Cash for Grades or Money for Nothing? Evidence from Regression Discontinuity Designs

Cristian Crespo

LSE Department of Social Policy

The Department of Social Policy is an internationally recognised centre of research and teaching in social and public policy. From its foundation in 1912 it has carried out cutting edge research on core social problems, and helped to develop policy solutions.

The Department today is distinguished by its multidisciplinarity, its international and comparative approach, and its particular strengths in *behavioural public policy, criminology, development, economic and social inequality, education, migration, non-governmental organisations* (NGOs) *and population change and the lifecourse.*

The Department of Social Policy multidisciplinary working paper series publishes high quality research papers across the broad field of social policy.

Abstract

This paper estimates the impact of a Chilean cash for grades programme, the *Bono por Logro Escolar (BLE)* in 2013, on future educational outcomes. The cash transfer was targeted using two scores from 2012, an income index and academic performance. I implement a sharp regression discontinuity design along these two running variables. I show that students marginally at each side of the two thresholds used only differed in receiving the BLE in 2013. The main causal estimates for the outcomes are not statistically significantly different from zero. Additionally, the main causal estimates are centred around zero and their standard errors are small. If a local average effect of the BLE in 2013 exists this is at best modest in magnitude. Any potential impact of the BLE in 2013 would have been at least smaller than those found in developing countries, where effects of at least 0.17 of a standard deviation on test scores have been observed.

Key words: cash for grades, regression discontinuity, *bono por logro escolar*, cash transfers

Author

Cristian is currently a PhD candidate at LSE. His thesis studies questions located at the intersection of conditional cash transfers (CCTs), targeting mechanisms of CCTs, and educational outcomes. Cristian has experience as a program/policy evaluation consultant for universities, governmental agencies, and non-profit organizations. Additionally, he served as the Head of the Research Department in the Ministry of Social Development in Chile.

Introduction

Should we pay children to learn? Few people are indifferent in regard to this controversial question. Finding a response to this question is particularly relevant for the United States where, by 2008, at least twelve states had adopted schemes where primary and secondary school students were rewarded with money for obtaining good grades or test scores or passing exams (Toppo, 2008).

Over the last eleven years, the public debate around this issue has been noticeable. Several articles have been published in the United States' mainstream newspapers (Calefati, 2008; Guttenplan, 2011; Higgins, 2015; Roberts, Becker, & Ibanga, 2008). The discussion seemed to reach its peak in 2010. During that spring the question: should schools bribe kids? and the corresponding reply "it can work, if it is done right" made it to the cover of Time Magazine (Ripley, 2010), while in the autumn of the same year, the Annual Poll of the Public's Attitudes Toward the Public Schools revealed that 76% of United States' adults opposed the idea of school districts paying small amounts of money to students to read books or get good grades (Bushaw & Lopez, 2010).[1]

How cash for grades programmes work is also a matter of active deliberation. Kremer, Miguel, and Thornton (2009), Gneezy, Meier, and Rey-Biel (2011) and Fryer (2011) offer syntheses about how these policies operate. On the one hand, incentives provide a price effect that makes specific behaviours more attractive. These economic effects are expected to increase individuals' effort and performance. On the other hand, a psychological effect potentially exists. This could undermine intrinsic motivation, negatively affecting the desired behaviours, especially after the reward is removed.[2]

The evidence regarding the effectiveness of cash for grades schemes is mixed. In Ohio, an effect of 0.15σ is estimated for maths but no impact is observed in three other subjects (Bettinger, 2012). In Kenya, an effect of 0.19σ in test scores and other positive externalities have been found (Kremer et al., 2009). In Israel, the results are positive and statistically significant only for girls (Angrist & Lavy, 2009). In New York, mostly no effects have been observed (Fryer, 2011; Riccio et al., 2013). In Mexico, effects from 0.17σ to 0.30σ in maths test scores have been estimated but these are partly explained by students' cheating (Behrman, Parker, Todd, & Wolpin, 2015). In lab experiments, characterised by immediate but lower rewards, some positive effects have been found in Chicago (Levitt, List, Neckermann, & Sadoff, 2012) and India (Hirshleifer, 2017).[3]

---

[1] Educational philosophers have not been able to agree on this topic either. To illustrate, Sidorkin (2007, 2009) argues that moral reasons exist for paying students. He suggests that, at least for low-income students, most of the economic value of their education is received by the society, not themselves. Accordingly, for these students, schooling does not represent a consumer good but a form of labour that deserves fair compensation. Conversely, Warnick (2017) questions cash for grades programmes on ethical grounds. His main argument is that cash incentives are uniquely corruptive of key educational aims and values. Specifically, he claims that these schemes hinder the development of students' self-control, promote advancing private versus public ends and additionally reinforce a perception of students as market actors.

[2] This last argument is quite contested in psychology. After conducting a meta-analysis, a group of authors conclude that, generally, extrinsic rewards are not harmful to motivation (Cameron, 2001; Cameron, Banko, & Pierce, 2001; Cameron & Pierce, 1994). However, Deci, Koestner and Ryan (1999; 2001) and Kohn (1999) have claimed for substantial weakening effects after reassessing evidence on this topic.

[3] I provide more information about the design and results of all these interventions in Appendix A. All these evaluations are randomised control trials. In all cases, randomisation did not occur at the individual level.

My paper adds to this body of knowledge. Specifically, I assess the effect of a cash for grades intervention in Chile on subsequent educational outcomes. The *Bono por Logro Escolar* consists of a one-off cash transfer of up to $50,000 CLP ($100 USD approximately) given to high achieving students from fifth to twelfth grade. The programme intends to provide an incentive or reward for students' effort and overall academic achievement. My assessment is mostly done for the first version of the BLE, which was implemented in the middle of the 2013 academic year (and used information from 2012 to determine eligibility). Using large and rich administrative datasets, the outcomes I analyse are attendance and average grade for the academic years 2013 and 2014.

I use a sharp regression discontinuity (RD) design to recover the causal estimates. To be eligible for the BLE, students needed to belong to the poorest 30% of the population and be in the top 30% in terms of performance within their cohort. I compare students just below and above each of these two thresholds. Students at one side or the other of each threshold only differ in whether they receive the BLE. Thus, any differences in outcomes can be attributed to the programme. This empirical strategy is suitable as the BLE was implemented retrospectively. Students did not know that information from 2012 would determine whether they would be a recipient in 2013.

Each main causal estimate is not statistically significantly different from zero for both types of outcomes. Additionally, the main causal estimates are centred around zero and their standard errors are small. As a result, each 95% confidence interval contains values of a small magnitude exclusively. To illustrate, the highest upper bound I find for any 95% confidence interval is 0.035 (0.056σ) for average grade and 0.49% (0.052σ) for attendance. In practical terms, these estimates are near a third of the minimum distance between reported grades in the country's educational system and are equivalent to a day of school attendance within an academic year.

If a local average effect of the BLE in 2013 exists this is at best modest in magnitude. Therefore, I am unable to detect it with statistical certainty. Additionally, the analysis by subgroups does not consistently show estimates that are statistically significantly different from zero. These results cannot be generalised for the entire population who received the BLE in 2013 given that the RD estimates are only valid for students near the two thresholds used to target the BLE. Among these students, any potential impact of the BLE in 2013 would have been at least smaller than those found in developing countries, where effects of at least 0.17σ on test scores have been observed.

A possible reason for these results is that the programme was not very salient for the population. If children and adolescents were unaware of the implementation of the BLE then it would not be expected to observe changes in their behaviour. An alternative explanation is that students and their families were aware of the cash transfer but unresponsive to its size. Another potential cause of these results is that the BLE provided two types of effects that cancelled each other out overall.

The rest of the paper is structured as follows. The second section briefly describes the design and implementation of the *Bono por Logro Escolar*. The third section introduces the data sources, provides summary statistics and explains the empirical strategy. The fourth section presents the results of the impact assessment. The last section discusses the results and main implications.

Programme Description

The design of the *Bono por Logro Escolar* attempted to provide a cash transfer for high achieving students in primary and secondary schools belonging to the poorest segments of the population. Only students enrolled between the fifth and twelfth grades and not older than 24 years old in year $t-1$ were eligible to receive the cash transfer in year t. According to the Chilean educational system this meant that, conditional on age, only students in the last four years of primary school (fifth through eighth grade) and all students in secondary education were eligible in principle.[4]

In addition, to be eligible for the programme in year t students needed to belong to the poorest 30% of the population. The PFSE index measured this in year $t-1$.[5] In practice, the threshold used in this relative income index was 98 points. Accordingly, to be eligible for the BLE in 2013, students needed to have a PFSE score in the year 2012 equal to or lower than 98 points.

The academic performance requisite needed to receive the BLE in year t depended on students' average grade in year $t-1$. Students needed to be in the top 30% of their cohort to be eligible for the BLE.[6] Within each cohort the students were ranked. The number one was assigned to the student with the highest average grade. The average grade in Chile ranges from a minimum of one to a maximum of seven, is generated within each school, and is reported to the central level using only one decimal place. Ties were allowed in the ranking. For example, if two students had the same average grade and this grade was the highest in their cohort both received a value of one in the rankings. Also, in this scenario, the student or students with the second highest average grade received the third position. To be eligible for the BLE in year t the proportion between the student ranking and his or her cohort size in year $t-1$, the relative ranking, had to be no higher than 0.3.

In 2013, the BLE provided a lump sum of $50,000 CLP ($100 USD approximately) for students in the top 15% of highest performance and $30,000 CLP for students within the 15% and 30% range. The size of the cohort needed to be at least seven students for this last rule to hold. If the cohort size was between two and six, only the first-ranked student in the cohort was eligible.

The Ministry of Social Development first delivered the cash transfer in July 2013. Given that the Chilean academic year starts in March and ends in December, the BLE in 2013 was implemented in the middle of the academic year. Thereafter, the BLE has been given once per year with payments occurring between September and November. When it first started in 2013, the programme was implemented retrospectively. The rules regarding eligibility were established after December 2012, which marked the conclusion of the 2012 academic year.

---

[4] The Decree number 24 of the Chilean Ministry of Social Development, published in June 2013, regulates general aspects of the cash transfer (Biblioteca del Congreso Nacional de Chile, 2013).

[5] The PFSE index score was the result of the combination of a proxy means test and a means test. The former variable was the Social Protection File score. This proxy means test score provided the relative position of Chilean households regarding income by needs. Not having a Social Protection File score translated automatically in not being eligible for the BLE. The other variable in the PFSE formula was income per capita. The Ministry of Social Development built this variable from multiple administrative datasets.

[6] Three variables define an academic cohort: i) the school, ii) the type of education provided by that school (for example traditional or adult education, scientific-humanistic or technical-professional), and iii) the grade in which the students were enrolled. Students belonging to the same cohort have these three characteristics in common.

Until 2014, the BLE was paid in cash. Beneficiaries needed to collect their payments in person from local government agencies. From 2015, the BLE progressively adopted bank transfers. The BLE is paid to the student if they are at least 18 years old. Otherwise, a member of the student's household, most likely the mother, receives the payment. In the Chilean educational system, students are, in theory, expected to graduate from their secondary studies at the age of 18. Hence, the majority of BLE payments are not received by the students but by another household member.

The conception of the BLE is related to the *Seguridades y Oportunidades* law approved in May 2012. Within this large piece of legislation, the *Bono por Esfuerzo* (Cash Payment for Effort) was created. This law establishes that the State can provide conditional cash transfers in diverse areas of social policies to foster social achievements. The *Bono por Logro Escolar* (Cash Payment for Student Achievement) was created in 2013 following this logic. Consequently, the public discourse from politicians and public managers about the goals of the BLE has been that the programme is a tool to appreciate, incentivise or reward students' achievement and effort. This logic has been unaltered despite successive changes of coalitions in government.

I assess whether receiving the BLE in 2013 impacted attendance and academic performance in 2013 and 2014. I analyse the programme in 2013 because I can guarantee that students were not aware that their academic performance in 2012 would have an impact on whether they received the BLE in 2013. This is not necessarily true for the programme in 2014 (which was implemented using information from the academic year 2013) and for its later versions.

The research questions respond to the design, implementation and goals of the BLE but also the literature on cash for grades. Students who received the BLE in 2013 were expected to increase their effort (which, in this paper, is measured through attendance) and future academic performance to receive it again in the future. Additionally, the cash transfers could have had an effect through investments leading to better educational outcomes. From the psychological standpoint, a common concern related to these types of programmes is that they could have an adverse effect through decreased motivation after the incentive is withdrawn. Given that the BLE has remained through the years there is less support for an argument of this kind.

Data and Methods

This section describes the data and methods. The first subsection introduces the data. The second part of the section explains how I structure the dataset for the analysis. The third subsection provides descriptive statistics. Finally, the fourth part discusses the methodological approach I use to recover the causal estimates of the BLE, the regression discontinuity design.

Data

The Ministry of Social Development (MSD) provided most of the datasets I use in this paper. I combine the datasets using the individual ID number provided by the Chilean State. For privacy purposes the ID numbers were changed by the MSD using an algorithm that is unknown to me. The three primary sources of information for this research are as follows:

Bono por Logro Escolar Dataset

Created since 2013 and replicated annually by the MSD, this dataset contains information for all students between the fifth and twelfth grades in year $t−1$. The dataset excludes students in flexible adult and differential education. Each dataset has approximately 1,900,000 students. I requested two versions of this dataset (the years 2013 and 2014) for this paper. Some variables available in this dataset are: i) school ID, ii) type of school (with categories such as traditional primary education, scientific-humanistic or technical-professional secondary education), iii) grade, iv) average grade, v) attendance, vi) student ID, vii) age, viii) student ranking in the cohort, ix) cohort size, and x) PFSE score. All these variables refer to year $t−1$. Another key variable in this dataset denotes whether the student was a recipient of the BLE in year t.

Ministry of Education (ME) Performance Dataset

This dataset is created each year by the Ministry of Education, which later shares it with the MSD. The dataset contains information for all students who finish the academic year from the first through to the twelfth grade (except for flexible adult and differential education). Each version of this dataset has approximately 2,950,000 students. I have at my disposal eight datasets (from 2009 until 2016). The variables available in this dataset are the same as in points i) to vi) of the BLE Dataset.[7] More educational information at the school level is available from public sources. Using the variable school ID, as a key to merge, I obtain the schools': i) administrative dependency (such as public or private subsidised), ii) geographic location, and iii) urban or rural status.

Social Protection File (SPF) Dataset

This dataset contains information for Chilean households and all their members. Each observation represents an individual (adult or child) who lives in a household. Each household has a unique ID number that allows for the identification of all the individuals who belong to it. Households voluntarily requested the SPF at the local government level. The SPF information was essential to be eligible for multiple social policies. From January 2010, the dataset had 10,782,270 individuals (Comité de Expertos Ficha de Protección Social, 2010), approximately 63.5% of Chile's population. I use two versions of this dataset (years 2012 and 2013) in this research. The MSD administers the dataset. Some of its variables are household structure, gender and schooling. With this information I can generate additional variables such as household size, female head of household and years of schooling of the head of the household

Dataset Structure and Sample

To carry out the impact assessment, I structure the dataset by cohorts (years t: 2013 and 2014). Each year-cohort uses information from years $t−1$, t and $t+1$. The BLE Dataset of year t provides the programme recipients in year t and, among other variables, the academic performance and

---

[7] All these variables are from year t. For example, the 2013 ME Performance Dataset provides the average grade for the academic year 2013. The 2013 BLE Dataset provides the average grade for the academic year 2012.

PFSE scores from year $t-1$. These last two variables are useful to assess eligibility for the cash transfer. I use the information from the SPF of year $t-1$, the same year as the PFSE score, for characterisation. Finally, future average grade and attendance come from the ME Performance Datasets of years t and/or $t+1$. This organisation is presented in detail in Table 1.

Table 1: Dataset Structure by BLE Year-Cohorts

BLE Cohorts	Previous Academic Performance and PFSE Score	SPF Information	BLE Recipient	Future Average Grade and Attendance
Year 2013	2012	2012	2013	2013 & 2014
Year 2014	2013	2013	2014	2015

The body of the paper presents the results for the 2013 cohort. This cohort is more likely to provide valid causal estimates as students were not aware that their academic performance in 2012 would affect whether they received the BLE in 2013. I show the results for the 2014 cohort in an appendix.

The sample excludes students in the eleventh and twelfth grades in year $t-1$. This restriction limits the sample to students that would not (or were unlikely to) graduate from secondary education in years $t-1$ or t. Therefore, these students were more likely to have been enrolled in primary or secondary education in years t and $t+1$, which reduces sample attrition. Additionally, I exclude from the analysis those students whose cohort size in year $t-1$ was lower than seven.

I also exclude students who were at least 18 years old in the month of year t in which the BLE was paid. This action intends to restrict the sample to students that were unable to collect their payments personally. Hence, the impact estimates are only valid for those students whose cash transfer was collected by a household member. The effect for students aged 18 years or older, who could collect their payments on their own, could not be estimated due to low statistical power.

Another essential characteristic of the sample is that it is comprised exclusively of students that had a PFSE score, and accordingly a valid SPF score. 76.9% of the students in the BLE 2013 Dataset had a valid PFSE score. This is not a representative sample of the population of Chilean students, as households with higher earnings were less likely to request a Social Protection File.

These characteristics of the sample do not favour making inferences about the whole student body. However, this is not problematic if the main findings of this study are linked only to students from the fifth to the tenth grades with a Social Protection File in 2012 who were younger than 18 years old in 2013. Given that this subset of students is more likely to be recipients of cash for grades programmes, the relevance of the findings of this study holds.

Descriptive Statistics

Table 2 provides descriptive statistics for the BLE in 2013. The first four columns of the table display the mean values for four subsets of students. I obtain these subsets after splitting the sample by PFSE score (below or equal to 98 points and above this threshold) and by relative ranking (equal to or lower than 0.3 and values higher than this threshold). Thus, the first column only contains students who were eligible for the BLE in 2013. The last four columns of the table give the mean, standard deviation, minimum and maximum values for the entire sample.

Panel A shows the mean for the variable BLE recipient in 2013. The mean is one in the first column and zero in the second, third and fourth columns. These results show full compliance for the BLE in 2013. Every student in the sample who was eligible for the cash transfer was provided with it (if the cash transfer was collected by an adult member of his or her household). Conversely, students who were ineligible for the BLE in 2013 did not have any access to the cash transfer that year. The fifth column of the table shows that approximately 14% of the students in the sample were provided with the BLE in 2013.

Panel B presents descriptive statistics for the variables influencing eligibility for the BLE in 2013. The mean PFSE score in the sample is 113 points with a minimum of 24 and a maximum of 769. Among the BLE recipients, the mean PFSE score is 58.6 points. The mean value for average grade in 2012 is 5.48. For students in the highest 30% of academic performance the mean is higher than 6.00, while for students who are not in this group the mean is around 5.20. The average cohort size is 86.2 students while the average student ranking is 41.1.

Panel C shows that students in the highest 30% in terms of academic performance had a higher percentage of attendance relative to students outside of this group. Regarding the type of school attended, the poorest 30% were more likely to attend public schools and approximately seven out of ten students were enrolled in a traditional primary school. Three out of ten students attended traditional secondary schools. Among this group of students, enrolment in scientific-humanistic (SH) schools was nearly twice that in technical-professional (TP) schools. Students with a PFSE score higher than 98 points were slightly more likely to be enrolled in a secondary SH school relative to their peers with a PFSE score lower than or equal to 98 points.

From Panel D, we see that the sample mean age in 2012 is 12.53 years and that boys were less likely to be part of the highest academically achieving group than girls. Additionally, students in the poorest 30% of the population had a head of household with fewer years of schooling, who was more likely to be female and less likely to be employed compared to students who were not among the poorest 30%. In relative terms, students with a PFSE score no higher than 98 were also more likely to attend a rural school and to live outside the metropolitan (or capital) region.

Table 2: Descriptive Statistics for the BLE in 2013

Variables	Relative Ranking ≤ 0.3		Relative Ranking > 0.3		Total			
	PFSE ≤ 98	PFSE > 98	PFSE ≤ 98	PFSE > 98				
	Mean	Mean	Mean	Mean	Mean	Std. Dev.	Min.	Max.
Panel A: BLE Recipient in 2013								
BLE Recipient in 2013	1	0	0	0	0.141	0.348	0	1
Panel B: BLE Eligibility Variables in 2012								
PFSE Score	58.6	165.4	57.9	161.9	113.0	67.3	24	769
Average Grade	6.04	6.09	5.18	5.23	5.48	0.55	4.0	7.0
SR: Student Ranking	12.6	13.5	51.0	57.4	41.1	50.1	1	786
CS: Cohort Size	82.9	91.6	79.5	91.2	86.2	79.3	7	786
SR/CS: Relative Ranking	0.156	0.149	0.646	0.630	0.483	0.284	0.001	1
Panel C: Attendance and School Information in 2012								
Attendance (%)	93.9	94.4	91.3	92.2	92.5	7.0	1	100
Public School	0.521	0.425	0.520	0.409	0.465	0.499	0	1
Primary Traditional Education	0.685	0.682	0.717	0.678	0.693	0.461	0	1
Secondary SH Traditional Education	0.197	0.223	0.177	0.227	0.206	0.404	0	1
Secondary TP Traditional Education	0.112	0.091	0.096	0.088	0.094	0.292	0	1
Panel D: Demographic Information								
Age in 2012 (Years)	12.44	12.39	12.58	12.60	12.53	1.82	8	16
Male	0.418	0.438	0.525	0.538	0.499	0.500	0	1
Head of Household Schooling (Years)	9.23	10.77	8.78	10.34	9.74	3.37	0	24
Household Monthly Income ($CLP)	112,418.9	272,970.4	108,020.8	256,265.9	189,150.7	194,888.1	0	15,276,972
Metropolitan Region	0.329	0.365	0.324	0.374	0.349	0.477	0	1
Rural School	0.116	0.070	0.121	0.067	0.093	0.290	0	1
Head of Household is Female	0.582	0.294	0.601	0.331	0.450	0.498	0	1
Head of Household is Employed	0.710	0.818	0.704	0.804	0.760	0.427	0	1
Household Size	4.18	4.19	4.28	4.24	4.24	1.45	1	33
Household Number of Rooms	1.95	2.23	1.93	2.20	2.08	0.93	0	63
Number of Observations	149,834	188,959	357,416	368,451	1,064,660			

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Methodological Approach

Regression Discontinuity Designs: Sharp vs Fuzzy, One vs Multiple Running Variables

A researcher interested in identifying the causal relationship between receiving the BLE and future attendance or academic performance needs a suitable strategy for doing so. For example, a simple comparison of outcomes among those who receive the cash transfer and those who do not is likely to provide biased estimates. Restricting the comparisons to subsets of the population is not likely to work either. Within students in the highest 30% of student achievement, BLE recipients differ relative to non-recipients regarding key observable characteristics such as the type of school they attend or the schooling of their household's heads. An analogous situation happens within the poorest 30%. In this case, BLE recipients and non-recipients are not comparable regarding their previous attendance and gender, among other variables.

An alternative approach is a regression discontinuity design. In all RD designs some exogenous variation in treatment occurs around a threshold of a running variable or rating score. Two types of RD designs exist: sharp and fuzzy. In fuzzy RD designs the running variable is a relevant factor, but not the only one, in explaining treatment status. In sharp RD designs treatment is fully explained by the running variable. In the context of RD designs to evaluate the impact of cash transfers, a sharp RD design is suitable for causal inference only if all units that meet the eligibility criteria (generally those having a score below a threshold) receive the cash transfer and if all units that do not meet the criteria never receive the cash transfer.

A sharp RD design is a suitable approach in this paper. In the simplest sharp RD design only one running variable and threshold exists. Here there are two running variables. Regardless, I can use two different sharp RD designs depending on how I restrict the sample. For example, if I utilise the subset of students in the highest 30% of academic performance, I can implement a sharp RD design using the PFSE scores as a running variable as the treatment changes from zero to one at the 98-point threshold (as shown in the first two columns of Panel A in Table 2). Similarly, I can execute another sharp RD design if I restrict the sample to the poorest 30%.

I employ these two alternatives in the paper. Their details are explained later in this subsection. This type of approach, with more than one running variable and where a restricted sample is used to implement an RD design, has been labelled a frontier RD estimation (Reardon & Robinson, 2012) and provides frontier-specific estimates (Wong, Steiner, & Cook, 2013). Consequently, I report two frontier-specific estimates in this paper. The first type of estimate informs us about the effect of the BLE in 2013 solely for students around the 30% of lowest income. The second type of estimate does so only for students near the 30% of highest achievement.

To obtain causal estimates, RD designs compare outcomes for units just above and below a threshold in a running variable. In a context of two running variables, the alternative to estimating two frontier-specific effects is to estimate a unique frontier-average treatment effect directly. This can be done either by estimating the discontinuity in the outcome along both thresholds

simultaneously using bivariate regressions or by collapsing the two running variables into a single one (Wong et al., 2013). Either of these two approaches will imply using more complex methods and assumptions. Given that the frontier-average treatment estimate is the result of a weighted average of frontier-specific estimates (Wong et al., 2013) I opt to provide only the latter type of estimate and use standard (univariate) RD estimation strategies and assumptions.

Local Randomisation vs Continuity-Based Framework in RD Designs

In RD designs, causal inference relies on the assumption that the average outcome for units marginally at one side of the threshold must represent a valid counterfactual for the group just at the other side of the threshold (Hahn, Todd, & Van der Klaauw, 2001; Lee, 2008). Despite this common theoretical understanding, there are disparities in, and a lack of consensus about how RD designs are interpreted and implemented in practice (Cattaneo, Idrobo, & Titiunik, 2018a). Two frameworks exist for RD analysis: local randomisation and continuity-based. These two frameworks rely on different identification assumptions and, consequently, differ in their strategies for estimation and the tests they use to assess the internal validity of their estimates (Cattaneo, Titiunik, & Vazquez-Bare, 2017b; Sekhon & Titiunik, 2017). In practice, the continuity-based framework is most commonly employed (Cattaneo et al., 2018a).

Researchers who adopt the local randomisation framework use the logic of experimental designs to recover causal estimates. They take the simple average of the outcome in a small window on either side of the index' threshold. Hence, the impact estimate is equivalent to the difference in means across the cutoff point. The same approach is used to assess the quality of the randomisation. The underlying assumption of this framework is that the average potential outcomes are uncorrelated with the running variable within a small neighbourhood close to the threshold (Cattaneo, Idrobo, & Titiunik, 2018b).[8] For this assumption to hold, the treatment needs to be at least as good as randomly assigned for units within a distance w from the cutoff C.

Researchers relying on the continuity-based framework use regression at each side of the threshold to predict the limiting value of the outcome precisely at the threshold. In this framework the running variable can be associated with average potential outcomes, but this association is assumed to be smooth at the threshold. Therefore, this continuity assumption allows us to interpret any discontinuity in the conditional expectation of the outcome (as a function of the running variable) at the threshold as causal evidence of the treatment (Imbens & Lemieux, 2008). To assess the internal validity of the causal estimates, it is necessary to check whether the distributions of the pre-treatment variables show discontinuities. Repeated discontinuities in these distributions at the threshold cast doubt on the plausibility of the continuity assumption.

Within the continuity-based framework, when the estimation uses only observations near the threshold the approach is known as local polynomial or non-parametric (or a combination of these

---

[8] Each observation i has two potential outcomes. $Y_i(1)$ is the hypothetical outcome that would be observed if assigned to treatment while $Y_i(0)$ would be observed in case of being assigned to the control group. In a sharp RD context, we can only observe $Y_i(1)$ for units at one side of the threshold (while $Y_i(0)$ remains unobserved for this group) and $Y_i(0)$ for units at the other side of the threshold (with $Y_i(1)$ being unobserved for this group).

two terms). Conversely, a global or flexible parametric model uses all or most of the running variable's support. The former approach has been recommended over the latter in the recent literature (Cattaneo et al., 2018a; Cattaneo et al., 2017b) due to its increased capacity to predict boundary points, the estimate of interest in the continuity-based RD framework.

The local randomisation framework requires a stronger assumption relative to the continuity-based framework (Cattaneo et al., 2018b). For the causal estimates to be unbiased in the local randomisation framework, the average potential outcomes need to be uncorrelated with the running variable along the whole interval $[C-w, C+w]$. If this assumption holds, then the continuity assumption on which the other RD framework relies holds. However, continuity does not assure local independence between the average potential outcomes and the running variable.

A priori, I am not able to guarantee that the local randomization RD framework central assumption holds. For example, by design PFSE scores are highly correlated with income per capita, a variable that in turn may be correlated with future academic performance and attendance within the neighbourhood around the 98-point threshold. Similarly, previous academic performance is likely to be a strong predictor of future academic performance within the interval $[C-w, C+w]$. In any of these cases, RD estimates equivalent to differences in means will most likely be biased.

In contrast, the continuity assumption holds more plausibly in the two sharp RD applications I propose. Accordingly, the body of the paper focuses on the continuity-based RD framework. Regardless, I discuss the local randomisation RD framework and its results in an appendix.

RD Application #1: Sharp RD Design Using PFSE Score in 2012 as a Running Variable

In the first RD application, I use the PFSE score as a running variable. I implement a sharp RD design after restricting the sample to students in the highest 30% of academic performance.

Because I use a continuity-based RD framework, first I need to choose the size h of the bandwidth. The size of the bandwidth determines which observations of the income index (around the 98-point PFSE threshold) are used in the local regression. Hence this design relies only upon units within the interval $[98-h, 98+h]$. I choose h in a data-driven way to avoid selecting it arbitrarily. More precisely, I choose the h that minimises the mean squared error of the local polynomial estimator. This is the most popular approach for bandwidth selection in RD designs (Cattaneo et al., 2018a). I obtain the causal estimates from the following regression:

$$Y_i = \alpha_1 + \beta_1 I_{1i} + \theta_1 f(\Delta PFSE_i) + \gamma_1 f(\Delta PFSE_i)I_{1i} + \varepsilon_{1i},$$

where Y_i is average grade or attendance for student *i* in year *t* or *t*+1. $\Delta PFSE_i$ is the 98−PFSE score in *t*−1 for student *i* (distance to the 98-point PFSE threshold). I_{1i} is a binary variable that takes a value of zero if $\Delta PFSE_i$ is negative. Otherwise, I_{1i} is one. $f(x)$ is a local polynomial of *x* of order *p*. ε_{1i} is the error term, the difference between the observed and predicted values.

In this approach β_1 corresponds to the causal effect. Cattaneo et al. (2018a) recommend using a triangular kernel in the regression. Additionally, they recommend that the order *p* of the local polynomial for use in the estimations should be one or two. For inference, I assume that the observations are clustered by schools. To assess the internal validity, I use the same local regression but I replace Y_i with each variable of X_i', a vector of pre-treatment variables. After running the regression for each pre-treatment variable, I perform a test of joint significance. If the hypothesis of no joint significance is rejected the continuity assumption is unlikely to hold.

RD Application #2: Sharp RD Design Using Average Grade in 2012 as a Running Variable

A second candidate for running variable is the relative ranking in 2012. Although students did not know that their relative academic performance in 2012 was being used for the BLE assignment in 2013, this is not a suitable running variable. The relative ranking is a result of a two-step administrative procedure that transforms the average grade of each student. The first step transforms the average grade into a ranking. The second step transforms the latter value into a relative ranking by dividing the student ranking by the number of students in the cohort. This two-step procedure non-randomly affects the position of students near the 0.3 threshold and causes them not to be comparable around it. Accordingly, there is no valid counterfactual as potential outcomes are likely to differ for units across the 0.3 threshold in the relative ranking.[9]

In my second RD application I use the average grade in 2012 as a running variable. I implement a sharp RD design after restricting the sample to the poorest 30% of the students. The average grade in 2012 is a suitable candidate for a running variable in a sharp RD design. This variable changed the eligibility for the BLE in 2013 deterministically. Among the 30% poorest students, those whose average grade in 2012 was equal to or higher than *T* received the BLE in 2013, while those whose average grade was lower than *T* did not receive it. *T* represents the average grade that determined which students were in the top 30% in terms of highest achievement in each academic cohort. Thus, various thresholds exist between cohorts. Unlike the relative ranking, the average grade is free from administrative sorting.[10] Students that differ in their average grade by a tiny fraction are not expected to differ in terms of potential outcomes and pre-treatment variables.

As there are different thresholds for multiple subgroups I implement a multi-cutoff RD design. The approach generally consists of normalising the running variable, for example assigning the value of zero to all units with a score of *T*, and then pooling all the observations (Cattaneo, Keele, Titiunik, & Vazquez-Bare, 2016a). This strategy has been used in multiple RD papers in topics such as education, poverty and health (Carneiro, Galasso, & Ginja, forthcoming; De la Mata, 2012; Lindo, Sanders, & Oreopoulos, 2010; Pop-Eleches & Urquiola, 2013). Accordingly, in my second RD

---

[9] Appendix B provides empirical support for this argument. First, I present differences in means in a small window around the 0.3 threshold in the relative ranking for key pre-treatment variables. I find multiple, large and statistically significant differences. Then, I provide estimates for the continuity-based framework. I find comparable results relative to the local randomisation framework. Finally, the appendix explains how the procedure affects the position of different types of students near the 0.3 relative ranking threshold systematically.

[10] Administrative sorting relates to procedures, beyond the control and knowledge of individuals, that affect the position of these individuals non-randomly near the threshold. Administrative sorting threatens the continuity assumption on which RD designs rely.

application I fit the following local regression (where β_2 is the causal estimate):

$$Y_i = \alpha_2 + \beta_2 I_{2i} + \theta_2 f(\Delta AG_i) + \gamma_2 f(\Delta AG_i)I_{2i} + \varepsilon_{2i},$$

where Y_i is the average grade or attendance for student i in year t or t+1. ΔAG_i is the average grade of student i in t−1 minus T (distance of average grade to the threshold). I_{2i} is a binary variable that takes a value of one if ΔAG_i is non-negative. Otherwise, the variable takes a value of zero. $f(x)$ is a local polynomial of x of order p. Finally, ε_{2i} corresponds to the error term.

I could not select the bandwidth h driven by the data on this occasion. There are not enough unique values in the running variable to implement the algorithm that estimates the optimal bandwidth. This is explained by the average grade in Chile being rounded and reported only with one decimal place (the average grade is the result of a simple average of multiple courses).[11] Instead, I opt for two values of h (h=0.2 and h=0.3). Given the average grade scale, these are the minimum bandwidths from which I can fit a local regression of order p one and two, respectively. The other technical aspects of the estimation (such as the type of kernel, the internal validity tests and clusterisation) are the same as in my first RD application.

RD Graphs and Running Variable Density Test

A key component of RD papers are graphs. I provide multiple types of figures to help the reader to assess the robustness of the key assumption, continuity, on which these designs rely.

The first type shows the relationship between the running variable and the outcomes. In this case a clear discontinuity at the threshold is suggestive of a treatment effect. The second type presents the relationship between the running variable and pre-treatment variables. Repeated discontinuities at the threshold cast doubt on the plausibility of the continuity assumption.

I build these two types of RD graphs following Cattaneo et al. (2018a). The running variable is shown on the horizontal axis while the outcome or pre-treatment variable is represented by the vertical axis. I calculate the average value of the vertical axis variable for a limited number of non-overlapping bins of the running variable. These values are shown by dots. I add a fourth-degree polynomial, fitted in the original data, at each side of the threshold. The polynomial represents the association between the variables in the horizontal and vertical axes. The dashed lines surrounding each polynomial represent the 95% confidence interval of the fitted function.

The third type of RD figures I provide in the paper are histograms of the running variable. A discontinuous density around the threshold usually indicates manipulation or sorting, which makes the continuity assumption on which the RD design mostly relies less likely to hold. I provide these

---

[11] Given that my second running variable is rounded, there is a potential risk of a rounding error in the RD estimations. One way to account for this is to follow Lee and Card (2008) and assume random deviations between the true regression function and the approximating function and estimate confidence intervals based on standard errors that are clustered by the running variable. However, Kolesár and Rothe (2018) recommend against this practice. Another approach is to follow Dong (2015), but this implies modelling the curvature of the outcomes by the running variable within the discrete values used, adding complex and untestable assumptions to the estimates.

graphs along with the results of a manipulation test for discrete running variables (Frandsen, 2017). I use this method instead of the McCrary test (McCrary, 2008) because the latter tool performs poorly when the running variable is not continuous (Frandsen, 2017).

Frandsen's test is based on smooth approximations of the running variable density close to the threshold of interest. Accordingly, the test detects deviations in the running variable density. Therefore, if the test is rejected this is interpreted as a sign of manipulation or sorting. In this sense this test is like McCrary's but Frandsen's test uses only points immediately adjacent to the threshold. If I assume that the density of the running variable is linear near the threshold (k=0) the test will detect small deviations from linearity at the threshold. This is the most rigorous criterion. If I allow any degree of curvature for the density of the running variable (k>0) the test is less likely to be rejected. I run the test using three values of k (k = 0, 0.1 and 0.2), which allows me to assess the sensitivity of the estimates to the curvature of the running variable.

Results

This section presents the results of the impact assessment. The first subsection focuses on the use of PFSE scores in 2012 as a running variable. The second part is centred around the use of average grade in 2012 as a running variable. Only continuity-based RD estimates are shown in these subsections. The third part incorporates the findings of the previous two subsections and synthesises the results of the local randomisation RD framework, the impact of the BLE in 2014 and the BLE in 2013 on subgroups (available in Appendixes C, D, and E, respectively).

PFSE Score in 2012 as a Running Variable

RD Estimates

Table 3 presents the RD estimates (β_1) for each outcome. The first four columns show the results for average grade in 2013 and 2014 while the last four columns do so for attendance. For each outcome, I present two estimates. Within each outcome, the first estimate uses a local quadratic regression (p = 2) while the second estimate uses a local linear regression (p = 1).

The estimates for future average grade are all close to zero and statistically insignificant. The estimates range from −0.010 to −0.003 points. Concerning the standard deviation of average grade, these local regression outputs range from −0.015 to −0.005. The estimates for attendance in 2013 and 2014 are also close to zero and not statistically significant at any level of confidence. The estimates are negative for 2013 and positive for 2014. Overall, the range goes from −0.098% to 0.108% where these results translate approximately into a fifth of a school day. On the scale of the standard deviation of attendance the estimates range from −0.011 to 0.012. Among the estimates, the lowest standard error is 0.013σ while the highest is 0.021σ. Therefore, if any RD estimate would have had an absolute value higher than 0.042σ this would have been statistically different from zero with a 95% level of confidence. Depending on the estimate I analyse, this could have been statistically significant with an absolute value as low as 0.026σ.

RD estimates are more sensitive to bandwidth selection than any other component (Cattaneo et al., 2018a). For this reason, these estimates are expected to be robust to different bandwidth sizes. Table 4 and Table 5 present again the results of the continuity-based framework for future average grade and attendance. On this occasion, I test two alternative bandwidths for each specification (column) of Table 3. I change the size of bandwidth h by 1.5 times and by half. Modifying the size of the bandwidth reveals few changes relative to the estimates in Table 3. Table 4 shows that the estimates for average grade in 2013 and 2014 remain near zero, are all statistically insignificant and mostly negative. Table 5 focuses on attendance. The estimates are close to zero, negative in 2013 and positive in 2014, and they all lack statistical significance.

Table 6 presents RD estimates for ten pre-treatment variables. None of these variables could have been affected by the BLE in 2013. Consequently, this exercise helps to assess the plausibility of the continuity assumption. I show two estimates per variable. While Panel A uses a local quadratic regression, Panel B uses a local linear regression. Overall, the conditional distributions of these variables do not show discontinuous behaviour at the threshold. I find no statistically significant coefficients for previous academic performance and attendance, age, enrolment in public or rural schools, household income and size, or the schooling and gender of the head of the household. Gender is the only variable for which I obtain a 95% statistically significant estimate. The differences in the estimated proportion of males at the threshold are 0.027 and 0.022 points. For both Panel A and Panel B, the joint tests of statistical significance of these ten variables are not rejected.

To sum up, the main causal estimates are not statistically significantly different from zero. If an effect of the BLE in 2013 exists (for students near the 30% relative income threshold), its size is likely to be no larger than a small fraction of a standard deviation and cannot be statistically detected. There is a small discontinuity in the distribution of gender, but no other pre-treatment variable shows this behaviour. In my sample, being male is not correlated with attendance; thus it is improbable that this outcome is affected. Being male is weakly negatively correlated with future average grade. This correlation may explain the negative coefficients for average grade. If this is the case the causal estimates could be slightly downwardly biased. However, the effects are most likely small in magnitude and not statistically significant.

RD Graphs and Running Variable Density Test

Figure 1 presents the density of the PFSE scores among the subgroup of students in the highest 30% of academic achievement and the results of the manipulation test proposed by Frandsen (2017). No abrupt changes in the density are observed along this figure or close to the BLE threshold (vertical line). The test fails to reject the hypothesis of no difference in the expected density at each side of the threshold. The most rigorous application, with no degree of curvature allowed, has a p-value of 0.187. The p-value increases when I partly relax this restriction.

**Figure 1: PFSE Score Density and Frandsen Manipulation Test
(Students in the Highest 30% of Academic Performance in 2012)**
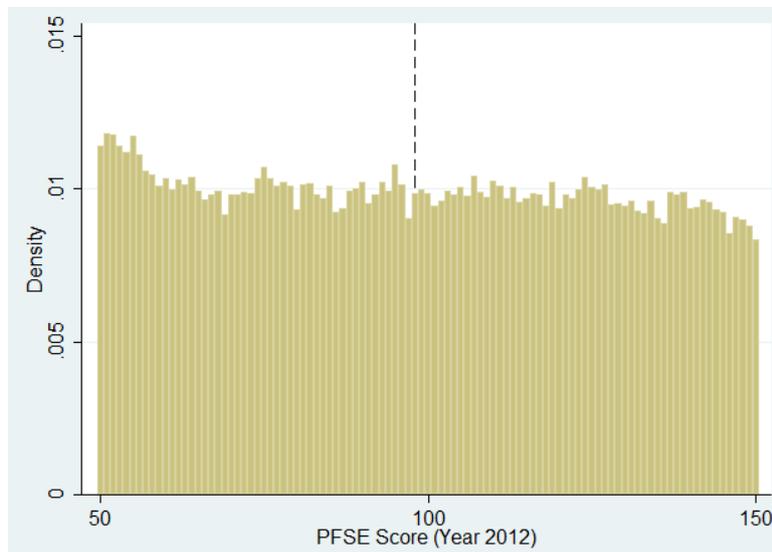
Manipulation test (*k*=0) *p*-value=0.187
Manipulation test (*k*=0.1) *p*-value=0.221
Manipulation test (*k*=0.2) *p*-value=0.293

Source: own calculations using administrative datasets, Chilean ME and MSD

Figure 2 presents a series of graphs that depict the relationship between the running variable and the outcomes of the assessment. Future average grade graphs can be found in the upper panel of the figure, while future attendance graphs are in the lower panel. In general terms, the figure shows a positive but weak association between the PFSE scores and both types of outcomes. More importantly, no graph shows a relevant discontinuity between the polynomial fitted functions at each side of the vertical line. Any detected discontinuity is small and not statistically significant as the confidence intervals of the fitted functions noticeably overlap at the threshold. This evidence is consistent with the RD estimates of Tables 3, 4 and 5.

Figure 3 shows eight graphs that illustrate the relationship between the PFSE scores in 2012 and the variables that could not have been affected by the BLE in 2013. Some variables, such as income and head of household's schooling, have an evident degree of association with the PFSE scores. Conversely, other variables have a weaker correlation with the PFSE index. From any of the eight graphs I present in Figure 3 it is possible to claim that discontinuous behaviour exists for a pre-treatment variable at the threshold. In all cases where a difference in the polynomial fit is observed, this is small in magnitude. Additionally, the 95% confidence intervals of the polynomial fits mostly overlap in each of these eight graphs. Although not strictly comparable with the estimates in Table 6, Figure 3 provides similar insights to this source.

Table 2: RD Estimates for Outcomes (Using Mean Squared Error Optimal Bandwidth)

Outcomes	average_grade2013		average_grade2014		attendance2013		attendance2014	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate: Original Outcome	-0.010 (0.012)	-0.008 (0.009)	-0.008 (0.013)	-0.003 (0.010)	-0.025 (0.146)	-0.098 (0.107)	0.108 (0.159)	0.106 (0.143)
RD Estimate: In Standard Deviations	-0.015 (0.019)	-0.012 (0.014)	-0.012 (0.021)	-0.005 (0.016)	-0.003 (0.017)	-0.011 (0.013)	0.012 (0.017)	0.011 (0.015)
Number of Observations	91,832	95,380	87,263	87,263	109,231	105,768	125,075	73,092
Bandwidth Size (*h*)	25.73	27.50	24.96	24.87	31.14	30.45	35.58	21.12
Order *p* of Local Polynomial	2	1	2	1	2	1	2	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 3: RD Estimates for Future Average Grade (Sensitivity Analysis to Bandwidth)

Outcomes	average_grade2013				average_grade2014			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate: Original Outcome	-0.010 (0.010)	-0.016 (0.017)	-0.005 (0.008)	-0.010 (0.011)	-0.005 (0.011)	-0.011 (0.018)	0.001 (0.009)	-0.011 (0.013)
Number of Observations	136,909	45,760	144,061	49,285	128,471	41,880	128,471	41,880
Bandwidth Size (*h*)	38.59	12.86	41.25	13.75	37.43	12.48	37.31	12.44
Order *p* of Local Polynomial	2	2	1	1	2	2	1	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

**Table 5: RD Estimates for Future Attendance (Sensitivity Analysis to Bandwidth)**

Outcomes	attendance2013				attendance2014			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate: Original Outcome	-0.090 (0.122)	-0.166 (0.200)	-0.090 (0.093)	-0.028 (0.143)	0.122 (0.134)	0.034 (0.219)	0.080 (0.121)	0.005 (0.196)
Number of Observations	166,185	56,237	162,448	52,798	185,597	62,667	111,405	38,417
Bandwidth Size (*h*)	46.71	15.57	45.68	15.23	53.37	17.79	31.69	10.56
Order *p* of Local Polynomial	2	2	1	1	2	2	1	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 4: RD Estimates for Pre-Treatment Variables (Using Mean Squared Error Optimal Bandwidth)

Pre-Treatment Variables (Var.)	(1) avg_grade2012	(2) attendance2012	(3) age	(4) male	(5) schoolpub	(6) schoolrural	(7) hmonthincome	(8) hsize	(9) hhschooling	(10) hhfemale
Panel A: Order p of Local Polynomial =2 (Local Quadratic Regression)										
RD Estimate: Original Var.	-0.007 (0.008)	-0.086 (0.130)	0.014 (0.043)	0.027** (0.011)	-0.007 (0.015)	-0.004 (0.008)	-1,024.1 (2,743.5)	0.014 (0.029)	-0.102 (0.080)	0.001 (0.010)
Number of Observations	120,566	96,218	117,115	100,039	106,700	103,351	100,039	106,652	79,028	110,005
Bandwidth Size (*h*)	34.09	26.94	32.68	28.77	29.84	28.66	28.50	31.37	22.87	32.40
Panel B: Order p of Local Polynomial =1 (Local Linear Regression)										
RD Estimate: Original Var.	-0.006 (0.007)	-0.135 (0.099)	0.012 (0.038)	0.022** (0.010)	-0.004 (0.014)	-0.002 (0.007)	-1,070.2 (2,364.1)	0.011 (0.025)	-0.085 (0.064)	-0.000 (0.008)
Number of Observations	85,347	92,635	85,347	65,158	74,553	81,678	61,849	72,109	61,849	79,028
Bandwidth Size (*h*)	24.20	25.86	23.84	18.58	20.98	22.84	17.81	20.55	17.95	22.77

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 2: Future Outcomes by PFSE Score in 2012 (Students in the Highest 30% of Academic Performance in 2012)
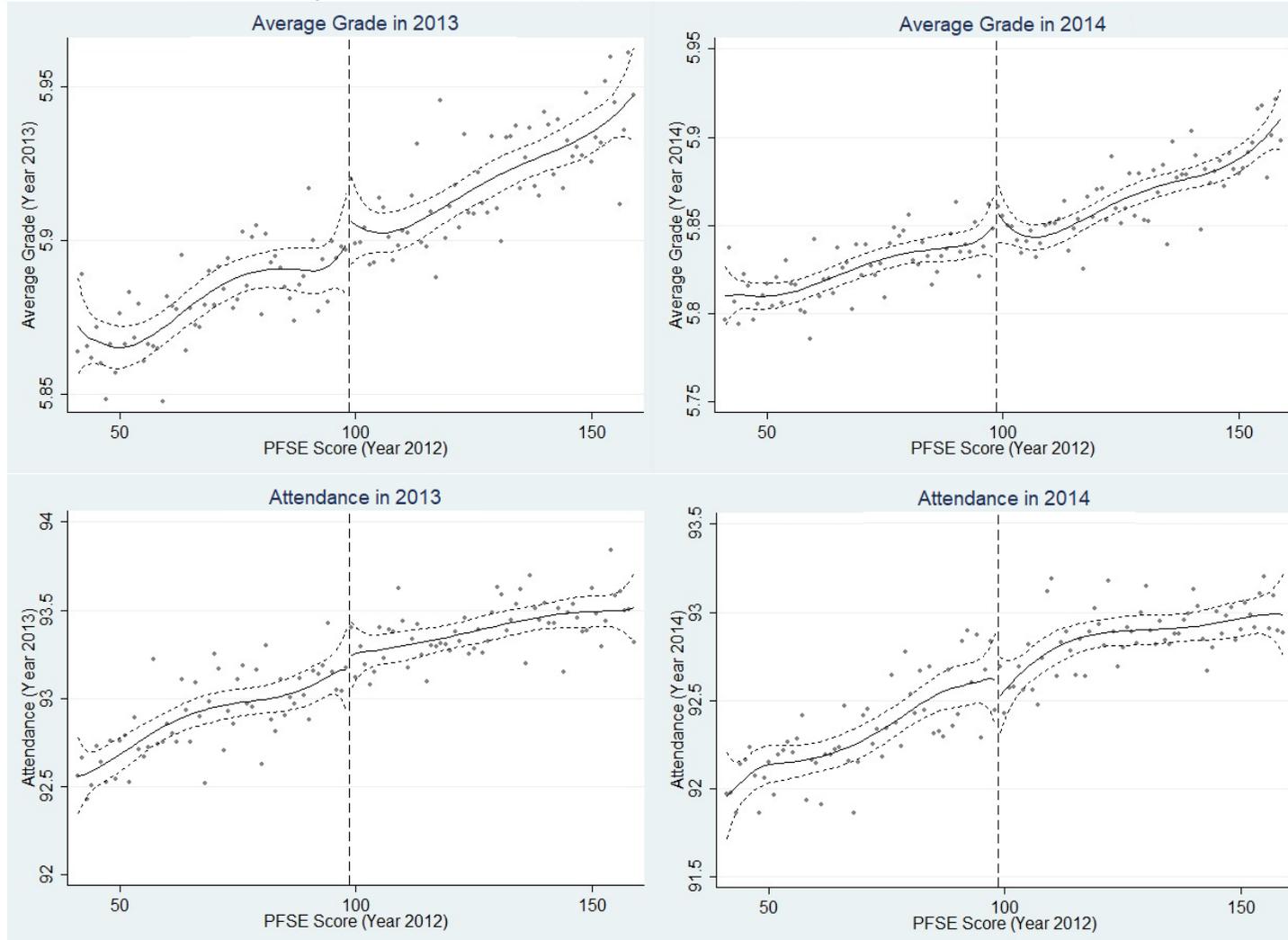
Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 3: Pre-Treatment Variables by PFSE Score in 2012 (Students in the Highest 30% of Academic Performance in 2012)
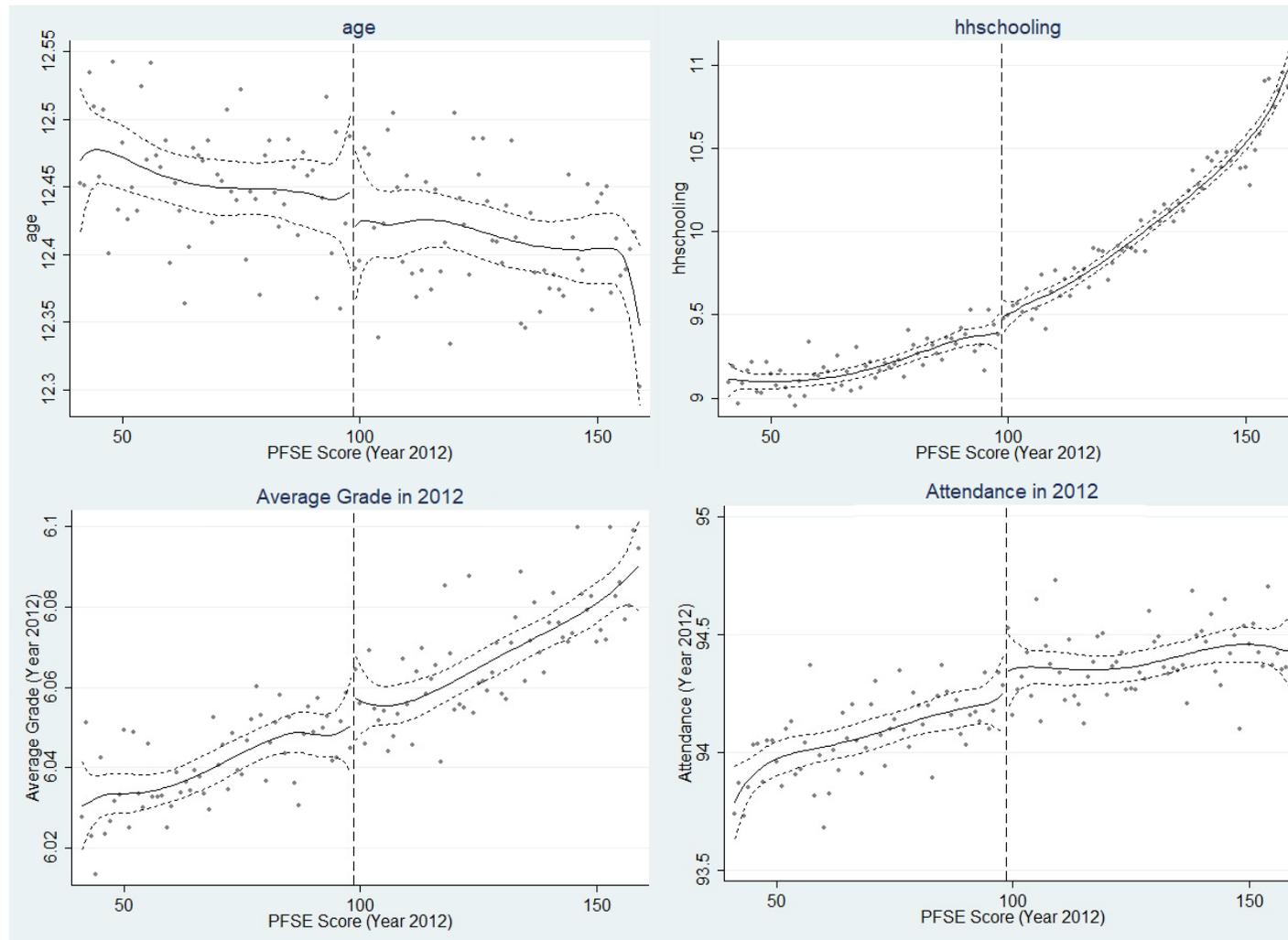
Figure 3 (continued): Pre-Treatment Variables by PFSE Score in 2012 (Students in the Highest 30% of Academic Performance in 2012)
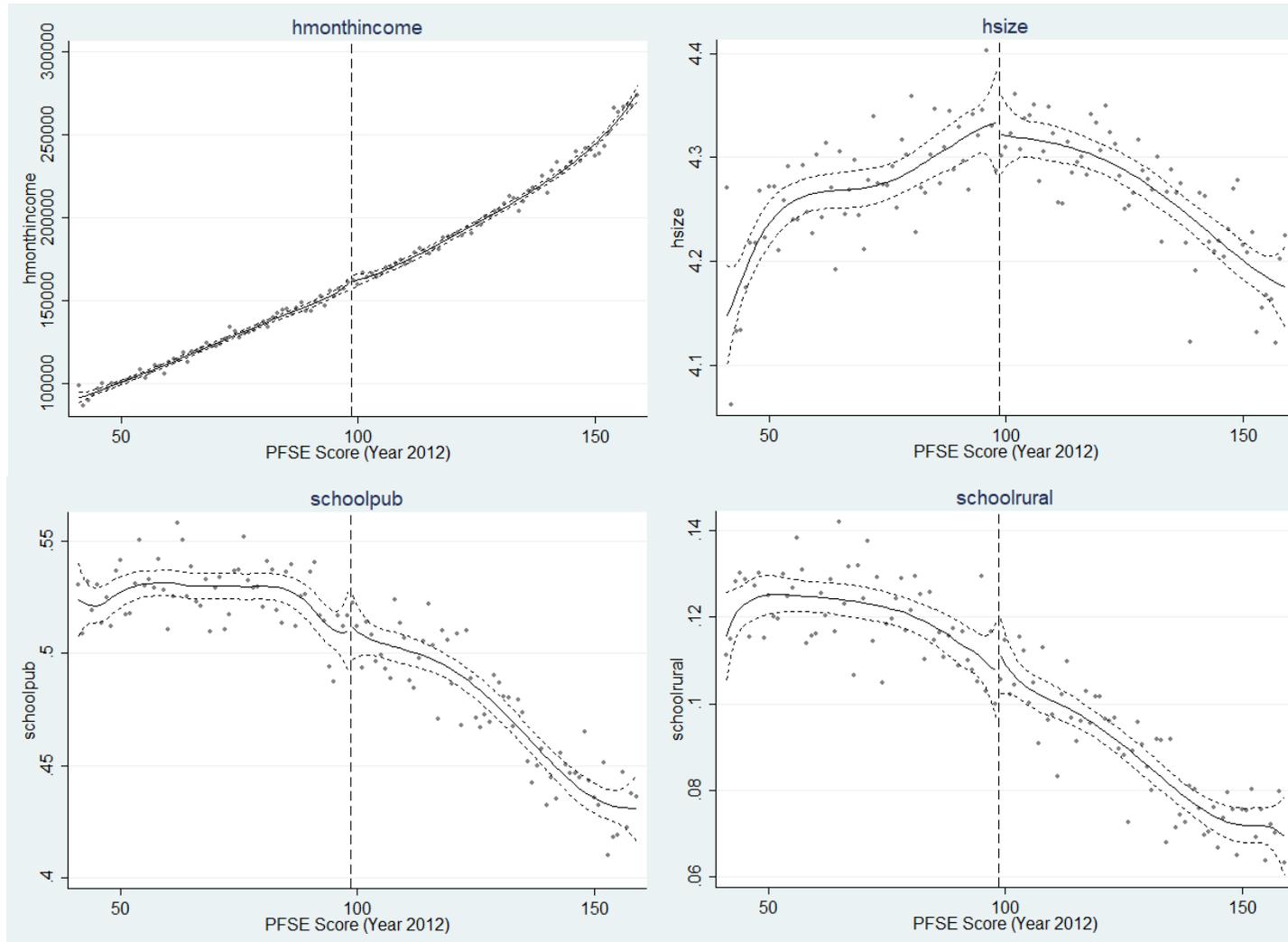
Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Average Grade in 2012 as a Running Variable

<u>RD Estimates</u>

Table 7 presents the results for future average grade and attendance. The first four columns focus on the former type of outcome, while the last four focus on the latter. The first type of RD estimate per outcome (β_2) uses a bandwidth h of 0.3 and a local quadratic regression. The second type of estimate comes from a local linear regression that uses a bandwidth h of 0.2.

All the main estimates for future average grade are small and not statistically different from zero at a 95% level of confidence. The estimates fluctuate from a minimum of −0.003 points (−0.004σ) to a maximum of 0.017 points (0.027σ). This last value is close to achieving a 95% level of statistical significance given that its standard error is 0.015σ. The range for the estimates of attendance goes from −0.043% to 0.147% (−0.005 to 0.016 of a standard deviation). Given that the highest standard error is 0.020σ, the approach would have detected statistical significance for any estimate higher than 0.040σ. Overall, these results are similar to the estimates in Table 3.

Table 8 and Table 9 summarise the sensitivity analysis by bandwidth choice. Overall, between the outputs of each of these tables and the estimates from Table 7, the differences in magnitude are positive but small. The estimates for future average grade in Table 8 are all close to zero (ranging from 0.002 to 0.023 points). Some estimates are statistically significant at a 95% level of confidence but overall these are not robust to alternative specifications. The estimates for attendance in 2013 and 2014 are small and lack any statistical significance. The estimates for 2013 range from 0.012% to 0.053%, and in 2014 from 0.138% to 0.189%.

Table 10 presents the estimates for ten pre-treatment variables. The specifications I use per variable follow a similar pattern to Table 7. The Panel A estimates are the result of a local quadratic regression run in a bandwidth of size 0.3. Panel B provides the results for a local linear regression run in a bandwidth of size 0.2. This approach helps to assess the likelihood of the continuity assumption to hold, as the BLE could have impacted none of these variables.

Overall, no clear discontinuities emerge in the distribution of any of these variables at the threshold. I find no significant differences from zero at a 95% level of confidence. The only estimate at a 90% level of confidence is average grade in 2012 for a local linear regression (p=1). Given that this coefficient is small and positive, if the average grade linear causal estimates in Tables 7 and 8 are not free of bias, these estimates would be slightly inflated. Accordingly, the potential unbiased estimates would be lower and less likely to be statistically significant.

The main causal estimates for both outcomes are not statistically significantly different from zero at a 95% level of confidence. I observe statistically significant coefficients for some alternative specifications of average grade, but these are not robust. The main estimates for both outcomes are near zero and have standard errors no larger than 0.020 of a standard deviation. Hence, for students around the 30% of highest achievement, the BLE in 2013 is unlikely to have caused a substantial effect.

RD Graphs and Running Variable Density Test

Figure 4 presents the density of the normalised average grade variable among the poorest 30% of students. Overall, the distribution of the distance of average grade to the cohort threshold T is smooth. The density decreases when approaching the zero threshold (the vertical line) from the left and increases when reaching this cutoff point from the right. The Frandsen test is not able to detect significant deviations in the expected density of the running variable. The minimum p-value of the test, for the most rigorous version of it, is 0.678.

Figure 4: Distance of Average Grade to the Threshold Density and Frandsen Manipulation Test (Students With PFSE $\leq$ 98 in 2012)

Manipulation test (k=0) p-value=0.678
Manipulation test (k=0.1) p-value=0.862
Manipulation test (k=0.2) p-value=0.989
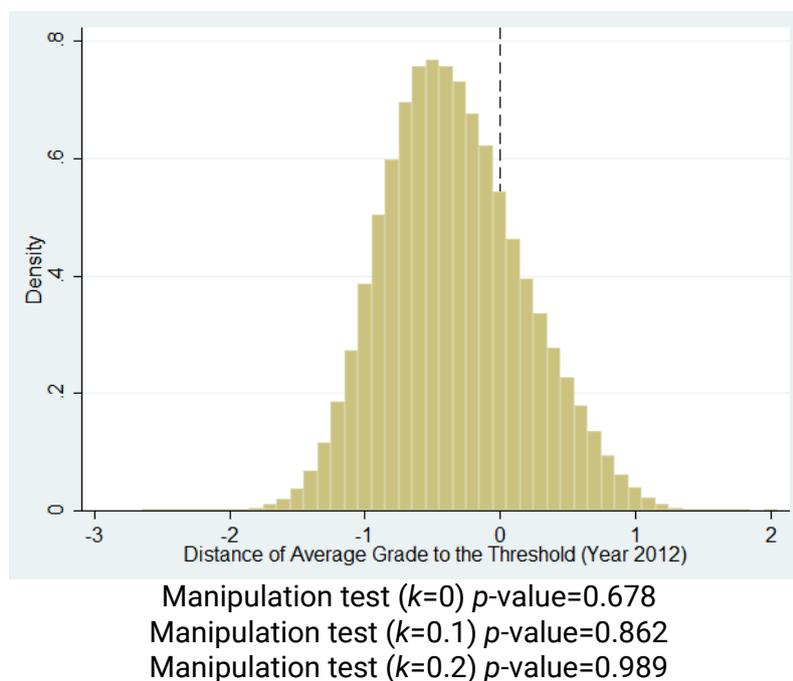
Source: own calculations using administrative datasets, Chilean ME and MSD

Figure 5 illustrates the relationship between the running variable and the outcomes. The upper panel shows the graphs for average grade while the lower panel focuses on attendance. The figure displays a positive association between the running variable and the outcomes. This association is strong as the confidence intervals of the fits are narrow. If the polynomial fits at each side of the vertical line were to be extended to the threshold it would not be possible to distinguish a clear discontinuity. In other words, any estimated discontinuity in the extrapolation would most likely be small. However, it does not seem possible to discard statistically significant differences only by looking at this figure. Although it is not strictly comparable, this graphic evidence is concordant with the results in Tables 7, 8 and 9.

Figure 6 shows eight graphs that illustrate the relationship between the running variable and other variables that could not have been affected by the BLE in 2013. The relationship between the running variable and each of these eight pre-treatment variables differs notoriously. The graphs

help to determine the plausibility of the continuity assumption to hold. No discontinuous relationships at the threshold are noticeable from the figure. No graph suggests an abrupt change in the projected distributions of the pre-treatment variables at the threshold. The results from Table 10 are consistent with Figure 6. Both findings provide support for the suitability of this RD approach to identify the causal effects of the *Bono por Logro Escolar* in 2013.

Table 5: RD Estimates for Outcomes

Outcomes	average_grade2013		average_grade2014		attendance2013		attendance2014	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate: Original Outcome	-0.003 (0.011)	0.006 (0.009)	0.010 (0.012)	0.017* (0.009)	-0.043 (0.169)	0.009 (0.126)	0.119 (0.188)	0.147 (0.141)
RD Estimate: In Standard Deviations	-0.004 (0.017)	0.010 (0.014)	0.016 (0.019)	0.027* (0.015)	-0.005 (0.020)	0.001 (0.015)	0.013 (0.020)	0.016 (0.015)
Number of Observations	166,638	112,023	163,219	109,783	166,638	112,023	163,219	109,783
Bandwidth Size (h)	0.300	0.200	0.300	0.200	0.300	0.200	0.300	0.200
Order p of Local Polynomial	2	1	2	1	2	1	2	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 6: RD Estimates for Future Average Grade (Sensitivity Analysis to Bandwidth)

Outcomes	average_grade2013				average_grade2014			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate: Original Outcome	0.004 (0.009)	0.002 (0.009)	0.016** (0.007)	0.011 (0.008)	0.016* (0.009)	0.014 (0.010)	0.023*** (0.008)	0.021** (0.008)
Number of Observations	269,585	219,355	219,355	166,638	263,547	214,693	214,693	163,219
Bandwidth Size (h)	0.500	0.400	0.400	0.300	0.500	0.400	0.400	0.300
Order p of Local Polynomial	2	2	1	1	2	2	1	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 7: RD Estimates for Future Attendance (Sensitivity Analysis to Bandwidth)

Outcomes	attendance2013				attendance2014			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate: Original Outcome	0.020 (0.124)	0.012 (0.138)	0.053 (0.101)	0.043 (0.110)	0.189 (0.139)	0.183 (0.155)	0.138 (0.113)	0.160 (0.123)
Number of Observations	269,585	219,355	219,355	166,638	263,547	214,693	214,693	163,219
Bandwidth Size (h)	0.500	0.400	0.400	0.300	0.500	0.400	0.400	0.300
Order p of Local Polynomial	2	2	1	1	2	2	1	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 8: RD Estimates for Pre-Treatment Variables

Pre-Treatment Variables (Var.)	(1) avg_grade2012	(2) attendance2012	(3) age	(4) male	(5) schoolpub	(6) schoolrural	(7) hmonthincome	(8) hsize	(9) hhschooling	(10) hhfemale
Panel A: Bandwidth h=0.3 & Local Quadratic Regression (p=2)										
RD Estimate: Original Var.	0.002 (0.008)	-0.152 (0.146)	0.070 (0.045)	0.017 (0.011)	0.012 (0.015)	0.004 (0.009)	827.5 (2,175.9)	0.022 (0.031)	-0.047 (0.070)	-0.007 (0.010)
Number of Observations	169,944	169,944	169,943	164,310	169,944	169,944	164,310	164,310	164,310	164,310
Panel B: Bandwidth h=0.2 & Local Linear Regression (p=1)										
RD Estimate: Original Var.	0.012* (0.007)	-0.064 (0.113)	0.044 (0.037)	0.013 (0.008)	0.008 (0.014)	0.002 (0.007)	1,028.4 (1,525.1)	0.010 (0.022)	-0.000 (0.052)	-0.005 (0.008)
Number of Observations	114,153	114,153	114,152	110,369	114,153	114,153	110,369	110,369	110,369	110,369

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Figure 5: Future Outcomes by Distance of Average Grade to the Threshold in 2012 (Students With PFSE ≤ 98 in 2012)
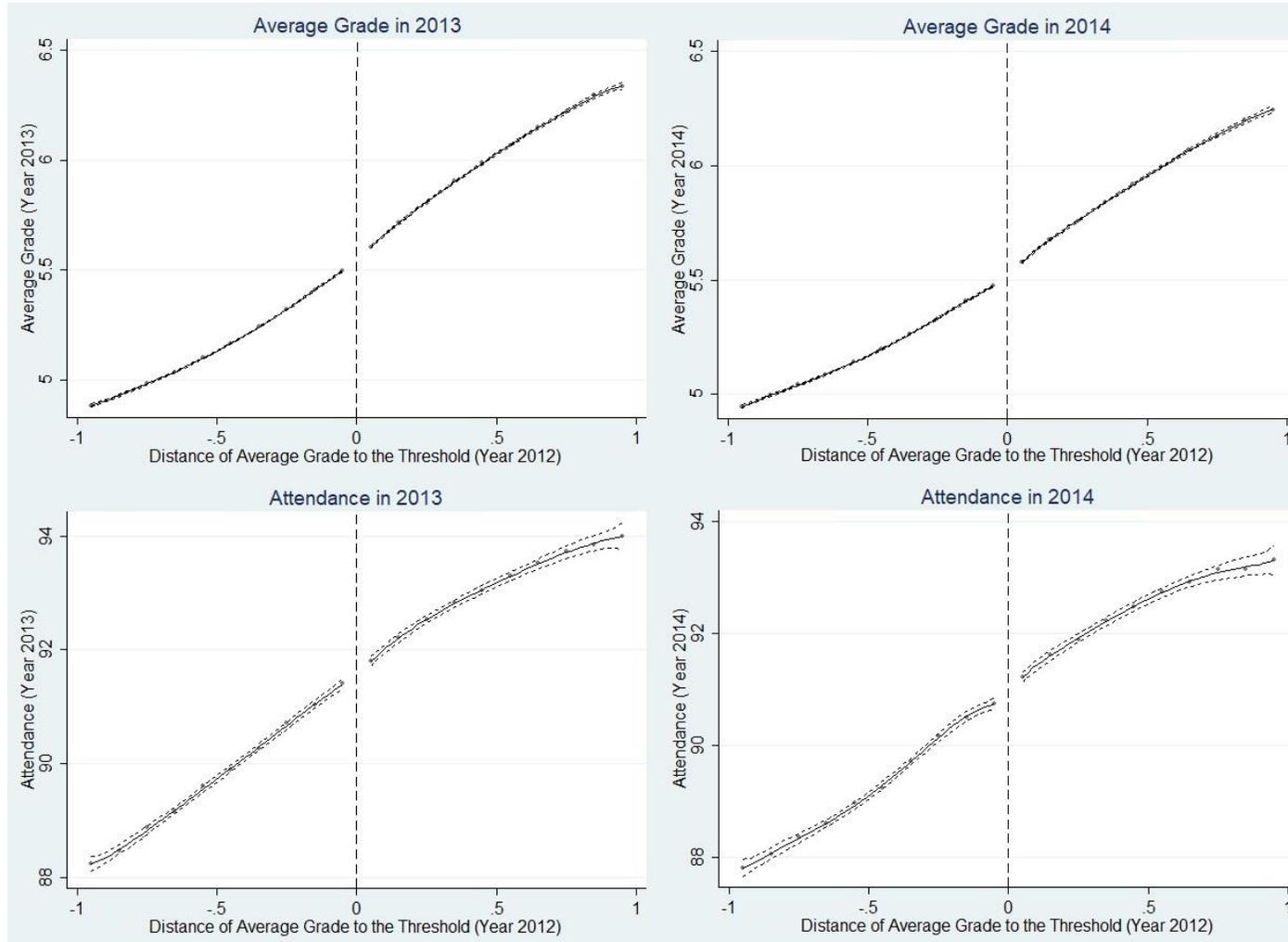
Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development
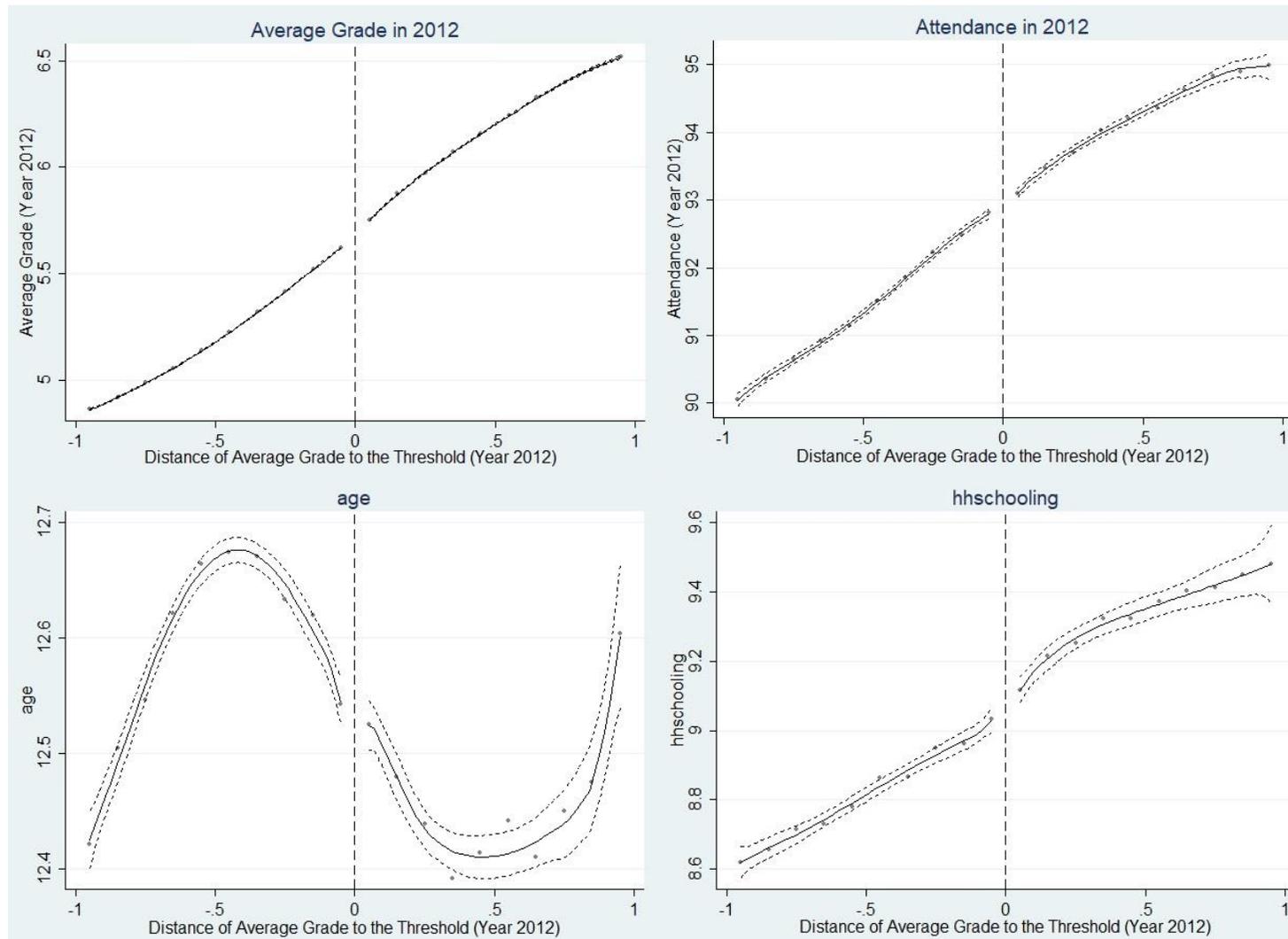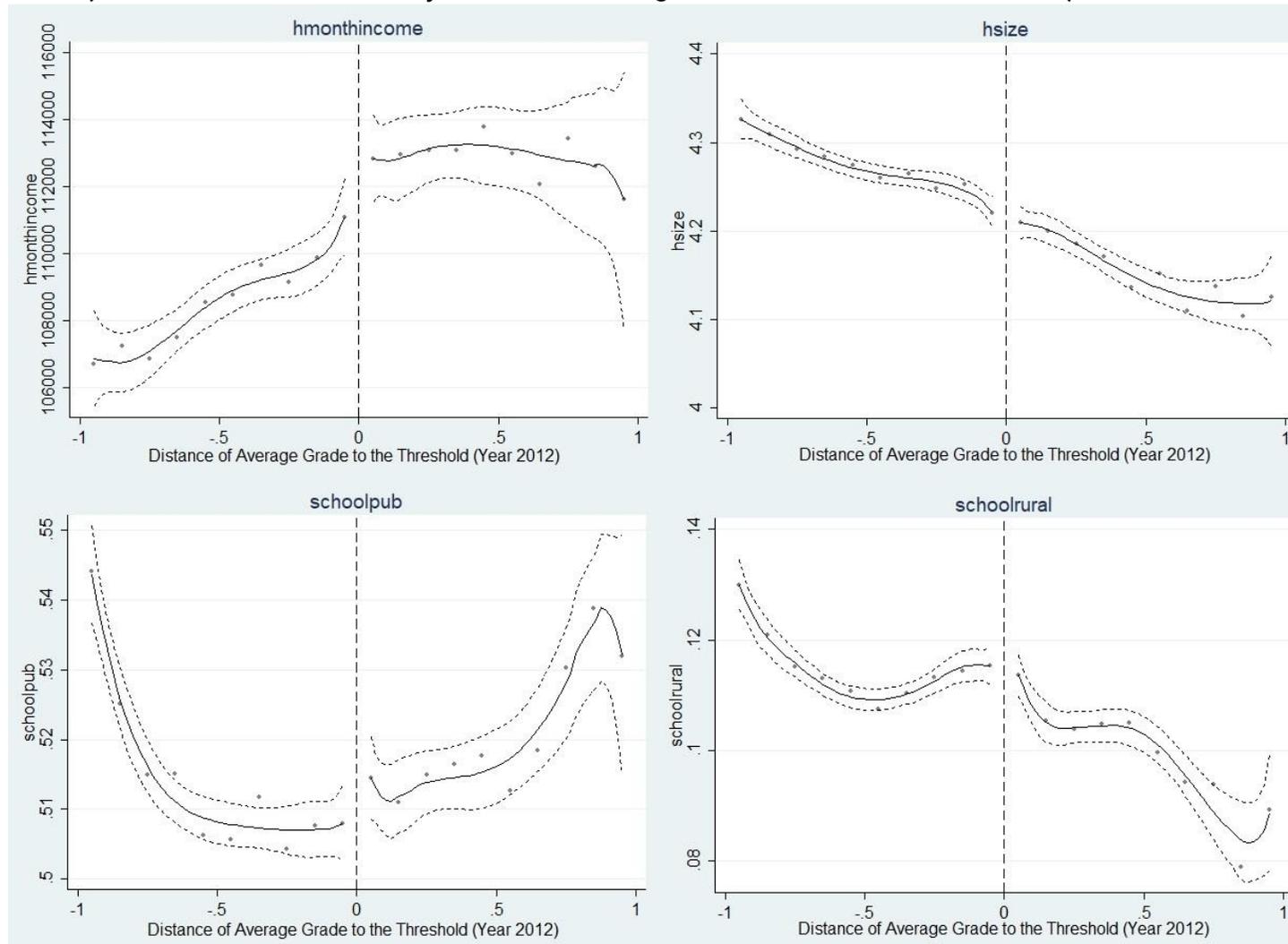
Figure 6: Pre-Treatment Variables by Distance of Average Grade to the Threshold in 2012 (Students With PFSE $\leq$ 98 in 2012)

**Figure 6 (continued): Pre-Treatment Variables by Distance of Average Grade to the Threshold in 2012 (Students With PFSE $\leq$ 98 in 2012)**

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Summary

I analyse different effects of receiving the BLE in 2013 over two groups of students, those around the 30% threshold of lower income and around the 30% threshold of highest academic achievement. Despite using different approaches, the value of zero is part of the 95% confidence interval in every main causal estimate. Additionally, the standard errors range from 0.013σ to 0.021σ.

Figure 7 summarises the main estimates for average grade. The left panel of the figure uses the outcome original scale. The right-hand panel shows the estimates in standard deviations. The most negative estimate of the 95% confidence interval lower bound is -0.034 (-0.053σ) while the highest estimate of the upper bound is 0.035 (0.056σ). These values represent nearly one third of the shortest distance between two grades in the country's educational system.

Figure 7: Summary of RD Point Estimates and Confidence Intervals for Average Grade
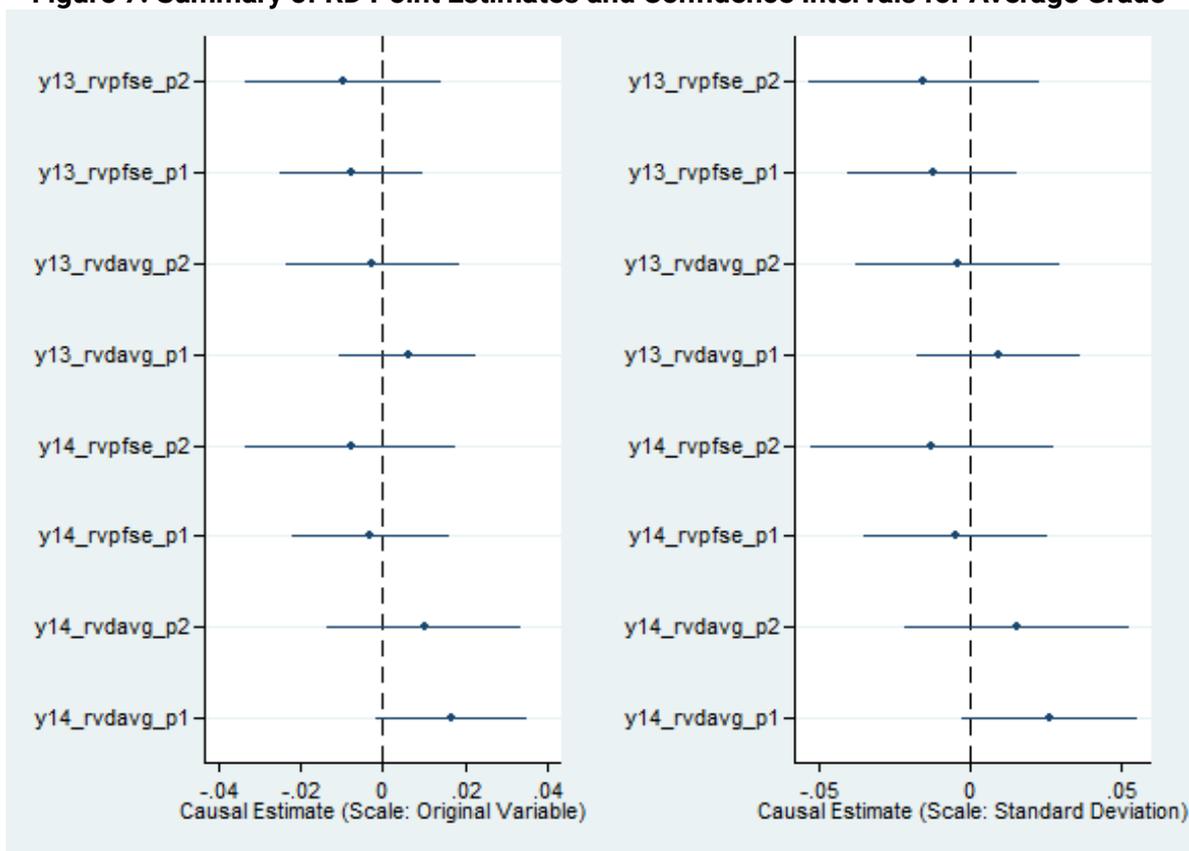
y: year of outcome is 2013 or 2014;
rv: running variable is PFSE score or distance of average grade;
p: order of local polynomial is one or two.

Source: own calculations using administrative datasets, Chilean ME & MSD

Figure 8 summarises the main estimates for attendance. The left and right-hand panels of the figure present the estimates using the outcome original scale and in standard deviations, respectively. The smallest estimate of the lower bound of the 95% confidence interval is -0.37% (-0.044σ), while the largest estimate of the upper bound is 0.49% (0.052σ). In practice, the latter value is equivalent to one day of attendance at school within an academic year.

Figure 8: Summary of RD Point Estimates and Confidence Intervals for Attendance
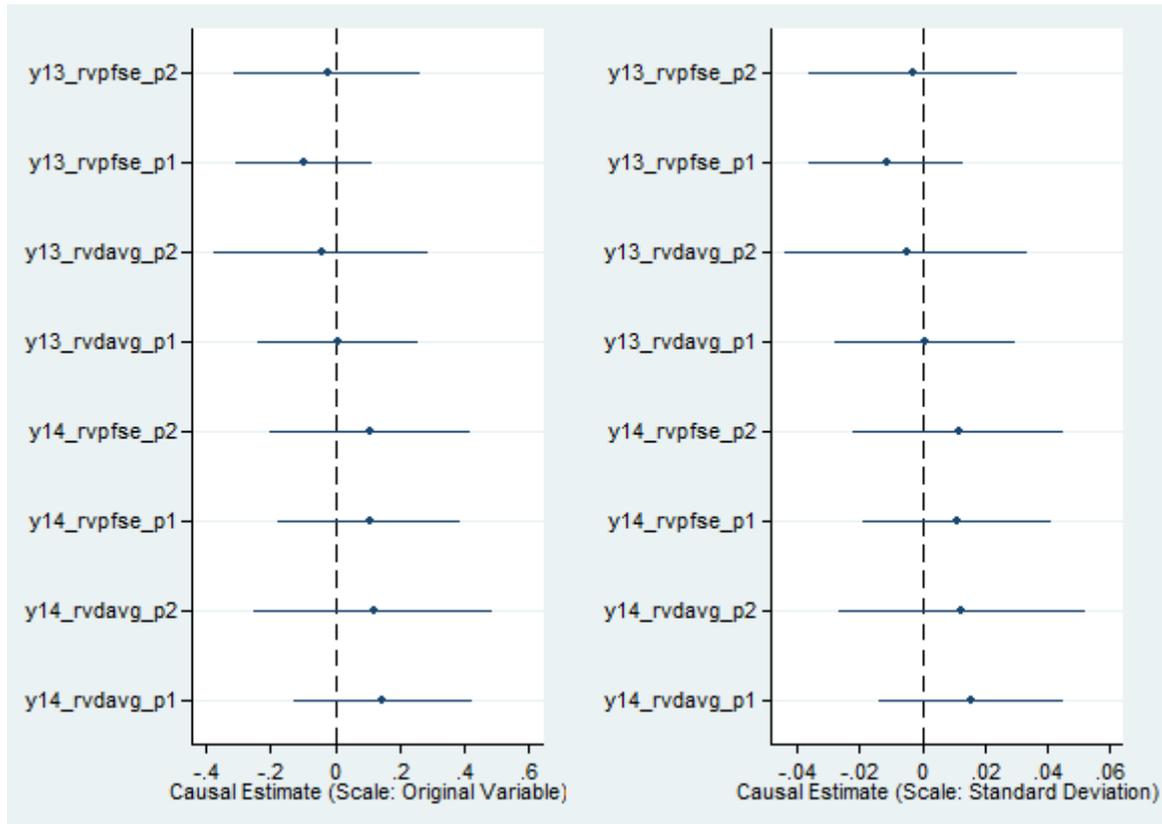
y: year of outcome is 2013 or 2014;
rv: running variable is PFSE score or distance of average grade;
p: order of local polynomial is one or two.

Source: own calculations using administrative datasets, Chilean ME & MSD

Given these estimates, if frontier-specific average effects of the BLE in 2013 exist these are most likely to be modest in magnitude. Thus, I am unable to detect them with statistical certainty. These findings also hold after I apply the local randomisation RD framework (in Appendix C) and calculate the effect of the BLE in 2014 (in Appendix D). The RD design provides average estimates that may miss causal effects on some population subgroups. Additionally, the RD estimates are only valid for observations near the threshold, which are not the poorest of the population in this paper. Appendix E analyses the impact of the BLE in 2013 over some population subgroups. This analysis does not consistently show estimates that are statistically different from zero. Therefore, any effects of this kind are unlikely to be large and could not be captured with statistical certainty.

Conclusion

This paper contributes to the empirical literature on cash for grades impact assessments. I estimate the effect of a Chilean cash for grades programme on subsequent attendance and academic performance. Specifically, I evaluate the impact of the *Bono por Logro Escolar* in 2013. As the cash transfer in 2013 was targeted using two scores from 2012, an income index and academic performance, it is possible to implement a sharp RD design along these two running variables. The differences in students' outcomes at the two thresholds used have a causal interpretation.

The main causal estimates are not statistically significantly different from zero for both types of outcomes. If anything, the BLE frontier-specific average effects are modest and as a result I am unable to detect them with statistical certainty. The size of these potential effects is smaller than those statistically significant effects of near 0.20σ found for interventions of this kind in developing countries (Behrman et al., 2015; Kremer et al., 2009). The results by subgroups do not consistently show estimates that are statistically significantly different from zero.

RD estimates are informative for the population around the thresholds but not necessarily away from them. By design my BLE impact estimates only provide information for students around the poorest 30% threshold and around the top 30% in terms of highest academic achievement. As RD estimates provide average effects for the population near the thresholds, the features of the design do not facilitate observing effects for entire subgroups. For example, we cannot learn about the effects of the BLE for those at the lower end of the income distribution who are at the median or the bottom end in terms of academic achievement. These subgroups of the population may be the ones who are more susceptible and would benefit the most from cash for grades programmes.

Given these caveats it is not possible to generalise my results for the entire population who received the BLE in 2013. Future research could overcome these limitations by introducing some degree of randomisation, allowing for obtaining estimates for the entire population that are expected to be eligible for the cash for grades intervention or entire subgroups of interest.

A possible explanation for the results is that the programme was not very salient for the targeted adolescents. The sample I analyse were not able to collect payments on their own. These were received by an adult member of their household. Moreover, students may not have learnt about the existence of the programme when it was first implemented in 2013. If children were unaware of the implementation of the BLE then it would not be expected to observe changes in their behaviour. If this is still the case, then programme managers could implement actions that increase students' awareness of the benefits (for example, by also giving a diploma to students).

An alternative reason is that children were aware of the cash transfer but unresponsive to its $100 USD maximum size. The monthly minimum wage was $210,000 CLP (approximately $410 USD) in August 2013. If this is a likely scenario then raising the amount of the cash transfer could lead to increased effects. However, whether this is a cost-effective initiative compared to others available deserves more analysis. Another potential explanation for my results is that the BLE provided two

types of effects that cancelled each other out overall. The price effect may have incentivised effort (measured by attendance) while a psychological factor could have reduced it.

All these different hypotheses deserve further exploration in future research. Unlike my paper (which addresses the question of whether the BLE worked), this research will need to focus on a different question: why is the BLE producing little effect on educational outcomes? Interviewing parents and students to find out how aware they are of the BLE implementation could prove useful. Additionally, these interviews could help to understand the causal mechanisms (or lack of them) between providing this cash transfer and subsequent improvements in effort and academic performance. Given the nature of this inquiry, this research should probably be qualitative.

Unlike the other programmes analysed in the literature, the BLE treatment assignment was not randomised at a higher level than students. Other schemes had a group of courses or schools that were part of a treatment group while the rest belonged to a control group. Conversely, in Chile within the same course it is possible to observe BLE eligible and ineligible students. Hence, in this context students may compete for access to the BLE. Students that did not receive the programme in 2013 may have observed classmates or other students accessing the cash transfer. This could have influenced awareness about the BLE and future academic performance in terms of accessing the cash transfer in the future (for example in 2014 and 2015). If this hypothesis is correct, then we would expect to observe higher estimates using the PFSE index as a running variable than for average grade (as eligibility by PFSE score is harder to modify by the student relative to academic performance). However, this is not the case and this hypothesis is unlikely to hold.

This paper provides multiple other contributions to the cash for grades evaluation literature. Beyond analysing the effect of a cash for grades intervention in a context of competition, I study the impact on an overall measure of academic performance, the average grade. This contrasts with the subject-specific criteria (for example test scores in maths, reading or writing) commonly found in the previous literature. Then, an additional potential explanation for the lack of results for the BLE is that for students it implies too much effort, or it is too hard to improve in all subjects (or to improve largely in a few subjects) to increase their average grade.

Whether to reward children according to their academic performance remains a hotly debated and unresolved topic. I observe no significant effects on educational outcomes for the first Chilean cash for grades programme. Further research and an enhanced BLE design may be needed to deliver grades for cash. Otherwise, the country risks little return on its money.

Appendix A. Summary of Cash for Grades Evaluations in Schools

Table 9: Summary of Cash for Grades Evaluations (Randomised Control Trials) in Primary and Secondary Education

Author(s)	Country/State	Programme Description[12]	Main Results
Kremer, Miguel and Thornton (2009)	Kenya	Girls who performed well in academic exams had their school fees paid for two years (7th and 8th grade) and received a grant of $19.20 per year. The scholarship schools were randomly selected from two Kenyan districts. The scholarship was awarded to the highest scoring 15% of 6th grade girls in the programme schools within each district.	An overall effect of 0.19σ is found on academic exams. The results are statistically significant in one out of two districts. Positive externalities are observed among girls with low pre-test scores and for some boys.
Angrist and Lavy (2009)	Israel	Treatment assigned at the school level (among very low performing schools). The programme lasted for three years. The awards were given to high school students. A student who passed all achievement milestones (mainly exams related to obtaining a high school matriculation certificate) could obtain just under $2,400.	Positive results in certification rates for girls (on the order of 0.10 percentage points). The results are mainly driven by the group for whom the certification is "within reach". No effects on boys.
Fryer (2011)	Chicago and New York	The experiment had two cash for grades arms. In Chicago, 9th graders were paid every five weeks upon performance in five core courses. The maximum a student could have won in a year was $2,000. In New York City, 4th and 7th grade students were rewarded based on internal assessments. The maximum amount students could have made in a school-year was $250 and $500, respectively. School-based randomised control trials determined treatment.	In both states, no effects are found in maths or reading achievement tests. In Chicago, marginally significant effects (0.10 of a standard deviation) are observed for grades in the five core subjects. In New York, the effects on the interim assessments are, if anything, negative.
Bettinger (2012)	Ohio	Cash payments (as much as $100 per student) were given to students in the 3rd through to the 6th grade for scoring "proficient" or "advanced" in their (state level) standardised testing. Eligibility by randomisation at the school-grade level.	Positive effects (0.15 of a standard deviation) for maths but no impacts are observed in reading, social science and science test scores.

---

[12] The currency of all the cash transfers described in Table 11 is United States Dollars.

Table 11 (continued): Summary of Cash for Grades Evaluations (Randomised Control Trials) in Primary and Secondary Education

Author(s)	Country/State	Programme Description	Main Results
Levitt et al. (2012)	Chicago	Low-income children and adolescents were offered cash ($10 or $20) for an improvement in a computer test score. These tests lasted between 15 to 60 minutes. Randomisation occurred at the class or school-grade level.	An effect of approximately a tenth of a standard deviation is observed for the $20 incentive. No effects for the $10 transfer. Secondary students are more responsive to the size of the transfer relative to elementary students.
Riccio et al. (2013)	New York	Payments, available for three years, were awarded when low-income households met specific education-based conditions of children. Among these conditions were superior attendance at school and certain performance levels in standardised tests. Each child was rewarded with between $600 and $700 per year for scoring proficient or above. The selection of families or households into the programme was made randomly.	The intervention does not improve outcomes for elementary and middle school students but shows effects among high school students who are more academically prepared than their peers (who entered high school as proficient readers).
Behrman et al. (2015)	Mexico	Mexican high schools with over 40,000 students were assigned to three treatment groups and a control group at random. In one of the treatment groups the payments depended on student performance in mathematics tests in 10th to 12th grade. Payments ranged from $227 up to $1,363 (depending on the level of progress in the tests).	Effects ranging between 0.17 and 0.30 of a standard deviation on maths test scores. All the estimates are statistically significant, but these results are partly explained by students copying.
Hirshleifer (2017)	India	The intervention was carried out among school children (4th through 6th grade) as an experiment (randomisation at the classroom-level). Two tests were implemented. The cash transfer, up to $3, depended on the result of the first test. A second test was only used to measure students' learning.	A positive result (0.24 of a standard deviation relative to the control group) is found but this is not statistically significant due to low statistical power.

Appendix B. Using Relative Ranking in a Regression Discontinuity Design

If I use a local randomisation RD framework, then I should expect not to observe statistically significant differences in the pre-treatment variables across the 0.3 threshold in relative ranking. This logic follows what is commonly shown in experimental designs. I choose two tiny windows of relative ranking (w=0.002 and w=0.005) to implement this RD framework. Therefore, this approach only uses observations whose relative ranking lies within the interval [0.3−w, 0.3+w]. I obtain the RD estimates in this framework from the following regression:

$$X_i = \alpha_3 + \beta_3 I_{3i} + \varepsilon_{3i},$$

where X_i is a pre-treatment variable for student i in year t−1, I_{3i} is a binary variable that takes a value of one if the relative ranking for student i in year t−1 is equal to or lower than 0.3. Otherwise, this variable takes a value of zero. ε_{3i} represents the error term of the regression.

Table 12 shows β_3. Each estimate is equivalent to the difference in means across the threshold. Panel A shows that students with a relative ranking in the [0.298,0.3] interval notably differ compared to those in the]0.3,0.302] interval. The first group, relative to the second, had higher average grades and attendance levels in 2012. These differences are statistically significant at a 99% level of confidence. Additionally, students in the first group were younger, and most likely to be enrolled in primary education, in a rural school and to belong to smaller cohorts relative to their peers. Panel B shows that the differences become smaller as the window w increases, but many estimates remain statistically significant at a 95% level of confidence.

In a continuity-based framework, a discontinuous relationship between the relative ranking and each pre-treatment variable will cast doubt on the plausibility of the continuity assumption. The estimates for the pre-treatment variables in this framework are provided by the regression:

$$X_i = \alpha_4 + \beta_4 I_{4i} + \theta_4 f(\Delta Rel\_Rank_i) + \gamma_4 f(\Delta Rel\_Rank_i)I_{4i} + \varepsilon_{4i},$$

where ΔRel_Rank_i is 0.3-relative ranking in t−1 for student i, I_{4i} is a binary variable that takes a value of one if ΔRel_Rank_i is non-negative. Otherwise, the variable takes a value of zero. Additionally, $f(x)$ is a local polynomial function of x of order p and ε_{4i} models the error term. Table 13 shows β_4. I present two estimates per variable, both using an ad-hoc data-driven bandwidth and small order polynomials. Multiple statistically significant estimates are observed, for example average grade and attendance in 2012, cohort size and rural school.

The evidence I present in Tables 12 and 13 suggests that the variation in BLE treatment cannot be considered as good as random near the 0.3 relative ranking threshold. There are systematic differences between students in the neighbourhood of this cutoff. These differences are very likely to affect potential outcomes and will bias the RD estimates in both frameworks.

The relative ranking is the result of a two-step transformation. In the first step the average grade is transformed into a ranking. In the second step the ranking is divided by the cohort size. This

procedure defines some characteristics for observations around a relative ranking value of 0.3. Table 14 summarises possible values for academic cohort size in relative ranking intervals.

Observations with a relative ranking of 0.3 can only come from a cohort whose size is a multiple of ten. Observations in the [0.298, 0.3[or the]0.3, 0.302] intervals necessarily come from cohorts of at least 57 and 53 students. Expanding the observations to the [0.295, 0.3[and]0.3, 0.305] intervals reduces the minimum cohort sizes to 27 and 23. Therefore, students whose relative ranking is 0.3 come from cohorts with distinctive characteristics relative to those who are very close to this threshold. Table 15 provides summary statistics to prove this point.

Students whose relative ranking is 0.3 belong to cohorts whose average size is 43.63. The average cohort size near this threshold increases sharply. Moreover, other characteristics of the schools and students are substantially different. Students with a relative ranking of 0.3 are more likely to be enrolled in a primary school, to be younger and to attend a rural school. Additionally, these students had a higher academic performance and attendance in 2012.

Tables 14 and 15 provide evidence against comparability between students whose relative ranking is 0.3 and their peers just above and below this threshold. In an RD design for the BLE, students with a relative ranking of 0.3 will be pooled together with those who are slightly below this value. This will attenuate the problems but will not solve them. Overall, there are no comparable observations for students whose relative ranking is 0.3. An additional challenge is that students at each side of the threshold are almost certain to come from different cohorts. Given all this evidence, the relative ranking is not a suitable running variable for an RD design.

Table 10: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework

Pre-Treatment Variables (Var.)	(1) av_grade2012	(2) attendance2012	(3) age	(4) schoolprimary	(5) schoolrural	(6) cohort_size	(7) hmonthincome	(8) hhschooling	(9) hhfemale
Panel A: Size of Window w=0.002									
RD Estimate:	0.094***	1.040***	-0.396***	0.142***	0.206***	-48.89***	-10,488.7*	-0.380**	-0.021
Original Var.	(0.023)	(0.359)	(0.134)	(0.039)	(0.016)	(4.07)	(5,494.6)	(0.160)	(0.025)
Number of Observations	2,348	2,348	2,348	2,348	2,348	2,348	2,271	2,271	2,271
Panel B: Size of Window w=0.005									
RD Estimate	0.046***	0.434**	-0.084	0.033	0.098***	-13.03***	-4,222.5	-0.119	-0.012
Original Var.	(0.014)	(0.213)	(0.076)	(0.020)	(0.012)	(2.40)	(3,374.4)	(0.100)	(0.015)
Number of Observations	5,082	5,082	5,082	5,082	5,082	5,082	4,917	4,917	4,917

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 11: RD Estimates for Pre-Treatment Variables in Continuity-Based Framework (Using Mean Squared Error Optimal Bandwidth)

Pre-Treatment Variables (Var.)	(1) av_grade2012	(2) attendance2012	(3) age	(4) schoolprimary	(5) schoolrural	(6) cohort_size	(7) hmonthincome	(8) hhschooling	(9) hhfemale
Panel A: Order p of Local Polynomial =2 (Local Quadratic Regression)									
RD Estimate: Original Var.	0.025*** (0.010)	0.282** (0.134)	-0.085* (0.051)	0.031** (0.015)	0.084*** (0.011)	-16.79*** (2.18)	-1,374.0 (1,987.0)	-0.126** (0.064)	-0.025*** (0.010)
Number of Observations	97,336	143,886	109,592	97,240	78,797	61,704	126,062	127,140	117,446
Bandwidth Size	0.094	0.139	0.106	0.093	0.075	0.060	0.126	0.127	0.117
Panel B: Order p of Local Polynomial =1 (Local Linear Regression)									
RD Estimate: Original Var.	0.020** (0.008)	0.232* (0.120)	-0.037 (0.036)	0.016 (0.011)	0.081*** (0.010)	-14.29*** (1.82)	-739.5 (1,449.6)	-0.111* (0.057)	-0.020*** (0.007)
Number of Observations	70,027	87,241	101,376	87,304	48,226	42,843	110,176	76,554	102,748
Bandwidth Size	0.068	0.084	0.099	0.084	0.047	0.041	0.110	0.076	0.100

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 12: Theoretical Academic Cohort Size by Intervals of Relative Ranking

Interval of Relative Ranking	Minimum Academic Cohort Sizes	Comments
0.3	10, 20, 30, 40, 50, 60	Only academic cohorts whose size is a multiple of 10 can have observations whose relative ranking equals 0.3.
[0.298, 0.3[	57, 67, 77, 87, 97	First academic cohort size where it is theoretically possible to have observations in each
]0.3, 0.302]	53, 63, 73, 83, 93	of these two intervals of relative ranking is 255.
[0.295, 0.3[	27, 37, 44, 47, 54, 57	First academic cohort size where it is theoretically possible to have observations in each
]0.3, 0.305]	23, 33, 43, 46, 53, 56	of these two intervals of relative ranking is 105.

Table 13: Summary Statistics (Mean Values) by Intervals of Relative Ranking

Variables (Year 2012)	Interval of Relative Ranking				
	0.3	[0.298, 0.3[	]0.3, 0.302]	[0.295, 0.3[	]0.3, 0.305]
Cohort Size	43.63	123.35	115.79	96.01	88.12
Average Grade	5.75	5.64	5.63	5.68	5.66
Attendance (%)	93.81	92.48	92.38	92.76	92.74
Primary School	0.808	0.566	0.595	0.655	0.683
Age (Years)	12.16	12.84	12.76	12.58	12.50
Rural School	0.278	0.060	0.008	0.081	0.062

Source: own calculations using administrative datasets, Chilean ME and MSD

Appendix C. Local Randomisation Regression Discontinuity Framework

<u>Using PFSE Scores as a Running Variable</u>

To implement a local randomisation RD framework, first I need to choose the size w of the window. The size of this window determines which observations of the running variable I use in the estimation. Hence, this RD design relies only on observations within the interval $]98-w, 98+w]$ of PFSE scores. I choose two values of w (w=1 and w=2). The justification behind this selection is twofold. Firstly, these values are the two minimum w available (no decimal places are available for the PFSE scores in my dataset). Secondly, these values are still likely to be small enough for the assumption on which the local experiment RD framework relies to hold. I obtain the RD impact estimates (β_5) in this framework from the following regression:[13]

$$Y_i = \alpha_5 + \beta_5 I_{5i} + \varepsilon_{5i},$$

where Y_i is the average grade or attendance for student i in year t or t+1. I_{5i} is a binary variable that takes a value of one if the PFSE score for student i in year $t-1$ is equal to or lower than 98. Otherwise, this variable takes a value of zero. The error term corresponds to ε_{5i}.

The first four columns of Table 16 show the results for average grade in 2013 and 2014 while the last four columns show the estimates for future attendance. The estimates for future average grade are all negative and statistically insignificant. The estimates range from −0.023 (−0.037σ) to −0.003 points (−0.005σ). The estimates for attendance in 2013 and 2014 are all close to zero and not statistically significant at any level of confidence. The estimates range from −0.244% (−0.026σ) to 0.071% (0.008σ). Overall, all these findings are similar to the ones observed in Table 3. For both types of outcomes, the estimates are closer to zero when w=2. In standard deviations, any estimate higher than 0.06σ will be statistically significant.

Table 17 presents the local randomisation estimates for ten pre-treatment variables. To assess the internal validity of the estimates, I use the same regression but replace Y_i for each pre-treatment variable within X_i'. These variables could not have been affected by the BLE in 2013. This step assesses the quality of the randomisation and, by extension, the internal validity of the causal estimates. The first two columns show the average grade and attendance in 2012. The third to the tenth columns present characteristics of the students, and their schools and households. Overall, the students at each side of the threshold tend not to differ in terms of these ten variables. For both Panels, the joint test of statistical significance of these ten variables is not rejected.

<u>Why can a Local Randomisation RD Framework not be used for Average Grade?</u>

In its original format the average grade is a continuous variable. A local experiment RD framework could be implemented using this variable in a neighbourhood around the threshold. However, previous academic performance is highly correlated with future average grades and attendance.

---

[13] This coefficient equals the difference in means in the threshold's neighbourhood. $\beta_5 = \bar{Y}_{]98-w,98]} - \bar{Y}_{]98,98+w]}$.

Consequently, the critical assumption of an RD local experiment framework is only likely to hold in a very small window around the threshold. Students who differ only by a very tiny fraction in their average grade in 2012 are more likely to be comparable. Unfortunately, the average grade in Chile is rounded at the schools and only one decimal place is reported to the central level. The smallest interval of units available is the $[T−0.1, T+0.1[$ neighbourhood.

Table 18 presents the local randomisation RD estimates. These are equivalent to differences in means between units in the $[T, T+0.1[$ and $[T−0.1, T[$ intervals of average grade in 2012. Among the poorest 30%, all students in the first group received the BLE in 2013 while the second group did not. Every estimate is statistically significant at a 99% level. These coefficients range from 0.102 to 0.106 points for average grade, and between 0.399% and 0.476% for attendance. These results cannot be interpreted as evidence of treatment effects from the BLE. The variation in treatment is not as good as random within the $[T−0.1, T+0.1[$ neighbourhood. Table 19 shows that students at each side of the threshold within this neighbourhood significantly differ regarding key pre-treatment variables such as their percentage of attendance in 2012 and the years of schooling of the head of their household. By construction, the students differ in their average grade in 2012.

Potential outcomes are likely to be correlated with the running variable in the $[T−0.1, T+0.1[$ neighbourhood. Given the characteristics of my dataset, it not advisable to implement a local randomisation framework using average grades as a running variable. The estimates from this framework will not be useful, as they will encompass both treatment effects and selection bias.

Table 14: RD Estimates for Outcomes in Local Randomisation Framework

Outcomes	average_grade2013		average_grade2014		attendance2013		attendance2014	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
RD Estimate: Original Outcome	-0.023 (0.018)	-0.012 (0.013)	-0.012 (0.019)	-0.003 (0.013)	-0.222 (0.233)	-0.146 (0.170)	-0.244 (0.279)	0.071 (0.195)
RD Estimate: In Standard Deviations	-0.037 (0.028)	-0.020 (0.020)	-0.020 (0.030)	-0.005 (0.021)	-0.026 (0.027)	-0.017 (0.020)	-0.026 (0.030)	0.008 (0.021)
Number of Observations	3,539	6,901	3,515	6,840	3,539	6,901	3,515	6,840
Size of Window w	1	2	1	2	1	2	1	2

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 15: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework

Pre-Treatment Variables (Var.)	(1) avg_grade2012	(2) attendance2012	(3) age	(4) male	(5) schoolpub	(6) schoolrural	(7) hmonthincome	(8) hsize	(9) hhschooling	(10) hhfemale
Panel A: Size of Window w=1										
RD Estimate: Original Var.	-0.019 (0.013)	-0.244 (0.192)	0.098* (0.058)	0.008 (0.017)	-0.006 (0.017)	-0.006 (0.010)	-1,532.4 (4,759.5)	-0.014 (0.048)	-0.095 (0.113)	-0.004 (0.017)
Number of Observations	3,567	3,567	3,567	3,430	3,567	3,567	3,430	3,430	3,430	3,430
Panel B: Size of Window w=2										
RD Estimate: Original Var.	-0.009 (0.009)	-0.033 (0.139)	0.064 (0.043)	0.024** (0.012)	-0.003 (0.012)	-0.002 (0.008)	-1,915.9 (3,227.2)	0.002 (0.034)	-0.078 (0.082)	0.002 (0.012)
Number of Observations	6,959	6,959	6,959	6,722	6,959	6,959	6,722	6,722	6,722	6,722

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean Ministry of Education and Ministry of Social Development

Table 16: RD Estimates for Outcomes in Local Randomisation Framework (*w*=0.1)

Outcomes	(1) average_grade2013	(2) average_grade2014	(3) attendance2013	(4) attendance2014
RD Estimates: Original Outcome	0.106*** (0.004)	0.102*** (0.004)	0.399*** (0.067)	0.476*** (0.076)
Number of Observations	56,775	55,659	56,775	55,659

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME & MSD

Table 17: RD Estimates for Pre-Treatment Variables in Local Randomisation Framework (*w*=0.1)

Pre-Treatment Variables (Var.)	(1) average_grade2012	(2) attendance2012	(3) hmonthincome	(4) hhschooling
RD Estimates: Original Var.	0.126*** (0.003)	0.291*** (0.055)	1,716.8* (899.4)	0.085*** (0.027)
Number of Observations	57,806	57,806	55,969	55,969

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME & MSD

Appendix D. Effects of Receiving the BLE in 2014 on Educational Outcomes of 2015

The body of the paper shows no statistically significant results for the BLE in 2013. A potential explanation for these results is that students were not aware enough about the implementation of the BLE in 2013. This was less likely to be the case after one year. Students who received the BLE in 2014 could have increased their effort and performance in 2015. This appendix explores the effects of receiving the BLE in 2014. In 2014 the BLE was implemented during September, three months before the end of the academic year. Given this late implementation, I only consider outcomes in 2015 in this assessment.

In practice, the estimates for the BLE in 2014 remain not statistically different than zero. For average grade, Panel A in Table 20 shows that the estimates range from −0.016 to 0.004 points. The standard errors of these estimates range from 0.010 to 0.020 points. Panel B in Table 20 presents the results for attendance. The estimates range from −0.042% to 0.041% while the standard errors are between 0.132% and 0.279%. The findings for the BLE in 2014 are like those in 2013. I find no statistically significant effects. Consequently, if an effect exists for the BLE, it is likely to be small in magnitude and cannot be detected statistically.

Table 18: RD Estimates for Outcomes in 2015

RD Framework and Running Variable (RV)	Local Randomisation RV: PFSE		Continuity-Based RV: PFSE		Continuity-Based RV: Average Grade	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Estimates for Average Grade in 2015						
RD Estimates:	-0.010	-0.007	-0.016	-0.013	-0.002	0.004
	(0.020)	(0.014)	(0.012)	(0.011)	(0.013)	(0.010)
Number of Observations	3,094	6,023	103,352	67,235	133,102	89,748
Window *w* / Bandwidth *h*	1.00	2.00	33.84	21.77	0.300	0.200
Panel B: Estimates for Attendance in 2015						
RD Estimates	-0.009	-0.025	0.039	0.041	-0.042	0.021
	(0.279)	(0.197)	(0.185)	(0.132)	(0.205)	(0.152)
Number of Observations	3,094	6,023	82,467	82,467	133,102	89,748
Window *w* / Bandwidth *h*	1.00	2.00	27.40	27.14	0.300	0.200
Order *p* of Polynomial	NA	NA	2	1	2	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean ME & MSD

Appendix E. Effects of Receiving the BLE in 2013 for Population Subgroups

I analyse the effect of receiving the BLE in 2013 on educational outcomes in 2014 for multiple subgroups. I divide the first two groups by income. The hypothesis to test in this case is that the BLE is most likely to have an impact at the lower end of the income distribution. There is limited information on this question for cash for grades programmes, though Galiani and McEwan (2013) and Maluccio and Flores (2005) find in Honduras and Nicaragua, respectively, that conditional cash transfers have a stronger impact on school enrolment among children living in the poorest households.

Analysis by Income (PFSE Scores)

Table 21 shows RD estimates for average grade and attendance in 2014 using average grade in 2012 as a running variable. The first row of Table 21 presents the results without using income subgroups. These results are equivalent to those shown in Table 7. Among those with a PFSE score equal to or lower than 98 points, I divide the sample into two. The first half is meant to include the poorest (approximately the bottom 15% in terms of income) while the other represents the second most deprived group (approximately between the bottom 15% to 30% in the income distribution).

Table 19: RD Estimates for Outcomes in 2014 by Halves of Lowest PFSE Scores in Continuity-Based Framework (Using Average Grade in 2012 as a Running Variable)

Subgroup	Outcomes			
	Average Grade in 2014		Attendance in 2014	
Total	0.010	0.017*	0.119	0.147
	(0.012)	(0.009)	(0.188)	(0.141)
1st Poorest Half	0.020	0.023**	0.234	0.277
	(0.016)	(0.012)	(0.257)	(0.185)
2nd Poorest Half	-0.000	0.010	-0.008	0.010
	(0.016)	(0.012)	(0.256)	(0.185)
Bandwidth h	0.300	0.200	0.300	0.200
Order p of Polynomial	2	1	2	1

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Source: own calculations using administrative datasets, Chilean ME and MSD

Table 21 shows that the first group has more positive estimates than the second group. The impact estimates for average grade in 2014 are 0.020 and 0.023 for the poorest. Only one estimate is statistically significant at a 95% level of confidence. Therefore, the significance is not robust to alternative specifications. The impact estimates for attendance in 2014 are 0.234% and 0.277% for the poorest, neither of which is statistically significant. Conversely, the RD estimates for the second poorest half are close to zero for both types of outcomes.

There is not enough evidence to claim that the BLE in 2013 had a statistically significant effect in 2014 on the poorest of the population. Given the standard errors shown, 0.012 and 0.016 for average grade and 0.185% and 0.257% for attendance, if an effect of the BLE exists on the poorest then it is at most modest in size and could not be captured consistently with statistical certainty.

Analysis by Gender and Educational Level

The next two groups are boys and girls. The final two groups of students are those who were either in the fifth or sixth grade and between seventh and tenth grade in 2012, respectively. In 2014 the former group was most likely to be enrolled in primary education while the latter group was most likely to be enrolled in secondary education. These two pairs of groups are justified in terms of the cash for grades literature, where heterogeneous results have been observed.

Table 22 and Table 23 present RD estimates for average grade and attendance in 2014, respectively. The first four columns in each table provide estimates that use PFSE scores as a running variable. The first and second columns focus on the local randomisation framework, while the third and fourth rely on the continuity-based framework. The last two columns of each table provide estimates for the continuity-based framework using the distance of average grade in 2012 as a running variable. The first row of each table provides the estimates without using subgroups.

These estimates are equal to those presented in Tables 3, 7 and 16.

From Table 22 it can be seen that girls have less negative estimates than boys in all the specifications and frameworks I use and that students between the seventh and tenth grades have mostly lower estimates relative to younger students. Only the local linear regressions using average grade as a running variable provide some statistically significant estimates. The estimates by subgroup mostly remain close to zero and are statistically insignificant. Overall, these estimates are not very different from those of the entire sample.

There is less consistent behaviour by subgroup for attendance in 2014. Table 23 shows that the RD estimates for boys are lower when I use the PFSE as a running variable but higher when the estimations rely upon previous academic performance. I observe a similar case for students in the fifth and sixth grades in 2012 relative to their peers in higher grades. Two estimates for the former subgroup show some degree of statistical significance in one type of RD framework.

Overall, the estimates by gender and educational level remain close to zero and are statistically insignificant for both outcomes. Naturally, the analysis by subgroup has wider confidence intervals given the smaller samples. The standard errors for average grade vary from a maximum of 0.029 points to a minimum of 0.012 points. Thus, the analysis by the subgroups of gender and educational level could have identified any estimate higher than 0.057 and as low as 0.024 points in average grade as statistically significant. Concerning attendance, the standard errors of the estimates range from 0.181% to 0.406%. Differences at the threshold higher than 0.800% and as low as 0.355% could have been statistically significant then.

For both types of outcomes, in the few cases where I observe statistically significant estimates, the significance is sensitive to the RD framework I use. Additionally, within each RD framework, the significance is also sensitive to the specification I utilise. Given my results and analysis, if an effect of the BLE exists by gender and educational level it is highly unlikely to be large. Any potential effect of this kind could not be captured with statistical certainty in this assessment.

Table 20: RD Estimates for Average Grade in 2014 (by Subgroups)

Subgroups	RD Framework and Running Variable					
	Local Randomisation RV: PFSE		Continuity-Based RV: PFSE		Continuity-Based RV: Average Grade	
	(1)	(2)	(3)	(4)	(5)	(6)
Total	-0.012	-0.003	-0.008	-0.003	0.010	0.017*
	(0.019)	(0.013)	(0.013)	(0.010)	(0.012)	(0.009)
Boys	-0.030	-0.011	-0.009	-0.010	0.010	0.017
	(0.029)	(0.021)	(0.020)	(0.015)	(0.017)	(0.013)
Girls	-0.003	0.006	-0.004	0.003	0.018	0.024**
	(0.026)	(0.018)	(0.017)	(0.012)	(0.016)	(0.012)
5th to 6th Grade in 2012	0.008	-0.002	-0.000	0.007	0.023	0.028**
	(0.029)	(0.020)	(0.019)	(0.014)	(0.017)	(0.013)
7th to 10th Grade in 2012	-0.017	-0.001	-0.011	-0.009	0.008	0.014
	(0.024)	(0.017)	(0.017)	(0.013)	(0.016)	(0.012)
Window w / Bandwidth h	1.00	2.00	24.96	24.87	0.300	0.200
Order p of Polynomial	NA	NA	2	1	2	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean ME & MSD

Table 21: RD Estimates for Attendance in 2014 (by Subgroups)

Subgroups	RD Framework and Running Variable					
	Local Randomisation RV: PFSE		Continuity-Based RV: PFSE		Continuity-Based RV: Average Grade	
	(1)	(2)	(3)	(4)	(5)	(6)
Total	-0.244	0.071	0.108	0.106	0.119	0.147
	(0.279)	(0.195)	(0.159)	(0.143)	(0.188)	(0.141)
Boys	-0.359	-0.049	-0.053	-0.060	0.268	0.142
	(0.391)	(0.279)	(0.236)	(0.214)	(0.270)	(0.205)
Girls	-0.164	0.101	0.178	0.177	-0.044	0.114
	(0.406)	(0.272)	(0.210)	(0.187)	(0.257)	(0.184)
5th to 6th Grade in 2012	-0.197	0.022	-0.012	-0.008	0.438*	0.388**
	(0.382)	(0.284)	(0.215)	(0.192)	(0.254)	(0.181)
7th to 10th Grade in 2012	-0.201	0.130	0.164	0.160	0.021	0.067
	(0.362)	(0.249)	(0.208)	(0.187)	(0.246)	(0.185)
Window w / Bandwidth h	1.00	2.00	35.58	21.12	0.300	0.200
Order p of Polynomial	NA	NA	2	1	2	1

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Source: own calculations using administrative datasets, Chilean ME & MSD

References

Angrist, J., & Lavy, V. (2009). The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial. American Economic Review, 99(4), 1384-1414.

Behrman, J., Parker, S., Todd, P., & Wolpin, K. (2015). Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. Journal of Political Economy, 123(2), 325-364.

Bettinger, E. (2012). Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores. Review of Economics and Statistics, 94(3), 686-698.

Biblioteca del Congreso Nacional de Chile. (2013). Decreto 24. Aprueba Reglamento que Regula el Bono por Esfuerzo.

Bushaw, W. J., & Lopez, S. J. (2010). Highlights of the 2010 Phi Delta Kappa/Gallup Poll. What Americans Said About the Public Schools. https://larrycuban.files.wordpress.com/2010/11/2010_poll_report1.pdf.

Calefati, J. (2008). Giving Students Cash for Grades. US News. Retrieved from https://www.usnews.com/education/articles/2008/11/28/giving-students-cash-for-grades

Cameron, J. (2001). Negative Effects of Reward on Intrinsic Motivation—A Limited Phenomenon: Comment on Deci, Koestner, and Ryan (2001). Review of Educational Research, 71(1), 29–42.

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues. The Behavior Analyst, 24(1), 1-44.

Cameron, J., & Pierce, W. D. (1994). Reinforcement, Reward and Intrinsic Motivation: A Meta-Analysis. Review of Educational Research, 64(3), 363– 423.

Carneiro, P., Galasso, E., & Ginja, R. (forthcoming). Tackling Social Exclusion: Evidence from Chile. The Economic Journal.

Cattaneo, M., Idrobo, N., & Titiunik, R. (2018a). A Practical Introduction to Regression Discontinuity Designs: Part I. In Cambridge Elements: Quantitative and Computational Methods for Social Science. Cambridge, UK: Cambridge University Press.

Cattaneo, M., Idrobo, N., & Titiunik, R. (2018b). A Practical Introduction to Regression Discontinuity Designs: Part II. In Cambridge Elements: Quantitative and Computational Methods for Social Science. Cambridge, UK: Cambridge University Press.

Cattaneo, M., Keele, L., Titiunik, R., & Vazquez-Bare, G. (2016a). Interpreting Regression Discontinuity Designs With Multiple Cutoffs. Journal of Politics, 78(4), 1229-1248.

Cattaneo, M., Titiunik, R., & Vazquez-Bare, G. (2017b). 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.

Comité de Expertos Ficha de Protección Social. (2010). Informe Final Comité de Expertos Ficha de Protección Social. http://www.ministeriodesarrollosocial.gob.cl/btca/txtcompleto/mideplan/c.e-fps-infinal.pdf.

De la Mata, D. (2012). The Effect of Medicaid Eligibility on Coverage, Utilization, and Children's Health. Health Economics, 21(9), 1061-1079.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. Psychological Bulletin, 125(6), 627– 668.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again. Review of Educational Research, 71(1), 1–27.

Dong, Y. (2015). Regression Discontinuity Applications With Rounding Errors in the Running Variable. Journal of Applied Econometrics, 30(3), 422–446.

Frandsen, B. (2017). Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable Is Discrete. In M. D. Cattaneo & J. C. Escanciano (Eds.), Advances in Econometrics (Vol. 38, pp. 281-315). Bingley, UK: Emerald Publishing Limited.

Fryer, R. (2011). Financial Incentives and Student Achievement: Evidence from Randomized Trials. Quarterly Journal of Economics, 126(5), 1755-1798.

Galiani, S., & McEwan, P. J. (2013). The Heterogeneous Impact of Conditional Cash Transfers. Journal of Public Economics, 103, 85-96.

Gneezy, U., Meier, S., & Rey-Biel, P. (2011). When and Why Incentives (Don't) Work to Modify Behavior. Journal of Economic Perspectives, 25(4), 191-210.

Guttenplan, D. D. (2011). Motivating Students With Cash-for-Grades Incentive. The New York Times. Retrieved from https://www.nytimes.com/2011/11/21/world/middleeast/21iht-educLede21.html

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects With a Regression-Discontinuity Design. Econometrica, 69(1), 201-209.

Higgins, L. (2015). Think and Grow Rich? Michigan School Offers Cash for Grades. USA Today. Retrieved from https://www.usatoday.com/story/news/nation-now/2015/11/01/michigan-high-school-cash-for-grades/75003438/

Hirshleifer, S. (2017). Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance. https://economics.ucr.edu/repec/ucr/wpaper/201701.pdf.

Imbens, G. W., & Lemieux, T. (2008). Regression Discontinuity Designs: A Guide to Practice. Journal of Econometrics, 142(2), 615-635.

Kohn, A. (1999). Punished by Rewards: The Trouble With Gold Stars, Incentive Plans, A's, Praise, and Other Bribes. Boston: Houghton Mifflin.

Kolesár, M., & Rothe, C. (2018). Inference in Regression Discontinuity Designs With a Discrete Running Variable. American Economic Review, 108(8), 2277-2304.

Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to Learn. Review of Economics and Statistics, 91(3), 437-456.

Lee, D. (2008). Randomized Experiments from Non-Random Selection in U.S. House Elections. Journal of Econometrics, 142(2), 675-697.

Lee, D., & Card, D. (2008). Regression Discontinuity Inference With Specification Error. Journal of Econometrics, 142(2), 655–674.

Levitt, S. D., List, J. A., Neckermann, S., & Sadoff, S. (2012). The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance. https://www.nber.org/papers/w18165.pdf.

Lindo, J. M., Sanders, N. J., & Oreopoulos, P. (2010). Ability, Gender, and Performance Standards: Evidence from Academic Probation. American Economic Journal: Applied Economics, 2(2), 95-117.

Maluccio, J., & Flores, R. (2005). Impact Evaluation of a Conditional Cash Transfer Program: The Nicaraguan Red de Protección Social. http://www.ifpri.org/publication/impact-evaluation-conditional-cash-transfer-program-2.

McCrary, J. (2008). Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. Journal of Econometrics, 142(2), 698-714.

Pop-Eleches, C., & Urquiola, M. (2013). Going to a Better School: Effects and Behavioral Responses. American Economic Review, 103(4), 1289–1324.

Reardon, S. F., & Robinson, J. P. (2012). Regression Discontinuity Designs With Multiple Rating-Score Variables. Journal of Research on Educational Effectiveness, 5(1), 83-104.

Riccio, J., Dechausay, N., Miller, C., Nunez, S., Verma, N., & Yang, E. (2013). Conditional Cash Transfers in New York City: The Continuing Story of the Opportunity NYC-Family Rewards Demonstration. https://www.mdrc.org/sites/default/files/Conditional_Cash_Transfers_FR%202-18-16.pdf.

Ripley, A. (2010). Should Kids Be Bribed to Do Well in School? Time Magazine.

Roberts, D., Becker, C., & Ibanga, I. (2008). Chicago Offers Students Cash for Good Grades. ABC News. Retrieved from https://abcnews.go.com/GMA/Parenting/story?id=6371073&page=1

Sekhon, J., & Titiunik, R. (2017). On Interpreting the Regression Discontinuity Design as a Local Experiment. In M. D. Cattaneo & J. C. Escanciano (Eds.), Advances in Econometrics (Vol. 38, pp. 1-28). Bingley, UK: Emerald Publishing Limited.

Sidorkin, A. M. (2007). Is Schooling a Consumer Good? A Case Against School Choice, But Not the One You Had in Mind. Philosophy of Education, 75–83.

Sidorkin, A. M. (2009). Labor of Learning: Market and the Next Generation of Educational Reform. Rotterdam: Sense.

Toppo, G. (2008). Good Grades Pay Off Literally. USA Today. Retrieved from https://usatoday30.usatoday.com/news/education/2008-01-27-grades_N.htm

Warnick, B. (2017). Paying Students to Learn: An Ethical Analysis of Cash for Grades Programmes. Theory and Research in Education, 15(1), 71–87.

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables: A Comparative Study of Four Estimation Methods. Journal of Educational and Behavioral Statistics, 38(2), 107-141.