluación en función de los objetivos planteados.	•		
Indica los resultados obtenidos por los estudiantes y justifica de manera clara las acciones tomadas a partir de ellos.			•
Menciona a quién o quiénes comunicó los resultados obtenidos por sus estudiantes y explica la importancia de que estos actores reciban dicha información.	•		
Analiza la implementación de su estrategia de evaluación, ilustrando claramente cómo esta favoreció el proceso de enseñanza aprendizaje.	•		

Producto 3: Filmación de una clase

El producto 3 está orientado a evaluar su capacidad para crear un entorno favorable al aprendizaje y presentar con rigurosidad los contenidos correspondientes. Se evaluó específicamente la forma en que usted desarrolla una clase, considerando su estructura y la organización eficiente del trabajo de sus estudiantes. Asimismo, se evaluó la forma en que retroalimenta sus aprendizajes, el grado en que promueve su participación, y el modo en que favorece el desarrollo de habilidades de pensamiento en función de los objetivos de la clase.

A continuación usted encontrará información acerca del rendimiento que alcanzó en los distintos aspectos evaluados en el Producto 3.

Dimensiones evaluadas en el Producto 3	no logrado	medianamente logrado	logrado
Desarrolla un inicio de clase que contiene motivación e información de los aprendizajes a lograr.		•	
Desarrolla un cierre de clase definido donde relaciona las actividades desarrolladas con los aprendizajes a lograr.		•	
Entrega orientaciones precisas y detalladas a sus estudiantes para el desarrollo de las actividades de enseñanza.		•	
Supervisa el desarrollo de las actividades de enseñanza para asegurar un adecuado desempeño de sus estudiantes.		•	
Explicita la vinculación entre los conocimientos previos de los estudiantes y los contenidos de la clase.	•		
Explica y relaciona los conceptos o procedimientos establecidos para la clase.		•	
Potencia habilidades cognitivas propuestas en el Marco Curricular a través de las actividades que desarrolla durante la clase.	•		
Establece un clima de relaciones respetuosas y empáticas, donde asegura la participación de todos los estudiantes de la clase.		•	
Escucha atentamente de manera equitativa los comentarios de todos			

sus estudiantes y los utiliza para profundizar o complementar los contenidos tratados.		●	
Responde de manera oportuna y adecuada las preguntas de sus estudiantes y las incorpra en las explicaciones de los contenidos.	●		
Enseña conceptos del Marco Curricular que se relacionan de manera coherente con los objetivos fijados para la clase.		●	
Enseña procedimientos del Marco Curricular que se relacionan de manera coherente con los objetivos fijados para la clase.			●
Modela actitudes del Marco Curricular que se relacionan de manera coherente con los Objetivos Transversales fijados para la clase.		●	

Producto 4: Análisis de la clase filmada

El Producto 4 está orientado a evaluar su capacidad para analizar críticamente la implementación de la clase filmada. A través de este producto, se evaluó específicamente su capacidad para analizar y justificar la contribución tanto positiva como negativa de diversos factores que son propios del proceso de enseñanza aprendizaje, como son: la implementación de la secuencia de actividades, la distribución del tiempo adoptada para la clase filmada, el manejo de las relaciones interpersonales y el establecimiento de normas de convivencia.

A continuación usted encontrará información acerca del rendimiento que alcanzó en los distintos aspectos evaluados en el Producto 4.

Dimensiones evaluadas en el Producto 4	no logrado	medianamente logrado	logrado
Elabora una secuencia de actividades coherente para la clase filmada y analiza su contribución al logro de los objetivos de aprendizaje planteados para ella.	●		
Justifica la pertinencia de una de las actividades incluidas en la secuencia y la justifica en función de su contribución al logro de los aprendizajes esperados para la clase.	●		

Evalúa y fundamenta claramente la administración del tiempo en la clase filmada en función de las metas logradas o no logradas en ella.		•	
Analiza el clima de relaciones interpersonales desarrollado en la clase filmada, considerando tanto su relación con los estudiantes, como la relación entre ellos.	•		
Identifica una estrategia que promueve el cumplimiento de normas de convivencia para establecer un buen ambiente de trabajo.			•
Justifica la estrategia de convivencia aplicada en la clase filmada en función de su contribución al logro de los objetivos de la clase.		•	

Producto 5: Reflexión pedagógica a partir de la unidad de aprendizaje implementada

El Producto 5 está orientado a evaluar la capacidad de realizar un análisis crítico de su propia práctica pedagógica. A través de este producto se evaluó específicamente el análisis de su influencia en el aprendizaje de sus estudiantes, así como la capacidad de generar medidas conducentes a mejorar o potenciar estos logros en sus alumnos. Además, se evaluó su capacidad de identificar fortalezas y debilidades de su práctica pedagógica y plantear una medida que le permita mejorar su desempeño docente.

A continuación usted encontrará información acerca del rendimiento que alcanzó en los distintos aspectos evaluados en el Producto 5.

Dimensiones evaluadas en el Producto 5	no logrado	medianamente logrado	logrado
Analiza críticamente su práctica de enseñanza, identificando y explicando claramente la mayoría de las decisiones pedagógicas que favorecieron el logro del aprendizaje de sus estudiantes en la unidad.		•	
Analiza críticamente su práctica de enseñanza, identificando y explicando claramente la mayoría			

explicando claramente la mayoría de las decisiones pedagógicas que dificultaron o impidieron el logro del aprendizaje de sus estudiantes en la unidad.	•		
A partir de los resultados de aprendizaje de sus estudiantes, propone una medida clara que contribuya a mejorar o potenciar sus niveles de logro.		•	
Reconoce el progreso obtenido por un estudiante en un aprendizaje específico y es capaz de explicar con claridad cómo se manifiesta este progreso.	•		
Reconoce e ilustra con situaciones concretas, la principal fortaleza y la principal debilidad de su ejercicio docente.		•	
Propone una medida clara para mejorar una debilidad de su ejercicio docente.		•	

Sus resultados en la Prueba de Conocimientos Disciplinarios y Pedagógicos

La prueba es un instrumento que evalúa el grado en que el profesor maneja los conocimientos disciplinarios y pedagógicos asociados al sector de aprendizaje y al nivel de enseñanza en que se desempeña. Los contenidos de este instrumento fueron definidos por el Ministerio de Educación a partir del currículo vigente y se publicaron en el sitio web del Programa AEP durante el mes de junio de 2010.

En el siguiente cuadro, usted encontrará información acerca del rendimiento que alcanzó en los ítemes cerrados de la prueba. Le agradeceremos considerar lo siguiente:

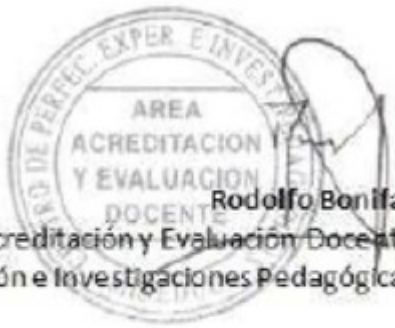
1. La segunda columna indica el porcentaje de respuestas correctas que usted obtuvo en cada uno de los ejes temáticos comprendidos en la prueba que rindió.
2. La tercera columna muestra el porcentaje de respuestas correctas que alcanzó, en promedio, el grupo de personas que rindió la prueba de **Educación Parvularia**

	Eje Temático	% respuestas correctas LILIANA ROSA SILVA SOTO	% respuestas correctas promedio
Preguntas Cerradas	Comunicación	46%	43%
	Conocimiento de las Bases Curriculares	33%	47%

	Formación Personal y Social	44%	46%
	Relación con el Medio Natural y Cultural	46%	45%
	Total	**45%**	**46%**

Si sus resultados en algunos de los ejes se encuentran sobre el promedio, puede considerarlos como fortalezas relativas y si son inferiores al promedio, constituyen debilidades relativas, que conviene reforzar.

Agradeciendo su motivación por postular e invitándolo desde ya a participar de nuevos procesos, se despide muy atentamente,



Rodolfo Bonifaz
Coordinador Nacional Área de Acreditación y Evaluación Docente
Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas

Imprimir Salir

Appendix C - Rewards

Figure C.1. The diploma.

Note: Retrieved from http://4.bp.blogspot.com/_KYQMVaNOfE0/SP5XBgiNY_I/AAAAAAAAU24/sieNcCh5olg/s1600-h/Diploma+AEP+2008+-+2.jpg

Figure C.2. The pin.

Note: Pin was provided by the Ministry of Education for the purpose of this dissertation

Figure C.3. The pin.

Note: Picture was retrieved from: http://www.educacionpaillaco.cl/noticia/2014/12/paillaco-cuenta-con-dos-nuevas-maestras-de-maestros.

Figure C.4. The pin and the diploma.

Note: Picture was retrieved from: http://www.eduglobal.cl/2013/07/04/mineduc-destaco-a-dos-profesores-que-lograron-acreditacion-de-excelencia-pedagogica/

Appendix D - Reward Process

Figure D.1. The reward ceremony

Note: Pictures were retrieved from http://www.lanacion.cl/mil-profesores-recibiran-asignacion-de-excelencia-pedagogica-en-2013/noticias/2013-04-09/143330.html%20%20%20%20

Figure D.2. Local award ceremonies

Note: Pictures were retrieved from http://deproveducaciontalagante.blogspot.com/2013/12/27-docentes-de-las-provincias-de.html http://colegiohispanoamericano.webescuela.cl/portfolio- item/fotos-del-inicio-de-clases-2014 http://www.voceroregional.cl/2014/11/24/docentes-de-la-provincia-de-valdivia-recibieron-distincion-aep/ http://educaciontecnologicaoctavos.blogspot.com/2008/10/aep-2008-asignacin-de-excelencia.html

Figure D.3. Media coverage.

Appendix E - Data

The Ministry of Education in Chile through the Statistics Department and the Center of Teaching Training and Improvement (CPEIP for the initials in Spanish) systematically collects the following datasets that are used in this research. Some of them were downloaded and others needed a special request to access and merge. The datasets were merged by school's, teacher's, and student's id correspondingly.

1. **STUDENTS ACADEMIC ACHIEVEMENT DATASET (SIMCE DATASET).** Database of students with their score in the national standardized test SIMCE. All students in either 4th, 8th, or 10th grade take the test.

Master Key: student id (KEY)

Secondary key: school id (RBD)

Tertiary key: level and classroom (curso-letra)

Reporting frequency: Annual Years: 2000-2013

Access: Via Information Request Form

2. **AEP DATASET.** Database of applicants for AEP identifies the teachers and the detailed results of the evaluation of the instruments. Importantly, this dataset includes the score obtained by the applicants. This score is a number going from 0 to 4 with two decimal numbers.

The dataset of the application year 2005 included a scores with six decimal numbers. Due to estimation issues, the score for that year was truncated to be with two decimal numbers as the other AEP datasets.

Master Key: teacher id (KEY)

Secondary key: school id (RBD)

Reporting frequency: Annual.

Years: 2002-2013

Access: Direct download from website.

3. **SCHOOLS, TEACHERS, CLASSROOMS, AND SUBJECTS DATASET (STCS DATASET).** Database of teachers working in public-funded schools identifying the school where they work, the classroom and subject they teach (e.g. teacher id 25, school id 123, 4th grade, classroom A, math). This data set is used to merge DATASETS 1 and 2.

Master Key: teacher id (KEY)

Secondary key: school id (RBD)

Thirdly key: level and classroom of application (curso-letra)

Reporting frequency: Annual.

Years: 2004-2012

Access: Via Information Request Form

4. **SIMCE TEACHERS SURVEY.** Yearly survey given to the teachers of the students who take the SIMCE. The dataset includes teachers' characteristics, teaching attitudes, students' attitudes towards learning, time distribution, school working environment and resources available, teachers' perception about the principal's work and relationship with the principal, among others.

Master Key: teacher id (KEY)

Secondary key: school id (RBD)

Tertiary key: level and classroom of application (curso-letra)

Reporting frequency: Annual

Years: 2002, 2009, 2010, and 2011

Access: Via Information Request Form

5. **TEACHER CENSUS DATASET.** Database of all teachers working in public-funded schools identifying the school where they work and their position, experience, among other teacher characteristics.

Master Key: teacher id (KEY)

Secondary key: school id (RBD)

Reporting frequency: Annual. (d) Years: 2003-2014

Access: Direct download from website

6. **SIMCE SCORES AT SCHOOL LEVEL (SCHOOL SIMCE DATASET).** Database of schools with their average SIMCE and their socioeconomic group as calculated by the Ministry of Education.

Master Key: school id (RBD)

Reporting frequency: Annual (c) Years: 2000-2013

Access: Direct download from website

Appendix F - Sample Statistics

Table F.1. Statistics in *t*, *t* + 1. *t* + 2, and *t* + 3.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	AEP Applicants	AEP Applicants teaching language	AEP Applicants teaching and applying for language			AEP Applicants teaching math	AEP Applicants teaching and applying for math		AEP Applicants teaching and applying for math or language		
Panel A. In t											
Application Year t	N	Teaching Year t	N	%	N	%	N	%	N	%	N
2003	935	2003									
2004	1,621	2004	855	53%	808	95%	804	50%	751	93%	1,559
2005	1,834	2005	633	35%	580	92%	630	34%	564	90%	1,144
2006	2,215	2006	686	31%	30	4%	648	29%	405	63%	435
2007	1,666	2007	597	36%	528	88%	579	35%	522	90%	1,050
2008	1,661	2008	530	32%	486	92%	496	30%	448	90%	934
2009	1,815	2009	501	28%	456	91%	479	26%	440	92%	896
2010	1,499	2010	467	31%	418	90%	447	30%	408	91%	826
2011	1,316	2011	379	29%	353	93%	403	31%	380	94%	733
TOTAL	16,468	TOTAL	4,648	28%	3,659	79%	4,486	27%	3,918	87%	7,577
Panel B. In t+1											
Application Year t	N	Teaching Year t+1	N	%	N	%	N	%	N	%	N
2003	935	2004	623	67%	552	89%	610	65%	533	87%	1,085
2004	1,621	2005	815	50%	760	93%	769	47%	712	93%	1,472
2005	1,834	2006	592	32%	538	91%	583	32%	519	89%	1,057
2006	2,215	2007	656	30%	36	5%	626	28%	384	61%	420
2007	1,666	2008	555	33%	480	86%	550	33%	481	87%	961
2008	1,661	2009	475	29%	432	91%	436	26%	401	92%	833
2009	1,815	2010	462	25%	420	91%	441	24%	401	91%	821
2010	1,499	2011	433	29%	383	88%	422	28%	383	91%	766
2011	1,316	2012	345	26%	322	93%	383	29%	352	92%	674
TOTAL	16,468	TOTAL	4,956	30%	3,923	79%	4,820	29%	4,166	86%	8,089
Panel C. In t+2											
Application Year t	N	Teaching Year t+2	N	%	N	%	N	%	N	%	N
2003	935	2005	600	64%	516	86%	586	63%	503	86%	1,019
2004	1,621	2006	758	47%	708	93%	710	44%	652	92%	1,360
2005	1,834	2007	576	31%	508	88%	547	30%	473	86%	981
2006	2,215	2008	582	26%	24	4%	562	25%	332	59%	356
2007	1,666	2009	516	31%	445	86%	496	30%	432	87%	877
2008	1,661	2010	454	27%	409	90%	421	25%	379	90%	788
2009	1,815	2011	440	24%	395	90%	430	24%	385	90%	780
2010	1,499	2012	395	26%	352	89%	402	27%	359	89%	711
2011	1,316	2013									
TOTAL	16,468	TOTAL	4,321	26%	3,357	78%	4,154	25%	3,515	85%	6,872
Panel D. In t+3											
Application Year t	N	Teaching Year t+3	N	%	N	%	N	%	N	%	N
2003	935	2006	568	61%	496	87%	549	59%	473	86%	969
2004	1,621	2007	711	44%	664	93%	663	41%	611	92%	1,275
2005	1,834	2008	515	28%	456	89%	500	27%	430	86%	886
2006	2,215	2009	568	26%	26	5%	519	23%	304	59%	330
2007	1,666	2010	483	29%	412	85%	475	29%	410	86%	822
2008	1,661	2011	424	26%	376	89%	404	24%	359	89%	735
2009	1,815	2012	411	23%	358	87%	393	22%	353	90%	711
2010	1,499	2013									
2011	1,316	2014									
TOTAL	16,468	TOTAL	3,680	22%	2,788	76%	3,503	21%	2,940	84%	5,728
Panel E. In t+1, t+2, t+3											
Application Year t	N	Year t+1/t+2/t+3	N	%	N	%	N	%	N	%	N
2003	935	2003/2004/2005	1791	192%	1564	87%	1745	187%	1509	86%	3,073
2004	1,621	2004/2005/2006	2284	141%	2132	93%	2142	132%	1975	92%	4,107
2005	1,834	2005/2006/2007	1683	92%	1502	89%	1630	89%	1422	87%	2,924
2006	2,215	2006/2006/2008	1806	82%	86	5%	1707	77%	1020	60%	1,106
2007	1,666	2007/2008/2009	1554	93%	1337	86%	1521	91%	1323	87%	2,660
2008	1,661	2008/2009/2010	1353	81%	1217	90%	1261	76%	1139	90%	2,356
2009	1,815	2009/2010/2011	1313	72%	1173	89%	1264	70%	1139	90%	2,312
2010	1,499	2010/2011/2012	828		735		824		742		1,477
2011	1,316	2011/2012/2013	345		322		383		352		674
TOTAL	16,468	TOTAL	12,957	79%	10,068	78%	12,477	76%	10,621	85%	20,689

Appendix G - Robustness Exercise. Covariate-adjusted Sharp Regression Discontinuity Estimates for Specification Checks.

Table G.1: Specification checks for teacher's, student's and school's characteristics in t. Covariate-adjusted SRD estimates using robust bias-corrected local linear regression.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}	(5) Robust 95% CI	(6) N_h	(7) N	(8) CI Length liff. w/ and w/o covs.
Panel A. Balance Variables								
Panel A.1 Teachers' characteristics								
Experience	-0.823	1.069	0.313	0.168	(-3.173, 1.017)	713	1,754	-55.2%
Female	-0.206	0.105	0.02	0.094	(-0.45, -0.039)	411	1,756	-0.2%
Age	0.502	2.38	0.828	0.149	(-4.147, 5.183)	624	1,756	0.3%
Spec. Degree	0.055	0.049	0.278	0.167	(-0.043, 0.149)	714	1,756	1.6%
Tenure	0.06	0.115	0.518	0.130	(-0.151, 0.299)	531	1,756	-1.7%
Metro. Region	0.077	0.099	0.409	0.179	(-0.112, 0.275)	750	1,756	1.8%
Rural	-0.006	0.077	0.989	0.2	(-0.151, 0.149)	865	1,756	-4.8%
# Schools by Teacher	0.042	0.059	0.449	0.18	(-0.071, 0.16)	750	1,756	-3.3%
# Classrooms by Teacher	0.132	0.438	0.996	0.155	(-0.857, 0.862)	677	1,756	1.4%
Teacher is a mover	-0.048	0.078	0.459	0.16	(-0.21, 0.095)	604	1,522	2.3%
Teacher is a leaver	0	0.028	0.832	0.178	(-0.05, 0.062)	671	1,522	3.7%
Panel A.2 Student's Characteristics								
Female	0.045	0.056	0.421	0.14	(-0.065, 0.156)	21,550	64,477	0.9%
Mother's Schooling>12	0.037	0.044	0.284	0.139	(-0.039, 0.134)	18,869	56,749	2.4%
Books at home>50	0.03	0.03	0.255	0.18	(-0.025, 0.093)	24,999	58,152	0.9%
Family Income Top Quintile	0.048	0.047	0.205	0.145	(-0.033, 0.152)	1,026	58,152	0.5%
Panel A.3 School's Characteristics								
Public School	-0.042	0.097	0.607	0.211	(-0.241, 0.141)	1142	2,244	-12.8%
Low-Medium SES	-0.029	0.122	0.899	0.152	(-0.255, 0.223)	869	2,244	1.1%
Panel B. Falsification Variables								
Topics Covered	0.014	0.062	0.989	0.17	(-0.12, 0.122)	792	1,807	-6.2%
Well Prepared	0.092	0.092	0.347	0.137	(-0.094, 0.266)	610	1,799	9.1%
Class Preparation>5	0.158	0.101	0.051	0.116	(-0.001, 0.394)	470	1,599	-0.5%
Test Score	0.038	0.122	0.658	0.172	(-0.185, 0.292)	27,795	64,477	0.0%

Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014); (iii) h_{MSE} corresponds to the-second generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (iv) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$; (v) standard errors (SE) are clustered at school level; (vi) teachers' age is included as covariate. Except in the case of age variable; (vii) application year, subject and grade dummies are included; (viii) the observation number (N) varies for several reasons. First, the variables in the table come from different datasets, which have different observations. For instance, test scores come from the SIMCE dataset, while teacher's characteristics come from administrative dataset. Second, variables coming from the same source have a different observation number because they were collected with different timings. For instance, the question regarding the number of hours spent preparing classes was part of the SIMCE teacher's survey in 2004, 2006 and 2010, while the question regarding topics covered was asked every year in the same survey; (ix) robust SE are estimated; and (x) column eight shows the increase rate of the confidence interval length if the program effect is estimated with and without covariates.

Table G.2: Attrition Test.

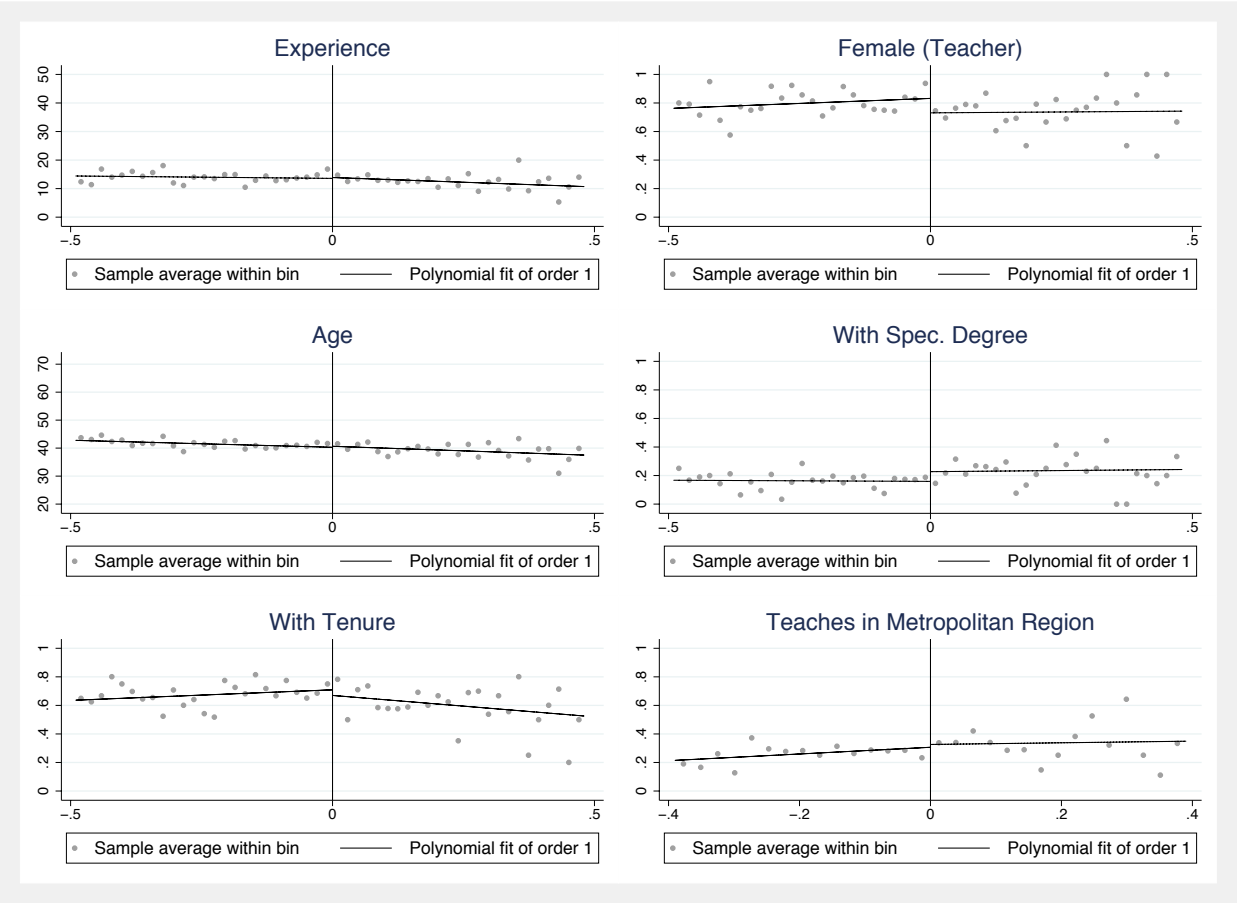
	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}	(5) Robust 95% CI	(6) N_h	(7) N
Panel A. Dependent Variables							
Probability of being assigned to a classroom that takes SIMCE in $(t + l)$ with l=1,2,3	-0.04	0.043	0.317	0.203	(-0.126, 0.041)	3,311	7,478

Notes: (i) the reported RD treatment effects are estimated from a local linear regression between the probability of finding a SIMCE test scores in years t + 1, t + 2 or t + 3 for a teacher who teaches math/language in those years and apply for math/language certification in t and the distance to the cutoff; (ii) the reported coefficient is estimated with a single regression between the AEP score ($\tilde{S}_t$) and dummy variable that takes the value 1 if the math/language teacher was assigned to a classroom that takes the math/reading SIMCE test in in years t + 1, t + 2 or t + 3, if $\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h_{MSE}$(certified teachers within the MSE bandwidth estimated above) and $\tilde{S}_t - h_{MSE} \leq S_{jt} < \tilde{S}_t$ (non-certified teachers within the MSE bandwidth); (iii) teachers' age is included as covariate; (iv) application year, subject and grade dummies are included; (v) standard errors (SE) are clustered at school level; (vi) robust SE are estimated; (vii) point estimators are constructed using local polynomial estimators with triangular kernel; (viii) robust p-values are constructed using bias correction with robust standard errors as derived in Calonico et al. (2014); (ix) h_{MSE} corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (x) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$.

Appendix H - Graphical Analysis. Regression Discontinuity Plots. Specification Checks.

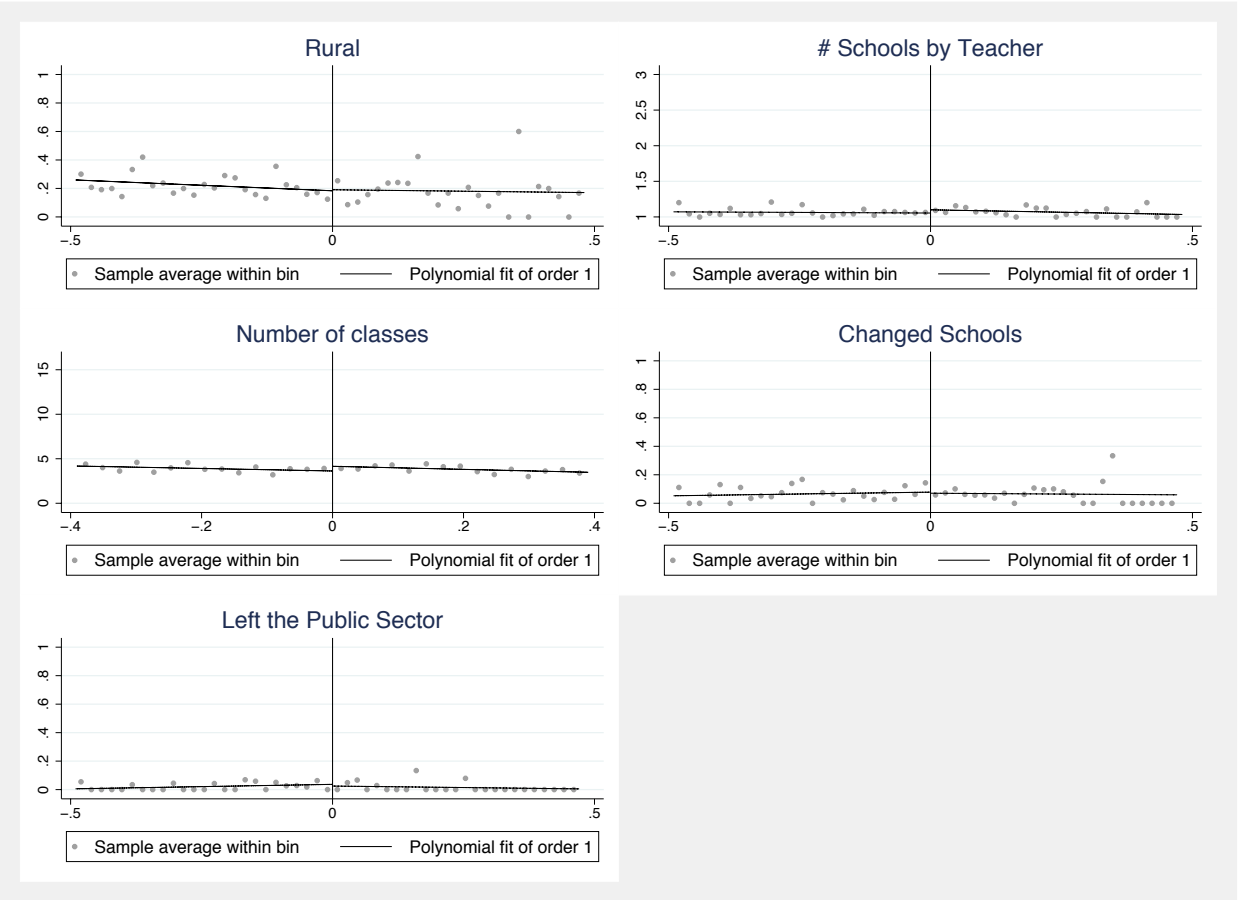
H.1. Balance Checks.

Figure H.1. RD plots for specification checks for teacher characteristics in t. Part I.



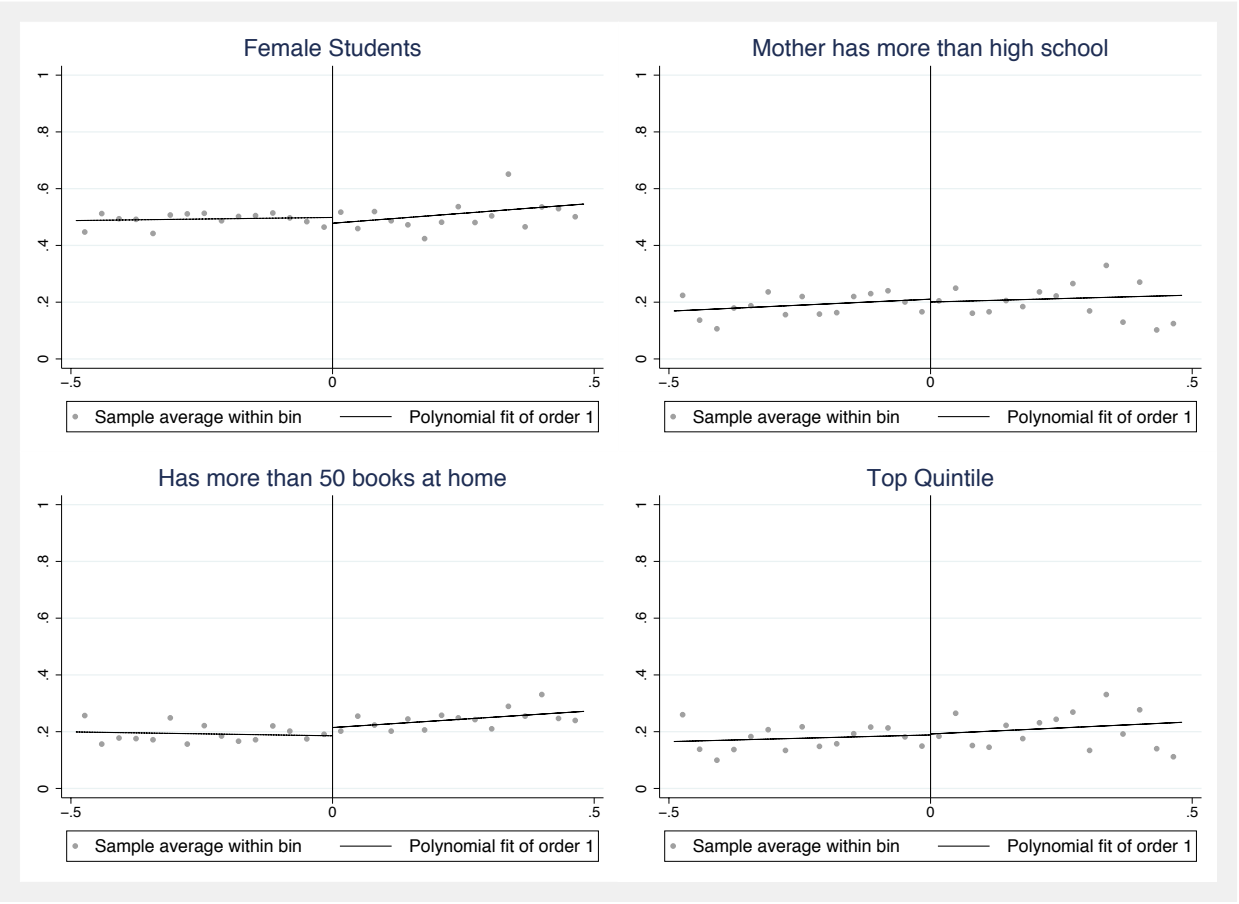
Notes: (i) data-driven RD plots using evenly spaced 25 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.5 points of a cutoff.

Figure H.2. RD plots for specification checks for teacher characteristics in t. Part II.



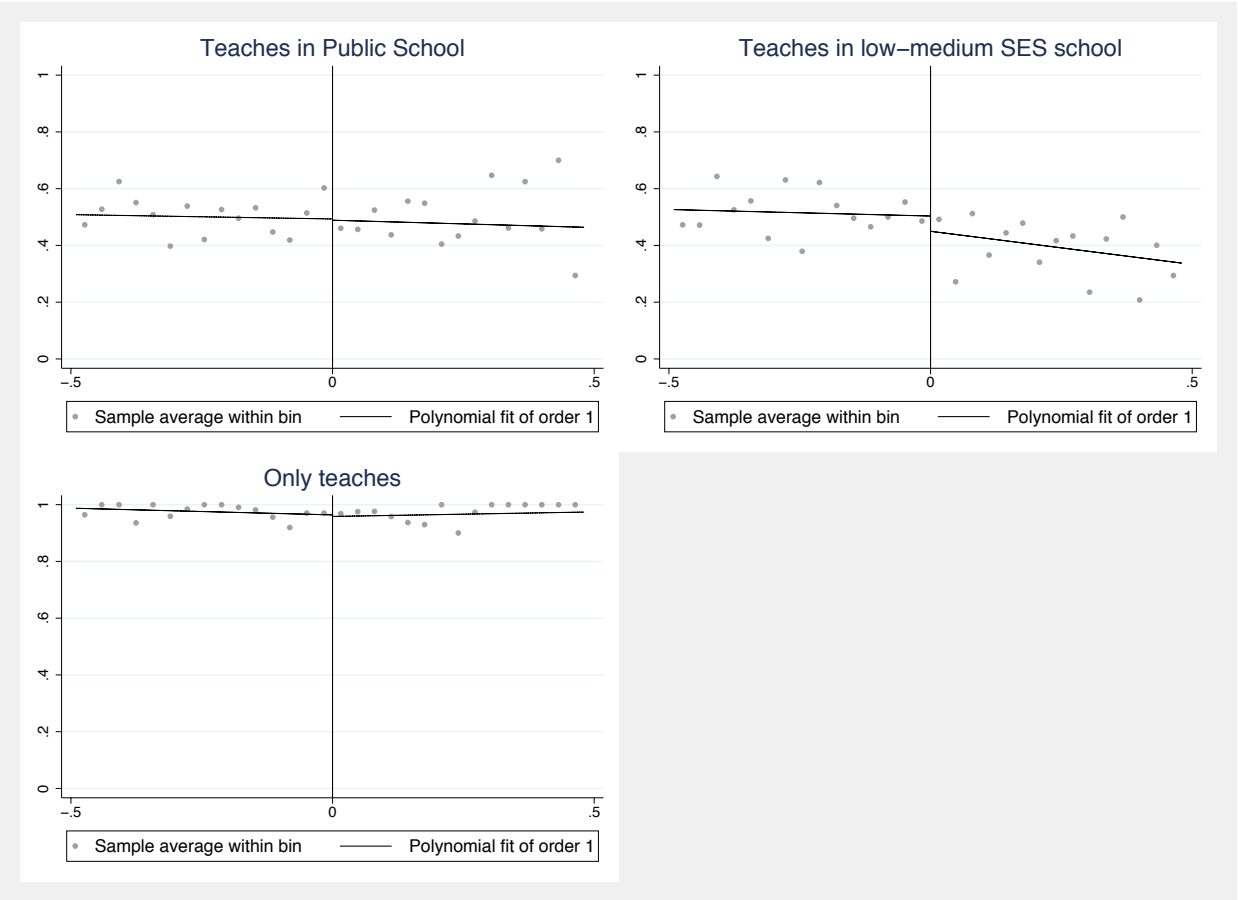
Notes: (i) data-driven RD plots using evenly spaced 15 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.4 points of a cutoff.

Figure H.3. RD plots for specification checks for student characteristics in t.



Notes: (i) data-driven RD plots using evenly spaced 15 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.5 points of a cutoff.

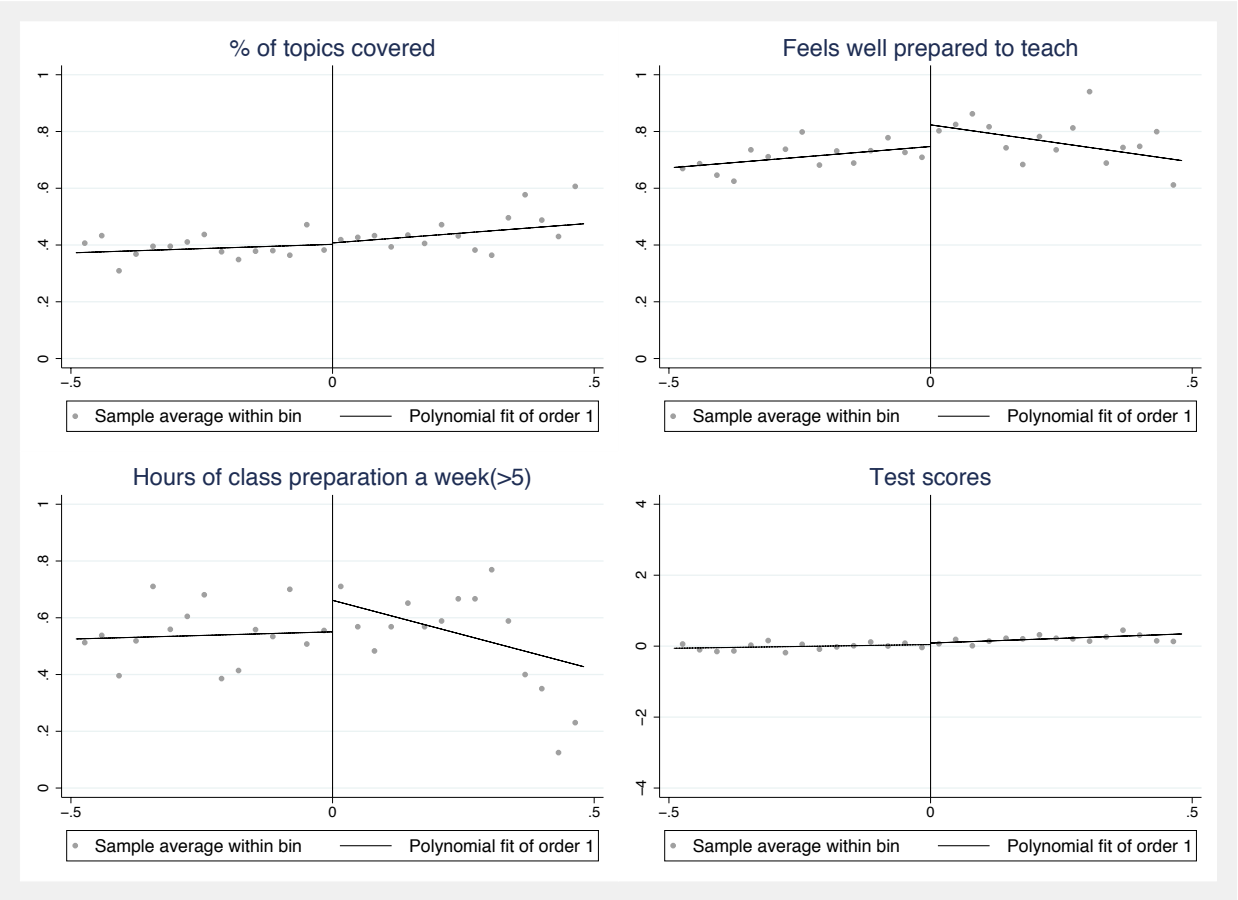
Figure H.4. RD plots for specification checks for school characteristics in t.



Notes: (i) data-driven RD plots using evenly spaced 15 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.5 points of a cutoff.

H.2. Falsification Tests

Figure H.5. RD plots for specification checks. Falsification variables in t.



Notes: (I) data-driven RD plots using evenly spaced 15 bins on each side of the cutoff; (ii) solid lines depict linear polynomial fits using control and treated units separately; (iii) dots depict sample average of outcome variable within each bin; and (iv) all panels are based on administrative data for the 2003-2011 applicant cohorts and restrict observations to individuals with AEP scores within 0.5 points of a cutoff.

Appendix I - Robustness Exercise. Sharp Regression Discontinuity (SRD) Estimates for Outcome Variable Analysis.

I.1. SRD estimates using the robust specification and covariates

Table I.1. Final program effect on standardized test scores in $t + l$. Covariate-adjusted SRD estimates using robust bias-corrected local linear regression with covariate.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}	(5) Robust 95% CI	(6) N_h	(7) N	(8) CI length diff. w/ and w/o covs.
Panel A. Final Outcomes								
Panel A.1 FULL PROGRAM								
Test Score if l=1,2,3	0.095	0.09	0.192	0.202	(-0.059, 0.294)	63,916	145,211	4.7%
Test Score if l=1	0.07	0.129	0.495	0.182	(-0.165, 0.341)	22,845	54,713	2.2%
Test Score if l=2	0.086	0.137	0.45	0.208	(-0.165, 0.372)	23,057	52,103	-0.7%
Test Score if l=3	0.112	0.152	0.368	0.224	(-0.161, 0.434)	17,367	38,395	1.4%
Panel A.2 FINANCIAL COMPONENT								
Test Score if l=1,2,3	-0.022	0.185	0.908	0.132	(-0.384, 0.341)	17,230	50,921	1.5%
Test Score if l=1	-0.088	0.21	0.569	0.128	(-0.531, 0.292)	5,463	16,935	-1.7%
Test Score if l=2	-0.158	0.238	0.542	0.105	(-0.611, 0.321)	4,778	18,904	2.2%
Test Score if l=3	0.342	0.242	0.111	0.131	(-0.089, 0.858)	5,139	15,082	-4.7%
Panel B. Placebo cutoff values: -0.2								
Panel B.1 FULL PROGRAM								
Test Scores if l=1,2,3	0.017	0.176	0.843	0.056	(-0.31, 0.38)	11,452	50,418	6.2%
Panel B.2 FINANCIAL COMPONENT								
Test Scores if l=1,2,3	0.186	0.214	0.263	0.031	(-0.18, 0.658)	1,922	14,325	-16.4%
Panel C. Placebo cutoff value: +0.25								
Panel C.1 FULL PROGRAM								
Test Scores if l=1,2,3	0.005	0.158	0.945	0.062	(-0.298, 0.32)	16,720	80,404	-0.8%
Panel C.2 FINANCIAL COMPONENT								
Test Scores if l=1,2,3	0.137	0.233	0.409	0.053	(-0.265, 0.65)	11,452	50,418	1.8%

Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014) ; (iii) h_{MSE} corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (iv) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$; (v) robust standard errors (SE) are clustered at school level; (vi) teachers' age is included as covariate; (vii) application year, subject and grade dummies are included; (viii) the observation number (N) varies for reasons such as data availability and changes in surveys; (ix) in Panel A each row reports coefficients from a single regression restricting the sample by the value of l =1, 2, 3. Then rows for test scores if l =1, 2 or 3 stand for the RD program effect on next-year, two-year-later, and three-year-later outcomes, respectively; (x) test scores are normalized within each grade and year to have a mean of 0 and a standard deviation of 1; (xi) the estimates for placebo cutoffs restrict the data to only certified teachers when the fake cutoff is lower than the actual cutoff (-0.2), and only non-certified teachers in the other case (+0.25); and (xii) column eight shows the increase rate of the confidence interval length if the program effect is estimated with covariates versus no covariates.

Table I.2. Intermediate program effect on math teachers' behavior in $t + l$. Covariate-adjusted SRD estimates using robust bias-corrected local linear regression.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}	(5) Robust 95% CI	(6) N_h	(7) N	(8) CI length diff. w/ and w/o covs.
Behavioral Outcomes								
Panel A. FULL PROGRAM								
Well Prepared	0.047	0.05	0.43	0.227	(-0.059, 0.138)	1,647	3,477	1.0%
Topics Covered	0.021	0.054	0.822	0.18	(-0.093, 0.118)	1,378	3,485	-3.7%
Class Preparation>2	0.013	0.084	0.741	0.182	(-0.137, 0.193)	807	2,147	-0.3%
Class Preparation>5	0.048	0.111	0.502	0.172	(-0.143, 0.291)	774	2,147	0.0%
Class Preparation>7	0.015	0.096	0.696	0.199	(-0.151, 0.225)	848	2,147	0.0%
Panel B. FINANCIAL COMPONENT								
Well Prepared	-0.074	0.103	0.356	0.126	(-0.297, 0.107)	369	1,225	0.2%
Topics Covered	-0.154	0.093	*0.059*	0.138	(-0.358, 0.007)	420	1,226	0.8%
Class Preparation>2	-0.041	0.066	0.534	0.146	(-0.171, 0.089)	513	1,433	4.0%
Class Preparation>5	0.132	0.14	0.214	0.127	(-0.1, 0.448)	444	1,433	2.4%
Class Preparation>7	0.221	0.124	*0.041*	0.163	(0.01, 0.497)	578	1,433	-1.0%

Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014) ; (iii) h_{MSE} corresponds to the-second generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (iv) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$; (v) standard errors (SE) are clustered at school level; (vi) teachers' age is included as covariate; (vii) application year, subject and grade dummies are included; and (viii) the observation number (N) varies for several reasons. First, the variables in the table come from different datasets, which have different observations. For instance, test scores come from the SIMCE dataset, while teacher's characteristics come from administrative dataset. Second, variables coming from the same source have a different observation number because they were collected with different timings. For instance, the question regarding the number of hours spent preparing classes was part of the SIMCE teacher's survey in 2004, 2006 and 2010, while the question regarding topics covered was asked every year in the same survey. (vii) Robust SE are estimated; and (ix) column eight shows the increase rate of the confidence interval length if the program effect is estimated with covariates versus no covariates.

Table I.3. Intermediate program effect on student and school characteristics in $t + l$.

Covariate-adjusted SRD estimates using robust bias-corrected local linear regression.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	RD treatment effect	**Robust SE**	**Robust p-value**	h_{MSE}	**Robust 95% CI**	N_h	**N**	**CI length diff. w/ and w/o covs.**
Alternative Mechanism								
Panel A. FULL PROGRAM								
Panel A.1 Student's Characteristics								
Female Student	0.002	0.03	0.89	0.294	(-0.056, 0.064)	85,791	145,211	-1.6%
Mother's Schooling>12	0.032	0.034	0.279	0.203	(-0.03, 0.103)	57,853	129,911	-5.7%
Books at home>50	0.014	0.024	0.488	0.213	(-0.03, 0.064)	60,792	130,822	0.0%
Family Income Top Quintile	0.039	0.041	0.246	0.199	(-0.033, 0.127)	55,947	130,822	-3.6%
Panel A.2 School's Characteristics								
Public School	-0.088	0.08	0.276	0.236	(-0.244, 0.07)	2,143	4,298	-5.4%
Low-Medium SES	-0.091	0.083	0.184	0.253	(-0.271, 0.052)	2,302	4,299	-3.3%
Panel A.3 Teacher's Characteristics								
Teacher is a school-mover	0.02	0.061	0.712	0.208	(-0.098, 0.143)	1,353	2,951	1.7%
Panel B. FINANCIAL COMPONENT								
Panel B.1 Student's Characteristics								
Female	0.031	0.039	0.319	0.117	(-0.037, 0.115)	14,522	50,921	0.0%
Mother's Schooling>12	-0.076	0.077	0.233	0.127	(-0.242, 0.059)	13,407	43,638	4.9%
Books at home>50	-0.004	0.033	0.754	0.133	(-0.076, 0.055)	15,682	45,899	-0.8%
Family Income Top Quintile	-0.051	0.068	0.335	0.132	(-0.2, 0.068)	15,682	45,899	3.5%
Panel B.2 School's Characteristics								
Public School	-0.109	0.147	0.477	0.174	(-0.392, 0.183)	2,143	4,298	-2.7%
Low-Medium SES	0.182	0.202	0.276	0.108	(-0.175, 0.615)	2,302	4,299	6.5%
Panel B.3 Teacher's Characteristics								
Teacher is a school-mover	-0.052	0.103	0.496	0.157	(0, -0.273)	538	1354	0.2%

Notes: (i) point estimators are constructed using local polynomial estimators with triangular kernel; (ii) robust p-values are constructed using bias-correction with robust standard errors as derived in Calonico et al. (2014) ; (iii) h_{MSE} corresponds to the second-generation data-driven MSE optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b); (iv) N is total number of observations while $N_h = N_h^+ + N_h^-$ where $N_h^- = \sum_{i=1}^{n} 1(\tilde{S}_t - h \leq S_{jt} < \tilde{S}_t)$, $N_h^+ = \sum_{i=1}^{n} 1(\tilde{S}_t \leq S_{jt} < \tilde{S}_t + h)$;;(v) standard errors (SE) are clustered at school level; (vi) teachers' age is included as covariate; (vii) application year, subject and grade dummies are included; (viii) the observation number (N) varies for reasons such as data availability and changes in surveys; (ix) robust SE are estimated; and (x) column eight shows the increase rate of the confidence interval length if the program effect is estimated with covariates versus no covariates.

I.2 SRD estimates using alternative specification with covariate

Table I.4. Final program effect on standardized test scores in $t + l$. Parametric RD method. Covariate-adjusted SRD estimates using alternative local linear regression.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}
Panel C. Final Outcomes				
Panel C.1 FULL PROGRAM				
Test Score if l=1,2,3	0.09	0.079	0.256	0.202
Test Score if l=1	0.066	0.114	0.561	0.182
Test Score if l=2	0.077	0.115	0.503	0.208
Test Score if l=3	0.111	0.132	0.4	0.224
Panel C.2 FINANCIAL COMPONENT				
Test Score if l=1,2,3	-0.026	0.158	0.871	0.132
Test Score if l=1	-0.094	0.191	0.624	0.128
Test Score if l=2	-0.156	0.204	0.447	0.105
Test Score if l=3	0.296	0.215	0.17	0.131
Panel C.3 NON-FINANCIAL COMPONENT				
Test Score if l=1,2,3	0.124	0.158	0.436	0.202
Test Score if l=1	0.131	0.207	0.528	0.182
Test Score if l=2	0.238	0.203	0.242	0.208
Test Score if l=3	-0.08	0.219	0.717	0.224

Notes: (i) all estimators are constructed using linear ordinary least-squares with robust standard errors (SE); (ii) h_{MSE} corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b) that is recovered from the estimation of the program effect on the corresponding outcome using a robust bias-corrected local linear regression with triangular kernel; (iii) robust SE are clustered at school level; (iv) teachers' age is included as covariate; and (vi) application year, subject and grade dummies are included.

Table I.5. Intermediate program effect on teachers' behavior in $t + l$. Covariate-adjusted SRD estimates using alternative local linear regression.

	(1) RD treatment effect	(2) SE	(3) p-value	(4) h_{MSE}
Panel A. Behavioral Outcomes				
Panel A.1 FULL PROGRAM				
Well Prepared	0.048	0.044	0.272	0.227
Topics Covered	0.019	0.045	0.674	0.18
Class Preparation>2	0.01	0.073	0.89	0.182
Class Preparation>5	0.049	0.096	0.607	0.172
Class Preparation>7	0.016	0.082	0.843	0.199
Panel A.2 FINANCIAL COMPONENT				
Well Prepared	-0.065	0.088	0.459	0.126
Topics Covered	-0.145	0.079	*0.066*	0.138
Class Preparation>2	-0.04	0.057	0.482	0.146
Class Preparation>5	0.132	0.126	0.296	0.127
Class Preparation>7	0.22	0.111	0.047	0.163
Panel A.3 NON-FINANCIAL COMPONENT				
Well Prepared	0.065	0.084	0.442	0.227
Topics Covered	0.167	0.082	*0.042*	0.18
Class Preparation>2	0.051	0.09	0.573	0.182
Class Preparation>5	0.014	0.144	0.923	0.172
Class Preparation>7	-0.162	0.13	0.214	0.199

Notes: (i) all estimators are constructed using linear ordinary least-squares with robust standard errors (SE); (ii) h_{MSE} corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b) that is recovered from the estimation of the program effect on the corresponding outcome using a robust bias-corrected local linear regression with triangular kernel; (iii) robust SE are clustered at school level; (iv) teachers' age is included as covariate; and (vi) application year, subject and grade dummies are included.

Table I.6. Intermediate student, school and teacher characteristics in $t + l$. Covariate-adjusted SRD estimates using alternative local linear regression.

	(1) RD treatment effect	(2) Robust SE	(3) Robust p-value	(4) h_{MSE}
Panel B. Alternative Mechanism				
Panel B.1 FULL PROGRAM				
Panel B.1.1 Student's Characteristics				
Female Student	0.001	0.026	0.966	0.294
Mother's Schooling>12	0.031	0.029	0.289	0.203
Books at home>50	0.013	0.021	0.52	0.213
Family Income Top Quintile	0.038	0.036	0.282	0.199
Panel B.1.2 School's Characteristics				
Public School	-0.091	0.066	0.169	0.236
Low-Medium SES	-0.093	0.07	0.189	0.253
Panel B.1.3 Teacher's Characteristics				
Teacher is a mover	0.021	0.052	0.689	0.208
Panel B.2 FINANCIAL COMPONENT				
Panel B.2.1 Student's Characteristics				
Female	0.027	0.034	0.423	0.117
Mother's Schooling>12	-0.066	0.063	0.295	0.127
Books at home>50	-0.011	0.031	0.728	0.133
Family Income Top Quintile	-0.049	0.057	0.396	0.132
Panel B.2.2 School's Characteristics				
Public School	-0.105	0.124	0.398	0.174
Low-Medium SES	0.194	0.173	0.263	0.108
Panel B.2.3 Teacher's Characteristics				
Teacher is a mover	-0.051	0.088	0.56	0.157
Panel B.3 NON-FINANCIAL COMPONENT				
Panel B.3.1 Student's Characteristics				
Female	-0.026	0.036	0.48	0.294
Mother's Schooling>12	0.074	0.059	0.212	0.203
Books at home>50	0.009	0.033	0.795	0.213
Family Income Top Quintile	0.067	0.063	0.29	0.199
Panel B.3.2 School's Characteristics				
Public School	0.033	0.124	0.792	0.236
Low-Medium SES	-0.128	0.132	0.334	0.253
Panel B.3.3 Teacher's Characteristics				
Teacher is a mover	0.072	0.093	0.44	0.208

Notes: (i) all estimators are constructed using linear ordinary least-squares with robust standard errors (SE); (ii) h_{MSE} corresponds to the second-generation data-driven MSE-optimal bandwidth selector proposed in Calonico et al. (2014) and Calonico et al. (2016b) that is recovered from the estimation of the program effect on the corresponding outcome using a robust bias-corrected local linear regression with triangular kernel; (iii) robust SE are clustered at school level; (iv) teachers' age is included as covariate; and (vi) application year, subject and grade dummies are included.