FACULTAD ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE
DEPARTAMENTO DE ECONOMÍA

SERIE DE DOCUMENTOS DE TRABAJO

SDT 429

The Impact of Grade Retention on Juvenile Crime

Autores:
Juan Díaz
Nicolás Grau
Tatiana Reyes
Jorge Rivera

Santiago, Septiembre de 2016

sdt@econ.uchile.cl
econ.uchile.cl/publicaciones

The Impact of Grade Retention on Juvenile Crime

Juan Diaz
Harvard University

Nicolas Grau
University of Chile

Tatiana Reyes
University of Chile

Jorge Rivera
University of Chile

This version: June, 2016

Abstract

Using detailed administrative and individual data on schooling and crime records from Chile, we estimate the effect of grade retention between 4th and 8th grade on juvenile crime. We base our research on the rule which specifies that students who fail more than one subject must repeat the year. We present two empirical strategies to address the strong evidence that the forcing variable is – locally – manipulated. First, we follow Barreca, Guldi, Lindo, and Waddell (2011) in implementing a *donut-hole* fuzzy regression discontinuity design (FRD). Second, we extend the approach developed by Keele, Titiunik, and Zubizarreta (2015) to implement a method that combines matching with FRD. These two methodologies deliver similar results and neither show a statistically significant effect on a placebo test. According to our results, grade retention increases the probability of juvenile crime by 1.6 percentage point (pp), an increase of 33% (higher for males and low SES students). We also find that grade retention increases the probability of dropping out by 1.5pp. Regarding mechanisms, our findings suggest that the effect of grade retention on crime does not only manifest itself indirectly as a result of its effect on dropping out. Furthermore, the effect of grade retention on crime is worsened when students switch schools right after failing the grade.

Keywords: Juvenile Crime, Grade Retention, Regression Discontinuity, and Matching.
JEL Classification: I21, K42, and C26 .

We thank Claudio Ferraz, Guido Imbens, Jose Zubizarreta, Miguel Urquiola, Zelda Brutti and seminar participants at Universidad de Chile and Sao Paulo School of Economics for valuable comments and suggestions. Juan Diaz and Nicolás Grau thank the Centre

for Social Conflict and Cohesion Studies (CONICYT/FONDAP/15130009) for financial support. All remaining errors are our own.

1 Introduction

Grade retention increases the education opportunity cost of repeaters, thereby illegal activities become more attractive (see, for instance, Lochner (2004)). Consequently, regardless of the grade retention controversy,[1] merely by an argument *"a la Becker"* there should not be debate about the positive effects that such a policy would have over the youth delinquency. However, what seems to be clear from the incentive side, as of yet does not have an overwhelming counterpart from the empirical front. This paper aims to contribute to this literature, by providing causal evidence on the positive effect that grade retention has on the level of youth delinquency.

Since the evidence has shown that the repeaters are more likely to have a lower innate ability and weaker social background than promoted students (see, for instance, Barrington and Hendricks (1989)), there is a challenge in estimating causal effects of grade retention on crime, given that the latent outcome – crime activity that would be observed in the absence of grade retention – and the propensity to fail a grade are simultaneously determined. A proper data-set to address this challenge is not always available by the researchers (indeed, rarely), and this could be the main reason behind the scarce literature on the estimation of the causal effect we are concerned with at this paper. Our investigation is based on administrative data on both academic records and crime committed by youths in Chile, individually recorded, and properly matched, for the entire population of them during the period 2007 – 2014.

Our identification strategy relies on a grade retention rule that creates a discontinuity on the probability of grade retention. This rule specifies that students who fail two or more subjects should repeat the grade, which appears to be an ideal situation for the implementation of standard RD methods. However, there is empirical and anecdotal evidence showing that the forcing variable (the student's second lowest score) is manipulated, because it is arguable that teachers' grading decisions at the margin of repetition may not sort students randomly.[2] To deal with this problem, we undertake two com-

[1] That is, the existence of ambiguous evidence, and even contradictory, over the effects that grade retention may has over some academic outcomes and/or socio-emotional variables of the youths. See Holmes et al. (1989), Jimerson (2001) and Rose, Medway, Cantrell, and Marus (1983); see also Reschly and Christenson (2013) for a fresh look on this debate.

[2] For the purposes of this paper, to avoid confusion with the concept of grade (level) with grade (performance) the latter will be referred to as "score".

plementary empirical strategies. In the first approach, we follow Barreca, Guldi, Lindo, and Waddell (2011) in implementing a *donut-hole* FRD, where, after removing observations in the immediate vicinity of the threshold for grade repetition, we run a standard FRD. This method delivers causal evidence to the extent that the manipulation is a local phenomenon, which is partially supported by the data, and to the extent that removing observations does not invalidate the RD assumption about the continuity of the outcome variable's expectation around the threshold, conditional to the same treatment status. In the second and preferred approach, we address the latter potential problem (that arises due to removing observations), by extending the method developed by Keele, Titiunik, and Zubizarreta (2015) to implement an estimation procedure that combines matching with fuzzy regression discontinuity design (FRD).

In our preferred specification (the RD-matching), the results show that grade retention between 4th and 8th grade increases the probability of juvenile crime by 1.8 pp, an increase of 37.5%.[3] Furthermore, we also examine the effect of grade retention on dropping out and future grade retention. According to our results, grade retention in primary school decreases the probability of grade retention by 5.9 pp (10.7%) in subsequent years and increases the probability of dropping out by 1.6pp (23.8%).[4] Thus, if we assume that grade retention in higher grades also impacts juvenile crime, then the effects on future grade retention suggest that we have found a lower bound for the effect of primary-school grade retention on crime, because those who did not repeat in primary school (who are *non-treated* in our estimation) had a higher probability of grade retention in the future, which also creates an impact on crime. To study the robustness of our results, we implement a placebo test by replicating the "Donut-hole" RD and the RD-matching estimations, but in this case we only consider students who did not repeat the grade, comparing those who scored below, with those whose scored above, the threshold. These two methods do not deliver a statistically significant effect in this placebo test for all three outcomes considered.

The literature and our paper have shown the relationship between grade retention and

[3]In order to study the robustness of our results, we implement a placebo test by replicating the "Donut-hole" RD and the RD-matching estimations, but in this case we only consider students who did not repeat the grade, comparing those who scored below, with those whose scored above, the threshold. These two methods do not deliver a statistically significant effect in this placebo test for all three outcomes considered.

[4]The effect on drop out is only statically significant for low SES students.

high school drop out.[5] This evidence, together with the studies showing the impact of high school drop out on crime,[6] could potentially place in doubt the novelty of our research: why to study the impact of grade retention on crime if we already know the effect of grade retention on school drop out and the effect of school drop out on crime? However, our findings suggest that the effect of grade retention on crime does not only operate through its effect on dropping out. Indeed, we find that the effect of grade repetition on crime occurring before (or simultaneously to) dropping out is found to be more relevant than the effect on crime that occurs after dropping out. We also show that the effect of grade retention on crime is worsened when students switch school right after failing the grade.

To the best of our knowledge, the closest to our paper was provided by Depew and Eren (2015), who estimate the impact of grade retention (with summer school) on juvenile delinquency (and school dropout) in Louisiana.[7] They assemble a novel data set after merging administrative information of educational outcomes with the criminal records of students attending schools in Louisiana. Then, taking advantage of the test-based grade promotion policy that has been applied in Louisiana as of a decade ago, the authors build an RD design, where the forcing variable is the score on a standardized test which determines whether or not a student is promoted. Their principal conclusion is that, for students attending eighth grade, the test-based grade retention policy decreases the likelihood of being involved in felony offences during their youth. Although the authors make a remarkable effort in identifying a causal effect of grade retention on juvenile delinquency, they do not correct the latent manipulation that the forcing variable suffers close to the cut off.[8]

Our paper is also related with Cook and Kang (2013), who merge administrative data of

[5]See King, Orazem, and Paterno (2015), Manacorda (2012), and Roderick (1994).

[6]See Anderson (2014), Fagan and Pabon (1990), and Thornberry, Moore, and Christenson (1985).

[7]At this part it is worth mentioning that there is a tangential to us literature where is investigated a sort of inverse problem than here, i.e., how criminal activities (either within the school or outside) affect some schooling outcomes. See, for instance, Burdick-Will **?**, Fagan and Pabon **?**, Hirschfield **?**.

[8]Indeed, the key assumption in the RD they run is that teachers (or someone else in charge) do not exercise precise control over the score in the standardized test near the cut-off point. If this holds, the variation in scores obtained at the threshold is as good as randomized (Imbens and Lemieux (2008) and Lee and Lemieux (2010)). Nevertheless, as in our own RD design, we believe that this essential assumption does not hold. As the figures A1 and A2 (page 47) suggest, it seems that a strategic allocation of students occurs around the cut-off score, as there are no students who score marginally below the minimum required.

academic performance with the criminal record of students attending public schools in North Carolina. They exploit the sharp RD design generated by the specific date which establishes the minimum age for school enrollment (cut date) and assess its effect on a number of educational outcomes, as well as on juvenile crimes committed. They present two main findings. First, during middle school, students born just after the cut date (the oldest) are more likely to outperform (in math and reading) those born just before (the youngest), and are less prone to be involved in juvenile delinquency. Second, those born after the cut date are more likely to drop out of school and commit a severe offence. Our research differs on at least two dimensions: we attempt to directly estimate the causal effect of grade retention on juvenile crime and instead of using the student's birthdate as a forcing variable, we employ the student's second lowest score to generate a fuzzy RD design, which in turn we improve by combining it with a matching method.

Our paper makes three main contributions. First, together with (Depew and Eren (2015)), it is the first that estimates a causal effect of grade retention on juvenile crime and it is the first evidence for a developing country, where the retention rates are higher.[9] Second, by extending the method developed by Keele, Titiunik, and Zubizarreta (2015) to the *fuzzy RD* case, we present a method that can be useful in many other contexts, where there is some evidence of manipulation in the forcing variable. Third, it sheds light on the mechanisms that explain the impact of grade retention on juvenile crime.

The paper proceeds as follows. Section 2 describes the main features of the data and the grade retention rule. Section 3 discusses the potential problems associated with implementing a standard RD approach. Section 4 presents the empirical approaches used in this paper. Section 5 details the results. Section 6 discusses possible mechanisms behind these results. Finally, Section 7 concludes and discusses future research.

2 Data and Grade Retention's Rule

In this section, we first describe the characteristics of our data set and then we explain the way that the grade retention rule operates.

[9] See Manacorda (2012).

2.1 Data

In this paper, we assemble administrative data sets from the Ministry of Education and the *Defensoria Penal Publica* (DPP). The DPP is the institution in Chile which provides free legal representation, for those who have been accused of committing a crime. The final data set includes over 1.2 million students and their juvenile criminal records (ages from 12 to 18) linked to a large set of demographic characteristics.

The information, collected from the Ministry of Education, is an administrative panel data set from 2002 to 2015, which for every student in the country, indicates the school attended every year, the grade level (and whether the student has repeated the grade), the student's attendance rate, some basic demographic information, and only for 2007 the annual average score for all subjects taken by the student (their cumulative grade point average). The latter is needed to establish which students are close to the threshold for grade repetition. We merge this panel with the information on performance on the national standardized test (the SIMCE), which is taken annually by all students in the 4th grade and every other year by all 8th grade students. When students take the SIMCE, a survey is administered to their parents. From these surveys, we obtain information about the mother's and father's education level and family income. We focus our attention on the students who, in 2007, were in 4th to 8th grade, attending public or subsidized schools.[10] Due to their high SES and low criminal rate, we excluded students attending private schools, which represent 8% of the national enrollment.

The DPP's records contain information on all defendants in criminal cases tried in Chile during the period of January, 2006 to December, 2014. This database includes information at the time of the accusation, the type of offence, and the verdict (including the length of the sentence). In this study, we consider only juvenile criminal cases and in order to focus on crimes that can be thought of as motivated by a cost-benefit analysis, we omit individuals who committed the most severe crimes, such as murder or rape.[11] Given that our "treatment" is grade retention in 2007, we also exclude students who were prosecuted before 2008. Thus, in all our estimations, the students who committed crimes are those who were prosecuted, between 2008 and 2014, for an offence with an *economic motivation*.

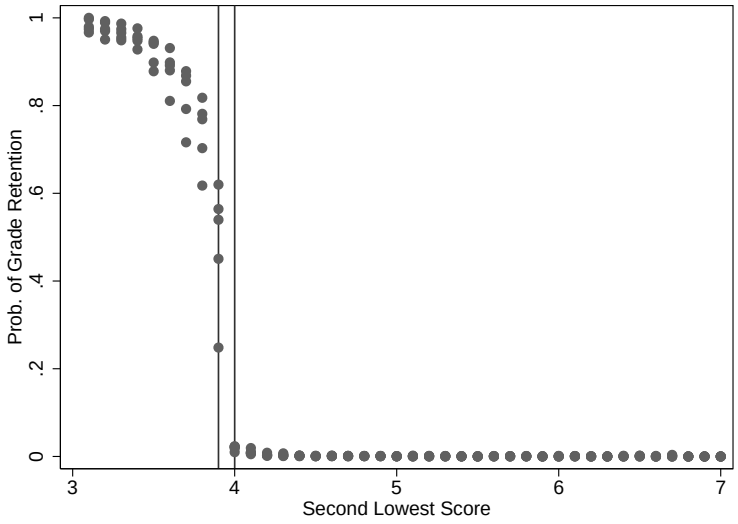
[10] We do not have SIMCE information for those students who were attending 7th grade in 2007. Thus, most of our estimations do not consider this group.

[11] We do not consider as crime the juvenile criminal cases where the verdict was *not guilty*.

2.2 Rules for Grade Retention

In Chile, student scores range from 1 to 7, with an increment of 0.1. Although there are other causes of grade retention, the most prevalent is scoring below 4, on two or more subjects (≤ 3.9). This rule suggests the possibility of implementing a regression discontinuity approach to study the causal effect of grade retention on crime. In this regard, Figure 1 shows a strong discontinuity in the probability of grade retention between 3.9 and 4. Each dot represents the grade retention rate of all the students in a particular grade who have a specific value on the second -lowest score.[12]

Figure 1: Grade retention rule



Note: This figure considers only schools which have at least one student scoring 4 or 4.1 and at least one student scoring 3.9 or 3.8 in their second-lowest score.

Although all schools must apply the 1-7 grading scale, they are free to set their own grading standards, which means that scores are not comparable across schools. This institutional feature explains why, in all of our estimations, we compare students – below and above the threshold – who attend the same school (and the same grade). This is also why, in all the plots that we present, we consider only students attending schools

[12]The main reason why this is not a sharp discontinuity, is because students who have two scores below 4 can pass the grade as long as their average across all subjects is greater or equal to 5.

with at least one student scoring 4 or 4.1 and at least one student scoring 3.9 or 3.8 in their second-lowest score.

3 Validity of the RD Design

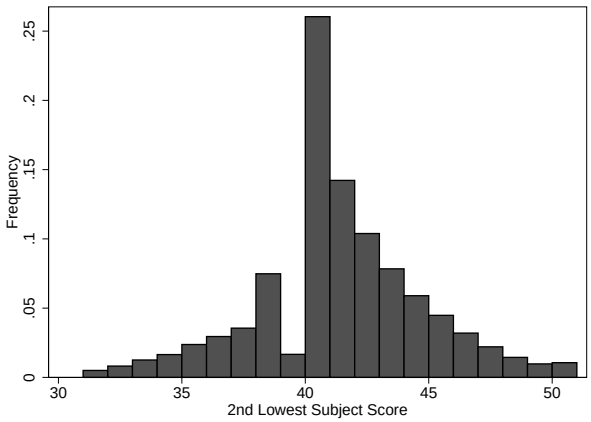
As we discuss in the following paragraphs, there are institutional reasons and empirical evidence to support the idea that the forcing variable – the second-lowest score – is manipulated around the threshold. However, we argue and also present evidence that this problem could be restricted to the scores closest to the threshold. The existence of this manipulation problem, and its local nature, is what determines our empirical strategies to estimate causal effects.

3.1 The Density of the Forcing Variable

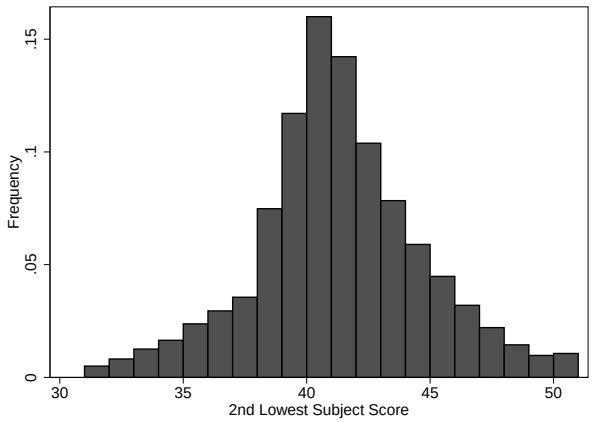
Figure 2 shows two histograms for the second lowest score. Panel (a) presents the real histogram (derived from data), while in panel (b) a hypothetical histogram is introduced, which is created from panel (a) by only moving a number of students from scoring 4 to 3.9. There are two lessons to be gleamed from these plots. First, there is a remarkable discontinuity in the histogram for the second-lowest score, around the threshold $(3.9-4)$. Second, the discontinuity (and possibly the manipulation) seems to be limited to the scores closest to the threshold $(3.9-4)$, in fact the histogram shown in panel (b) does not show any evidence of discontinuity.

Figure 2: Histograms for the 2nd Lowest Score

(a) Actual



(b) Hypothetical



The first point raises reasonable doubts about the internal validity of an RD estimator (see Lee and Lemieux (2010)), because it is arguable that teachers' grading decisions at the margin of repetition may not sort students randomly. The second point, which addresses local manipulation, is in line with the incentives that teachers face. In fact, even though the anecdotal evidence suggests that school leaders promote an upper bound for the rate of grade retentions, and, therefore, teachers may be *forced* to pass students who have a *real* score lower than 4, there is no reason to raise that score to a value higher than 4.[13] Moreover, if a student's real score (a latent variable) was 3.9 and to avoid any complain from their parents (asking for an small increase to pass the grade), her teacher manipulates that score grading 3.8, that is going to make our treatment and control groups more comparable. The relevance of this point is going to be more clear when we introduce our empirical strategy.

3.2 Graphical Test for Local Manipulation

We present direct evidence of local manipulation by taking advantage of the richness of our database. Intuitively, local manipulation should imply that the mapping from knowledge (a latent variable) to scores should be discontinuous around the threshold. In

[13]Teachers' grading behavior is not audited to find evidence of manipulation in their grading.

our framework, this means there should be a discontinuity in such a mapping between grades 3.9 and 4.

Fortunately, besides students' GPA at school, we have information on their standardized test scores (the SIMCE), where the latter can be thought as unbiased proxies of student's knowledge. Thus, we can test manipulation by studying the behavior of the mapping from SIMCE to GPA around the threshold, at each primary school. We do so in the following steps:

- To have the closest possible relationship between standardized tests and grade scores, we focus on the math Simce and math GPA for 8th grade students.[14]

- Let i index students, we run the following OLS regression for each school s:

$$MathSimce_{is} = \mu_0^s + \mu_1^s * MathGPA_{is} + \upsilon_{is}.$$

- We allow for a different mappings for each school, because schools may have different standards to grade their students.[15] Furthermore, to have enough precision in our estimated parameters, we exclude schools with less than 20 students. By doing so we drop 1625 schools, keeping 4125 for our OLS estimations.

- Given 4125 pairs of OLS estimations for μ_0 and μ_1, we calculate the residual for each student i, such that:

$$Residual_{is} = MathSimce_{is} - \hat{\mu}_0^s - \hat{\mu}_1^s * MathGPA_{is}.$$

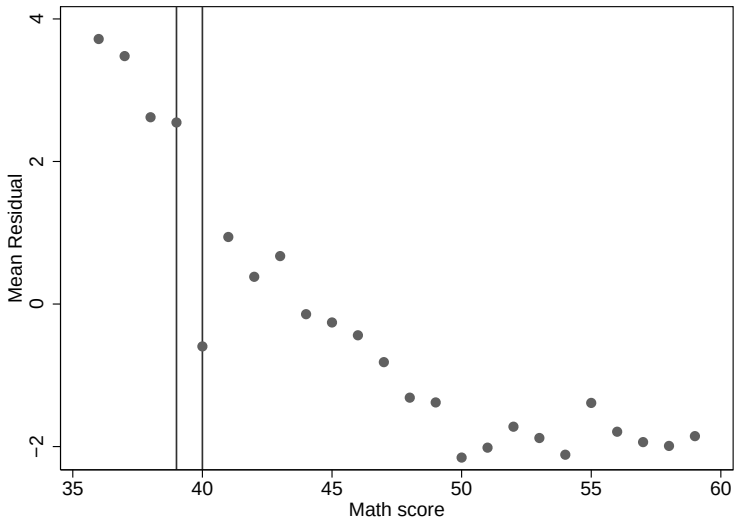
- In Figure 3 we present the mean of these residuals for each value of $MathGPA$. As can be seen in this figure, and even though there are other smaller jumps in

[14]To be clear, this means that the sample that we use to show local manipulation is different from the sample that we consider in our estimations of the effect of grade retention on crime. While in the former we only use 8th grade students and their math performance, in the latter we consider data from 4th to 8th grade and their average performance across subjects. Because we do not have standardized tests measuring all subject, we are constrained to show local manipulation in math and to assume that it is also local for the other subjects.

[15]In Figure 4 of Appendix A, we show how different are μ_0 (constant) and μ_1 (slope) across schools.

other parts of the Math score range, there is a clear discontinuity between 3.9 and 4.

Figure 3: Test for Local Manipulation



The simple test developed has clear limitations. The most important one is the assumption of a linear relationship between knowledge (measured by the SIMCE) and school GPA.[16] Indeed, this assumption is what determines the negative slope of the mean of the residuals. That said, it is remarkable that even imposing a linear relationship, the figure only show a clear jump between 3.9 and 4.

3.3 Tests Involving Covariates

To study the extent to which this manipulation could be a problem and how useful it is to use an RD approach in this context, Table 1 shows the differences in observables among different groups. In Group A, we compare students who were retained, in 2007, with students who were not. In this selected sample, the normalized differences in the means of the independent variables are all economically relevant, ranging from 1.49

[16] Another potential limitation is to assume that the SIMCE is an unbiased measure of the student knowledge. In our opinion what is relevant in this regard is that given the way in which the SIMCE is taken, where regular teachers are not in the classroom during the test (there are some external evaluators), there is no reason to think that SIMCE scores are manipulated.

to 0.29.[17] Moreover, all of these differences are in the same direction: the repeaters are students with characteristics highly correlated with future criminal behavior. They come from lower socioeconomic groups (measured by income and parents' education), they have lower levels of academic performance, their attendance rate is lower, and males are overrepresented in this group.

This story contrasts to that of Group B, where we compare students who, in 2007, scored 3.9 with those who scored 4, in their second-lowest subject. The stories from these two samples differ in two ways. First, the magnitudes of the normalized differences are remarkably smaller in Group B, where the largest normalized difference is 0.1. Second, in Group B, the signs of the differences in observables – between the highly probable repeaters and the rest– are in some cases in the opposite direction of those in Group A. For instance, students scoring 3.9 have a lower mean in repetition before 2007, and higher means in attendance in 2006 and in family income.

[17]The normalized difference in the mean is equal to $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(Sd(X_1)^2 + Sd(X_2)^2)/2}}$, where $\bar{X}_i$ is the sample mean for group i and $Sd(X_i)^2$ is the estimated variance for group i.

Table 1: Differences in covariates among different treatments and control groups

			Group A: All				
Variable	Non Repeaters	Repeaters	Norm. Dif.	Statistic	p-value	N (Non Rep.)	N (Rep.)
Repeated before 2007	0.09	0.29	-0.54	-65.25	0.000	683972	21293
Attendance 2006	94.6	92.3	0.38	49.71	0.000	683972	21293
Math SIMCE	0.01	-0.89	1.00	154.64	0.000	683972	21293
Language SIMCE	0.01	-0.88	0.99	151.55	0.000	683972	21293
Mother Education	10.78	9.50	0.37	52.61	0.000	683972	21293
Father Education	10.86	9.70	0.32	45.76	0.000	683972	21293
Family Income	98960	74412	0.28	43.81	0.000	683972	21293
Male	0.50	0.64	-0.30	-42.24	0.000	683972	21293
			Group B: second lowest subject score $\in \{3.9, 4.0\}$				
Variable	Mean (= 4.0)	Mean (= 3.9)	Norm. Dif.	Statistic	p-value	N (= 4.0)	N (= 3.9)
Repeated before 2007	0.25	0.20	0.10	2.65	0.008	2496	885
Attendance 2006	93.5	93.6	-0.02	-0.63	0.528	2496	885
Math SIMCE	-0.58	-0.65	0.08	2.09	0.037	2496	885
Language SIMCE	-0.58	-0.62	0.05	1.28	0.200	2496	885
Mother Education	10.43	10.29	0.04	1.07	0.283	2496	885
Father Education	10.54	10.39	0.04	1.08	0.282	2496	885
Family Income	91620	95597	-0.04	-1.08	0.280	2496	885
Male	0.56	0.60	-0.07	-1.68	0.093	2496	885
			Group C: second lowest subject score $\in \{3.8, 4.1\}$				
Variable	Mean (= 4.1)	Mean (= 3.8)	Norm. Dif.	Statistic	p-value	N (= 4.1)	N (= 3.8)
Repeated before 2007	0.21	0.23	-0.05	-2.71	0.007	7463	3889
Attendance 2006	93.2	93.1	0.01	0.61	0.540	7463	3889
Math SIMCE	-0.54	-0.73	0.23	11.89	0.000	7463	3889
Language SIMCE	-0.57	-0.73	0.19	9.89	0.000	7463	3889
Mother Education	10.36	10.17	0.06	2.82	0.005	7463	3889
Father Education	10.52	10.40	0.04	1.84	0.065	7463	3889
Family Income	88325	88132	0.00	0.11	0.913	7463	3889
Male	0.57	0.58	-0.00	-0.17	0.865	7463	3889

Note: Norm. Dif. is the normalized differences in the means.

The comparison between these two selected samples (Groups A and B) illustrates how much we gain by taking advantage of the discontinuity. Without an RD approach, the initial differences between the treated and the control groups – presented in Group A – would be too large to implement an empirical method based on controlling observables (*e.g.,* a type of matching), as a credible approach to estimate a causal effect. That said, as was anticipated in the density analysis and in our test for local manipulation, Group B shows some evidence of manipulation around the threshold, because theoretically,

without manipulation students scoring 3.9 should have – on average – worse performance and lower socioeconomic status than those students scoring 4, which is not always the case with our data. Of particular note, is the difference in the fraction of students who have previously repeated. A reasonable explanation for this difference is that teachers are more demanding with students who have not previously failed a grade, which creates a non random sorting around the threshold.

To address the sorting of students around the threshold, in Group C we compare students who scored 3.8 with those who scored 4.1 in their second-lowest subject. This selected sample has advantages and disadvantages, when compared to Group B. In regards to the former, all of the mean differences in observables between students below and above the threshold have the expected sign, which is consistent with the previous evidence that supports the idea that beyond 3.9 and 4 the data is free of manipulation. Regarding the disadvantages, we lose comparability between the groups below and above the threshold, particularly in respect to student performance. In sum, the remaining differences observed in Group C are much smaller than the ones observed in Group A and arguably free of manipulation. However, they are large enough to suggest the need to complement the RD design with another approach, to control for the differences in observables.

4 Empirical Approach

Considering the opportunities and problems with our data, we implement three different strategies to estimate the effect of grade retention on juvenile crime. In the first approach, which takes advantage of the local nature of the manipulation, we implement a standard FRD, but only using the students who scored 3.8 or 4.1 in their second-lowest score (Group C sample, Table 1). This RD method is known in the literature as the "Donut-hole" regression discontinuity; see Barreca, Guldi, Lindo, and Waddell (2011). In the second approach (*FRD-matching*), which addresses the differences in observables observed in the Group C sample, we combine a fuzzy regression discontinuity design (FRD) with the matching approach, named *design matching*, developed by Zubizarreta (2012).[18] This is our preferred empirical strategy. Finally, our third approach is to implement OLS estimator considering all of the students in our sample, and controlling

[18]This method is an extension of Keele, Titiunik, and Zubizarreta (2015), where the authors combine sharp regression discontinuity design with matching.

for all the variables used in the other empirical approaches. As opposed to the first two methods, this last approach is not implemented to deliver causal evidence, but is presented to give a reference point.

Given that in the first two methods we follow the FRD approach, which allows us to obtain the local treatment effect (LATE),[19] we begin this exposition by describing this empirical method. As detailed below, the difference between our two empirical strategies to estimate the causal effect ("Donut-hole" RD and FRD-matching) is in the procedure to define the sample used to implement the FRD estimation.

Let Y_i be a variable that takes the value one if the student committed a crime after 2007 and zero otherwise; Z_i a variable that takes the value one if the student's second-lowest subject score, in 2007, is below the threshold and zero otherwise; W_i a variable that takes the value one if the student repeats the grade, and zero otherwise; and X_i a set of covariates of student i. Hence, as is shown in Hahn, Todd, and der Klaauw (2001), when the sample considered is close to the threshold, the identification of the LATE parameter is given by a type of Wald estimator, such that:

$$\hat{\tau}_{FRD} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[W|Z=1] - E[W|Z=0]} \tag{1}$$

Furthermore, as pointed out by Imbens and Lemieux (2008), it is possible to obtain this Wald estimator by implementing a Two Stage Least Square method, where the first and second stages are described by:

$$First\ Stage: \quad W_i = \alpha_c^w + \alpha_z^w Z_i + \alpha_x^w X_i + \varepsilon_i^w, \tag{2}$$

$$Second\ Stage: \quad Y_i = \alpha_c^y + \tau_{FRD}\hat{W}_i + \alpha_x^y X_i + \varepsilon_i^y. \tag{3}$$

In this context, $\hat{\tau}_{FRD}$ is the estimation of the local average treatment effect.[20]

The first empirical approach, the "Donut-hole" RD, is the standard FRD but excluding the students whose second-lowest score is 3.9 or 4. Specifically, the sample consists of all the students who score 3.8 or 4.1 in their second lowest subject score, who belong to a

[19] See Imbens and Lemieux (2008) and Lee and Lemieux (2010) for reviews of RDD methods.

[20] This method is implemented in Stata using the command *ivreg*, with robust standard errors. It should be noted that this method of calculating robust standard errors does not take into account that the sample to implement the FRD estimation is built using a matching procedure.

schools-cohort with at least one student at each side of the threshold (Group C, Table 1).[21] Thus, given this restricted sample, the local average treatment effect is obtained by regressing equations (2) and (3).

The second empirical approach, the FRD-matching method, has as its starting point the same sample as the first approach (Group C sample). The difference lies in that in order to address the imbalance in observables between students scoring below and above the threshold, we use the *design matching* estimator to build similar groups. Unlike the standard matching methods, which attempt to achieve covariate balance by minimizing the total sum of distances between treated units and matched controls, this method achieves covariate balance directly by minimizing the total sum of distances while constraining the measures of imbalance to be less or equal to certain tolerances. In our implementation of this matching, we optimally find a pair for each student scoring 3.8, selected from those who are attending the same school-cohort and score 4.1,[22] by minimizing the weighted distance in math and language standardized test scores, parents education, previous retentions repetitions, attendance at past year, an income variable and gender; subject to mean balance on the same set of variables.[23]

We decided to implement this matching approach, as opposed to a more standard type, given that we have a relevant number of school-cohort clusters for which there are few students scoring 4.1 that qualify as a match for those scoring 3.8.[24] This situation creates an imbalance in observables that can only be reverted by a method that takes advantage of the school-cohort clusters with more options, by not only seeking more similar pairs, but also looking for pairs that compensate this imbalance. For instance, if at the school-cohort clusters with only one student scoring 4.1 (*i.e.,* with no option), there is also a higher fraction of males below the threshold, the design matching method will prefer –in some cases– to match males scoring 4.1 with females scoring 3.8, in the other school-cohort clusters, to compensate the former imbalance.

Table 2 presents the balance achieved by this matching procedure on the mentioned covariates. Comparing the differences observed in Table 2 with the differences presented

[21] That is, a cohort within a school with at least one student scoring 3.8 and one student scoring 4.1 in their subject with the second lowest score.

[22] In one specification, we also implement an exact match in gender.

[23] The details of this matching approach are described in Appendix B.

[24] In 12% of the cases there is only one, and in 29%, one or two.

in Group C in Table 1, it is clear that there is an improvement in terms of balance in observables.[25] However, there is an important reduction in the sample size (from 3889 to 2931 individuals below the threshold).

Table 2: Post matching differences in covariates

Variable	4.1	3.8	Norm. Dif.	Statistic	p-value	N (= 4.1)	N (= 3.8)
Repeated before 2007	0.20	0.20	-0.01	-0.55	0.582	2959	2959
Attendance 2006	93.3	93.2	0.02	0.70	0.481	2959	2959
Math SIMCE	-0.65	-0.68	0.03	1.15	0.248	2959	2959
Language SIMCE	-0.65	-0.68	0.03	1.28	0.202	2959	2959
Mother Education	10.33	10.26	0.02	0.79	0.430	2959	2959
Father Education	10.51	10.48	0.01	0.34	0.731	2959	2959
Family Income	89012	88476	0.01	0.23	0.821	2959	2959
Male	0.57	0.58	-0.01	-0.39	0.693	2959	2959

Note: Norm. Dif. is the normalized differences in the means.

Let N_{bt} be the number of students who score below the threshold and who have a match – above the threshold – found by the design matching procedure. Then, we estimate the local average treatment effect by implementing the 2SLS estimator described by equations (2) and (3), with a sample of $2 * N_{bt}$ students, where, for each of the N_{bt} students scoring below the threshold, we have one similar student scoring above the threshold.

5 Results

In this section, we present our findings on the impact of grade retention on juvenile crime, student drop out, and future grade retention. Moreover, we show the results of a placebo test.

[25] We also tried to achieve this balance by implementing a more standard matching approach (*e.g.,* minimizing the mahalanobis distance). However, in that case the improvement was only partial, probably due to the important number of school-cohort clusters for which there are few students scoring 4.1 that qualify as a match for those scoring 3.8.

5.1 Impact of Grade Retention on Crime

The main results of this paper are presented in Table 3, which shows the effect of grade retention on juvenile crime for different populations and under different empirical approaches. Focusing on the first two columns, which summarize the results of the empirical strategies intended to deliver causal effects, we find that the effect of grade retention on crime ranges from 1.6 to 3.7 pp, and in almost all specifications the effect is statistically significant.[26]

Table 3: Effect of grade retention on juvenile crime

Sample	(1) Donut-Hole FRD	(2) FRD-Matching	(3) OLS
All	0.016	0.018	0.035
	(0.0068)	(0.0082)	(0.0019)
	$N = 9681$	$N = 5130$	$N = 705261$
Low SES	0.024	0.037	0.044
	(0.0120)	(0.0142)	(0.0027)
	$N = 4527$	$N = 2330$	$N = 359021$
Males	0.024	0.025	0.042
	(0.0114)	(0.0147)	(0.0026)
	$N = 4187$	$N = 2176$	$N = 353552$
First Repetition	0.011	0.015	0.034
	(0.0079)	(0.0097)	(0.0021)
	$N = 6630$	$N = 3338$	$N = 638582$

Note: In the case of FRD-Matching there are $N_{bt}/2$ students with their second lowest score equal to 3.8 and $N_{bt}/2$ students with that score equal to 4.1. Standard errors in parentheses.

The effect is heterogeneous and economically significant. In particular, the impact – measured as percentage points – is larger for males and students from low SES.[27] Regarding the magnitudes, given that the crime rate for the students in this sample is about 4.8% (see Table 1, Group C), the estimates range from an effect of 33% to 77%.[28] Finally, the

[26]The difference between the sample size of columns one and two is due to the fact that in the former, most of the time, there are more than one student scoring 4.1 for each student scoring 3.8, which is not the case in column 2 (by construction).

[27]Low SES is defined as the group of students attending schools which fall below the median for a school's average income.

[28]A precautionary note about this range: these population groups also have different crime rates. For example, the male rate is 6.7% (the female rate is 2.2%) and the crime rate for students attending low SES schools is 6.8%.

effect is not statistically significant for those students this is their first time repeating a grade (row four). However, this last result should not be emphasized, because the point estimations are rather similar and the p-value are close to 0.1.

5.2 Effects on Other Outcomes

Given the informative nature of our panel data set, we can also examine the effect of grade retention on other outcomes.[29] Specifically, we could also focus on dropping out and future grade retention (after 2007). To do so, we implement the same empirical strategies that we followed to estimate the effect on crime. Table 4 shows the effect of grade retention in 2007 on the probability of future repetitions.[30] In particular, grade retention in 2007 decreased the probability of future repetitions from 2.3 to 10.4 pp (column (1)). Given that, in the estimation sample, 55% of the students repeat at least one grade after 2007, these figures represent a decrease from 4.1 to 18.9%.[31]

[29]We focus on the *Donut-Hole* RD method, as opposed to FRD-matching, given that this approach presents the smaller point estimates in the placebo analysis, and it also delivers the smaller effects in all the estimations.

[30]Given that there are drop-outs, there is a potential selection bias problem that we do not address in this paper.

[31]In the estimation sample, 55% of the students repeat at least one grade after 2007, which reflects two features of the data. First, the grade retention rate is remarkably high in Chile; in fact, the percentage for the entire population is 39%. Second, low performing students are overrepresented in the estimation sample.

Table 4: Effect of grade retention on future grade retention

Sample	(1) Donut-Hole FRD	(2) FRD-Matching	(3) OLS
All	-0.082	-0.059	0.116
	(0.0155)	(0.0191)	(0.0035)
	$N = 9681$	$N = 5130$	$N = 705261$
Low SES	-0.050	-0.023	0.107
	(0.0221)	(0.0272)	(0.0047)
	$N = 4527$	$N = 2330$	$N = 359021$
Males	-0.104	-0.096	0.111
	(0.0215)	(0.0278)	(0.0044)
	$N = 4187$	$N = 2176$	$N = 353552$
First Repetition	-0.068	-0.048	0.152
	(0.0189)	(0.0238)	(0.0041)
	$N = 6630$	$N = 3338$	$N = 638582$

Note: In the case of FRD-Matching there are $N_{bt}/2$ students with their second lowest score equal to 3.8 and $N_{bt}/2$ students with that score equal to 4.1. Standard errors in parentheses.

We define drop out as a situation in which the student does not attend school in the years corresponding to 11th and 12th grade. For instance, we say that a student who was attending 4th grade in 2007 dropped out, if she did not attend school in 2014 and 2015. We apply this definition, in order to have a comparable measure of drop out, among a group of students who were in different grades in 2007 (4th to 8th grade). Table 5 shows the effects of grade retention on dropping out. Specifically, grade retention in 2007 increases the probability of dropping out, from 1.2 to 3.2 pp (column (1)).[32] According to the drop out measure used in this paper, 6.3% of the students dropped out after 2007. Thus, these figures represent an increase from 19 to 51%. We find no effects for those students who repeated for the first time in 2007.

[32]These results are along the same lines as the findings of Manacorda (2012) and Jacob and Lefgren (2009).

Table 5: Effect of grade retention on Dropping out

Sample	(1) Donut-Hole FRD	(2) FRD-Matching	(3) OLS
All	0.015	0.016	0.059
	(0.0079)	(0.0096)	(0.0023)
	$N = 9681$	$N = 5130$	$N = 705261$
Low SES	0.030	0.032	0.081
	(0.0142)	(0.0174)	(0.0035)
	$N = 4527$	$N = 2330$	$N = 359021$
Males	0.010	0.019	0.057
	(0.0118)	(0.0141)	(0.0029)
	$N = 4187$	$N = 2176$	$N = 353552$
First Repetition	0.006	0.003	0.039
	(0.0076)	(0.0093)	(0.0022)
	$N = 6630$	$N = 3338$	$N = 638582$

Note: In the case of FRD-Matching there are $N_{bt}/2$ students with their second lowest score equal to 3.8 and $N_{bt}/2$ students with that score equal to 4.1. Standard errors in parentheses.

In addition to the discussion on magnitudes, there are several aspects of these results that are important to highlight. First, as in the crime estimation, the OLS estimation delivers larger effects, probably due to unobservable variables, because in that case we do not take advantage of the discontinuity in the grade retention probability (Column (3)), which adds further evidence supporting the soundness of the empirical method developed in this paper (*i.e.,* the FRD-matching, extension of Zubizarreta (2012)). Second, the effect of grade retention on dropping out suggests a relevant mechanism through which grade retention may affect juvenile crime: grade retention impacts on dropping out and dropping out impacts on crime. Third, if we assume that grade retention in higher grades also impacts on juvenile crime, then the results of Table 4 suggest that we are finding a lower bound for the effect of grade retention on crime, because those who did not repeat in 2007 (who are *non-treated* in our estimation) had a higher probability of grade retention in the future, which also impacts on crime.

5.3 Robustness Analysis

To examine the robustness of our results, we perform two empirical exercises. In the first one, we re-estimate the "Donut-hole" RD and the RD-matching, but now we restrict the

sample to the students whose final status at school is consistent with the retention rule. In practice, this is equivalent to re-estimating Columns (1) and (2) of Table 3, but now imposing a sharp RD design.

To be clear, we re-estimate the "Donut-hole" RD specification in two steps: (1) among all students whose second lowest score is 3.8 or 4.1, we only keep the students whose final status at school is consistent with the retention rule: below the threshold we drop the students who pass the grade, and above the threshold, we drop the students who repeat the grade; (2) given this sample, we estimate a standard sharp design RD, by regressing the following equation:[33]

$$Y_i = \alpha_c^y + \tau W_i + \alpha_x^y X_i + \varepsilon_i^y. \tag{4}$$

Along the same line, we re-estimate the RD-matching in two steps: (1) among all students whose second lowest score is 3.8 or 4.1, as before, we only keep the students whose final status at school is consistent with the retention rule; (2) given the matched sample, the LATE parameter (τ) is estimated by regressing equation 4.[34]

The second empirical exercise to review the robustness of our results is to implement a placebo test. In this case, we replicate the "Donut-hole" RD and the RD-matching estimations, but now we only compare students scoring below and above the threshold, who did not repeat the grade.[35] For instance, in the case of the "Donut-hole" RD, we proceed with the following two steps: (1) among all students whose second lowest score is 3.8 or 4.1, we only keep the students whose final status at school is *pass the grade*; (2) given this sample, we estimate $E[Y_i|Z_i = 0, W_i = 0, X_i]$ and $E[Y_i|Z_i = 1, W_i = 0, X_i]$ by regressing equation 4. To support our empirical approach, we should get that $E[Y_i|Z_i = 0, W_i = 0, X_i] = E[Y_i|Z_i = 1, W_i = 0, X_i]$.[36]

[33] Given step (1), this sample does not require a 2SLS estimator. Indeed, it is a sharp design RD.

[34] We are using the matched sample described in Table 2, as opposed to finding a new matched sample given the smaller number of students scoring below the threshold. These samples would be different due to the fact that design matching involves constraining the measures of imbalance to be less or equal to certain tolerances.

[35] In principle, we could do the same by comparing those who are below and above the threshold and repeated the grade. However, we do not have a sufficiently large sample size to do that.

[36] See Imbens and Rubin (2015).

Table 6: Effect of grade retention on juvenile crime (sharp design and placebo)

	Sharp Design		Placebo	
	(1)	(2)	(3)	(4)
	Donut-Hole RD	RD-Matching	Donut-Hole RD	RD-Matching
Sample				
All	0.015	0.018	0.002	-0.001
	(0.0058)	(0.0067)	(0.0082)	(0.0081)
	$N = 8694$	$N = 4421$	$N = 7054$	$N = 3236$
Low SES	0.018	0.033	0.015	0.014
	(0.0102)	(0.0118)	(0.0165)	(0.0166)
	$N = 4096$	$N = 2040$	$N = 3259$	$N = 1435$
Males	0.021	0.026	0.022	0.005
	(0.0101)	(0.0123)	(0.0158)	(0.0168)
	$N = 3783$	$N = 1910$	$N = 2891$	$N = 1340$
First repetition	0.013	0.016	-0.003	-0.004
	(0.0067)	(0.0080)	(0.0097)	(0.0092)
	$N = 5934$	$N = 2878$	$N = 4782$	$N = 2103$

Note: In the case of FRD-Matching there are $N_{bt}/2$ students with their second lowest score equal to 3.8 and $N_{bt}/2$ students with that score equal to 4.1. Standard errors in parentheses.

The results of these empirical exercises are presented in Table 6. In short, the figures of the first two columns, coming from the re-estimation of the "Donut-hole" RD and the RD-matching (but now imposing a sharp design), are remarkably similar to the results presented in Table 3. More importantly, the results of the placebo exercises (Columns (3) and (4)) show no statistical significance. Regarding the magnitudes, although all the estimates are not statistically significant, column (3) shows better (closer to zero) point estimates compared to column (4), namely, the "Donut-hole" RD seems more robust than the RD-matching. Overall, placebo results are rather important given that they reinforce the claim that the numbers presented in Columns (1) and (2) of Table 3 can be interpreted as (local) causal effects.[37]

Finally, in appendix C.2, we present the robustness analysis for drop out and grade retention after 2007 (Tables 11 and 12). As in the case of crime, in the placebo test

[37]That said, it is important to consider that the robustness of our approaches critically depend on the level of the initial imbalance in observables. For instance, we ran the same placebo approaches, but compared students who scored 4.1 to students who scored 4.4, and we found differences that were statistically significant. However, the differences in observables between these two groups (scoring 4.1 and 4.4) were much higher than the differences between the groups that were used in our estimation.

all parameters are statistically insignificant in the case of future grade retention and drop out. These results confirm the soundness of our empirical strategy to find causal estimates.

6 Mechanisms

In this section, we present two exercises to shed some light on what may explain the impact of grade retention on juvenile crime. First, we explore how grade retention increases the probability of the occurrence of negative trajectories after 2007. Second, we discuss the relevance of school switching in reinforcing the negative effect of grade retention on student's future.

6.1 The interaction between drop out and crime

The causal effect of grade retention on juvenile crime, documented in the previous section, could have operated through different mechanisms. For instance, grade retention could have impacted only through the effect of repetition on drop out, and the subsequent effect of drop out on crime. To explore what happens after grade repetition and how this event affects students' trajectories, we run a multinomial logit with four trajectories as possible outcomes, whose results are presented in Table 7. The possible trajectories after 2007 are: attending school in all periods of our sample, without committing a crime over those years (column (1)); dropping out in a year t (after 2007), without committing a crime in a year $t + a$, with $a > 0$ (column (2)); dropping out in a year t (after 2007) and committing a crime in a year $t + a$ (column (3)); and committing a crime in a year t (after 2007) and dropping out after that, simultaneously, or never (column (4)).

Given the non-linearity of this model we avoid the use of the score in the second lowest subject as the instrument in a fuzzy RD design, and instead we run a multinomial logit as if we had a sharp design RD environment. We do so by following the same approach described in section 5.3, namely, among all students whose second lowest score is 3.8 or 4.1, we only keep the students whose final status at school is consistent with the retention rule. Given this sample, we implement the design matching to optimally find a pair for each student scoring 3.8 among those who score 4.1. Therefore, and given this matched sample, the variable of interest is a dummy that takes one, if the student repeated the

grade in 2007, scoring 3.8 in the second lowest score, and it takes zero, if the student did not repeat the grade in 2007, scoring 4.1 in the second lowest score.[38] Besides this variable, the model includes the same controls as the models of the previous section.[39]

Table 7: Effect of grade retention on the probability of different teens' trayectories

	Always at School, no crime	Dropout, no crime	First Dropout, then crime	First crime, then, or simultaneously, dropout (if so)
Grade Retention (in 2007)	-0.0435	0.0282	0.0053	0.0100
	(0.0118)	(0.0108)	(0.0020)	(0.0048)
			Number of Obs. = 4961	Pseudo R2 = 0.11

Note: This is a multinomial logit model with a dependent variable with four categories. The model includes the following controls: gender, father education, mother education, math and language simce, family income, attendance in 2006, previous grade retentions. The table presents the marginal effects.

Table 7 shows the marginal effects of this multinomial logit. It says, for example, that grade retention increases the probability of committing a crime after 2007 before dropping out (if so) by 1 pp. Overall, the results show that grade retention increases the probabilities of "bad trajectories" (involving either dropping out or crime) and that the effect of grade retention on crime is not only through its effects on dropping out. In fact, the effect on crime occurring before (or simultaneously to) dropping out is more relevant than the effect on crime that occurs after dropping out.

6.2 Switching schools after grade retention

As documented in Hanushek, Kain, and Rivkin (2004), as a result of switching schools, students may experience a substantial pedagogical cost. If so, it could be the case that part of the effect of grade retention on crime found in this paper, is due to the fact that repetition may increase the probability of switching schools.

[38]It should be noted that even though this approach does not allow for a discussion on causality, since the right approach would be a FRD, the analysis of the effects of grade retention on crime, drop out, and grade retention after 2007 (presented in the previous sections) shows that this *fake* sharp design RD delivers rather similar results to the fuzzy Rd estimators.

[39]These are: gender, father's education, mother's education, math and language SIMCE, family income, attendance in 2006, previous grade retentions.

To explore the relevance of this mechanism, we run an OLS regression among all the students who repeated in 2007, scoring between 3.0 and 4.5 on the second lowest score,[40] and we study the correlation between switching schools (between 2007 and 2008) and juvenile crime, controlling for the same variables as in the models of the previous section, and including school-grade fixed effects. Considering that 8th grade a higher rate of student turnover, we present the results of this model including and excluding this grade.

Table 8: Crime and switching schools

Variables	4th to 8th grade		Excluding 8th grade	
Switching school	0.0220	(0.0068)	0.0224	(0.0080)
Attendance 2006	-0.0012	(0.0005)	-0.0014	(0.0006)
Mother Education	-0.0018	(0.0012)	-0.0022	(0.0013)
Father Education	-0.0019	(0.0011)	-0.0023	(0.0012)
Family Income	-0.0000	(0.0000)	-0.0000	(0.0000)
Male	0.0546	(0.0061)	0.0591	(0.0070)
Math SIMCE	0.0008	(0.0047)	0.0009	(0.0057)
Language SIMCE	-0.0011	(0.0044)	-0.0031	(0.0053)
Constant	0.1794	(0.0476)	0.2066	(0.0554)
N		18946		15171
R2		0.086		0.093

Note: These two estimations include school-grade fixed effects.
Standard errors in parentheses.

Table 8 shows that for those students who repeated a grade, switching schools increases the probability of crime by 2.2 pp, a result that does not change when 8th grade is excluded. Thus, in concordance with the literature, grade retention could be particularly negative for a student's future when directly followed by a change in school.

7 Conclusion

We exploit a discontinuity in the grade retention probability, given a repetition rule, to examine the effect of grade retention in primary school on juvenile crime. Due to clear evidence about – local – manipulation on the forcing variable, we depart from standard RD methods. First, we follow Barreca, Guldi, Lindo, and Waddell (2011)

[40] We use this set of students to have both a large enough sample size, and a group which is similar to the sample used in section 5.1.

to implement a *donut-hole* FRD, where, after removing observations in the immediate vicinity of the threshold for grade repetition, we run a standard FRD. Second, we extend the method developed by Keele, Titiunik, and Zubizarreta (2015) to implement a method that combines matching with a fuzzy regression discontinuity design.

This paper makes three main contributions. First, together with a recent paper (Depew and Eren (2015)), it is the first that estimates a causal effect of grade retention on juvenile crime and it is the first evidence for a developing country. This causal evidence calls into question the appropriateness of grade repetition as a public policy, a concern that is even more relevant in the context of Chile, a developing country with a high rate of grade retention.[41] That said, the interpretation of our findings should consider that we are not taking into account other aspects of this policy, *e.g.,* for example, the threat of retention could serve as an incentive for all students to exert more effort. Second, by extending the method developed by Keele, Titiunik, and Zubizarreta (2015) to the *fuzzy RD* case, we present an empirical approach that can be useful in many other contexts in which there is some evidence of manipulation in the forcing variable. Third, the paper sheds light on the mechanisms that could explain the impact of grade retention on juvenile crime. From this analysis, it is possible to infer relevant insights for public policy debate.

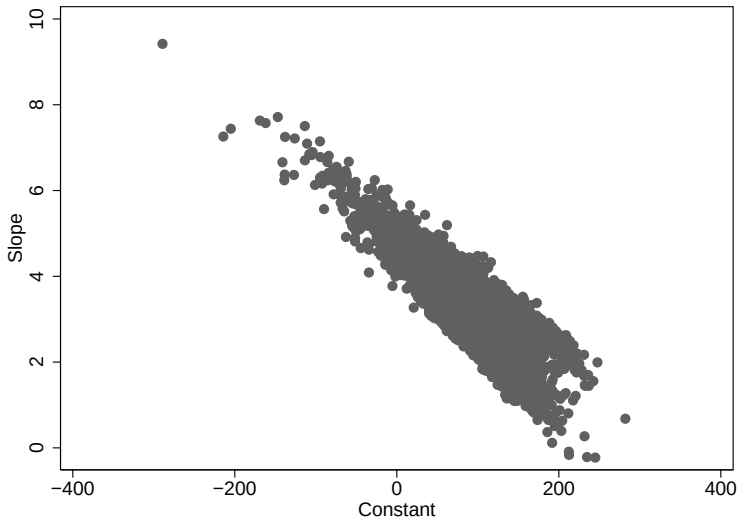
On one hand, because the effect of grade retention on crime does not only operate through its effect on dropping out, it is not possible to argue that rather than concern ourselves with grade retention, we should only start to worry when the student who has been held back, drops out. The evidence presented in this paper implies that policy makers should be concerned about high levels of grade retention on its own merit.

On the other hand, the relevance of the interaction between grade retention and school switching in the determination of juvenile crime, gives important clues about a possible avenue to attenuate the negative effects of grade retention, in case policy makers decide to continue supporting this practice. In particular, since repetition is a negative response from the education system, it may discourage students' commitment to their educational process. Thus, it is essential to design policies that will counteract this negative effect, breaking the connection between grade repetition and other causes of dropping out and juvenile crime. Particularly, for those students who have repeated more than once.

[41] In fact, in our sample 13.1% of the students repeated at least one grade between 1st and 8th grade.

A Grading Standards

Figure 4: Heterogenous Grading Standards Across Schools



B Design Matching

Let Z_1 denote the group of students whose second-lowest score in 2007 is below the threshold (*i.e.,* equal to 3.8), and let Z_0 denote the group of students whose second-lowest score is above the threshold (equal to 4.1).[42] Let j_1 index the members of group Z_1 and j_0 index the members of group Z_0. Define d_{j_1,j_0} as the covariate distances (in math and language standardized test scores, parents' education, previous repetitions, attendance during the previous year, per capita income, and gender) between unit j_1 and j_0. To enforce specific forms of covariate balance, define $e \in \varepsilon$ as the index of the covariate (school and grade identification) for which it is needed to match exactly, and $b_e \in B_e$ as the categories that covariate e takes, so that $x_{j_1;e}$ is the value of nominal covariate e for unit j_1 with $x_{j_1;e} \in B_e$. Finally, let $m \in M$ be the index of covariates for which it is desired to balance their means, in this case: math and language standardized test scores, parents' education, previous retentions, attendance during the previous year, per capita income, and gender. So that $x_{j_1;m}$ is the value of covariate m for unit j_1, and $x_{j_0;m}$ is the value of covariate m for j_0.

[42] We follow the notation and the description from Keele, Titiunik, and Zubizarreta (2015)

To solve the problem optimally, the following decision variables are introduced:

$$a_{j_1;j_0} = \begin{cases} 1 & \text{if unit } j_1 \text{ is matched to unit } j_0 \\ \\ 0 & \text{otherwise,} \end{cases}$$

Then, for a given scalar λ, the objective function to minimize is equal to:[43]

$$\sum_{j_1 \in Z_1} \sum_{j_0 \in Z_0} d_{j_1,j_0} a_{j_1,j_0} - \lambda \sum_{j_1 \in Z_1} \sum_{j_0 \in Z_0} a_{j_1,j_0}, \tag{5}$$

subject to pair matching and covariate balancing constraints. Under this penalized match, if distance can be minimized it will be, and if it cannot be minimized in every case, it will be minimized as often as possible. In particular, the pair matching constraints require each treated and control subject to be matched at most once,

$$\textstyle\sum_{j_0 \in Z_0} a_{j_1,j_0} \leq 1, \quad \forall j_1 \in Z_1 \tag{6}$$

$$\textstyle\sum_{j_1 \in Z_1} a_{j_1,j_0} \leq 1, \quad \forall j_0 \in Z_0 \tag{7}$$

This implies that it matches without replacement. The covariate balancing constraints are defined as follows

$$\textstyle\sum_{j_1 \in Z_1} \sum_{j_0 \in Z_0} \left| 1_{x_{j_1;e} = b_e} x_{j_1;e} - 1_{x_{j_0;e} = b_e} x_{j_0;e} \right| a_{j_1,j_0} = 0, \quad \forall e \in \varepsilon, \tag{8}$$

$$\left| \textstyle\sum_{j_1 \in Z_1} \sum_{j_0 \in Z_0} a_{j_1;j_0} x_{j_1;m} - \sum_{j_1 \in Z_1} \sum_{j_0 \in Z_0} a_{j_1;j_0} x_{j_0;m} \right| \leq \varepsilon_m \textstyle\sum_{j_1 \in Z_1} \sum_{j_0 \in Z_0} a_{j_1;j_0}, \quad \forall m \in M, \tag{9}$$

where 1 is the indicator function.

These constraints enforce exact matching and mean balance, respectively. More precisely, (8) requires exact matching by matching each subject in Z_1 to a subject in Z_0 in the same school and grade; and (9) forces the differences in means after matching to be less than

[43] We solve this optimization problem, by implementing the R package described in Zubizarreta and Kilcioglu (2016).

or equal to $\varepsilon_m = 0.03$ standard deviations apart for all $m \in M$, with M = standardized scores in language and math, parents' education, previous retentions, attendance during the previous year, an income variable and gender.

The Designmatch incorporates optimal subset matching into the integer programming framework in the objective function (5) via the λ parameter. The first term in (5) is the total sum of mahalanobis distances between matched pairs, and the second term is the total number of matched pairs. Therefore, λ emphasizes the total number of matched pairs in relation to the total sum of distances and, according to (5), it is preferable to match additional pairs if on average they are at shorter distances than λ. In our application, we choose λ to be equal to the median mahalanobis distance between j_1 and j_0 subjects so, according to (5), it is preferable to match additional pairs if on average they are at a shorter distance than the typical distance (as measured by the median).[44] Subject to the pair matching constraints (6) and (7) and the covariate balancing constraints (8) and (9), this form of penalized optimization addresses the lack of common support problem in the distribution of observed covariates of subject in Z_1 and Z_0.

Due to this penalty, the Design match keeps the largest number of matched pairs for which distance is minimized and the balance constraints are satisfied. This implies that as we alter the distances or the balance constraints, the number of j_1 and j_0 subjects retained changes. In particular, for stricter constraints we tend to retain a smaller number of subjects.

[44] λ can be thought of as a parametrization of the trade-off between bias and variance: a higher value of it would imply a bigger sample size, but more differences between treated and controls.

C Robustness Analysis

C.1 Results excluding students who repeated before 2007

Table 9: Effect of grade retention on juvenile crime (excluding who repeteated before 2007)

Sample	(1) Donut-Hole FRD	(2) FRD-Matching	(3) OLS
All	0.011	0.015	0.034
	(0.0079)	(0.0097)	(0.0021)
	$N = 6630$	$N = 3338$	$N = 638582$
Low SES	0.026	0.045	0.045
	(0.0153)	(0.0185)	(0.0032)
	$N = 2718$	$N = 1326$	$N = 312667$
Males	0.005	0.001	0.041
	(0.0133)	(0.0193)	(0.0030)
	$N = 2815$	$N = 1412$	$N = 313398$

Note: In the case of FRD-Matching there are $N_{bt}/2$ students with their second lowest score equal to 3.8 and $N_{bt}/2$ students with that score equal to 4.1. Standard errors in parentheses.

Table 10: Effect of grade retention on juvenile crime: sharp design and placebo, excluding who repeteated before 2007.

	Sharp Design		Placebo	
Sample	(1) Donut-Hole RD	(2) RD-Matching	(3) Donut-Hole RD	(4) RD-Matching
All	0.013	0.016	-0.003	-0.004
	(0.0067)	(0.0080)	(0.0097)	(0.0092)
	$N = 5934$	$N = 2878$	$N = 4782$	$N = 2103$
Low SES	0.020	0.041	0.013	0.010
	(0.0132)	(0.0157)	(0.0215)	(0.0203)
	$N = 2448$	$N = 1161$	$N = 1926$	$N = 814$
Males	0.008	0.011	0.001	-0.022
	(0.0116)	(0.0159)	(0.0191)	(0.0194)
	$N = 2544$	$N = 1243$	$N = 1931$	$N = 859$

Note: In the case of FRD-Matching there are $N_{bt}/2$ students with their second lowest score equal to 3.8 and $N_{bt}/2$ students with that score equal to 4.1. Standard errors in parentheses.

C.2 Other outcomes

Table 11: Effect of grade retention on future grade retention (sharp design and placebo)

	Sharp Design		Placebo	
	(1)	(2)	(3)	(4)
	Donut-Hole RD	RD-Matching	Donut-Hole RD	RD-Matching
Sample				
All	-0.087	-0.060	0.006	-0.005
	(0.0129)	(0.0151)	(0.0205)	(0.0212)
	$N = 8694$	$N = 4421$	$N = 7054$	$N = 3236$
Low SES	-0.058	-0.029	0.022	0.010
	(0.0190)	(0.0221)	(0.0321)	(0.0326)
	$N = 4096$	$N = 2040$	$N = 3259$	$N = 1435$
Males	-0.120	-0.101	0.048	0.007
	(0.0191)	(0.0227)	(0.0346)	(0.0334)
	$N = 3783$	$N = 1910$	$N = 2891$	$N = 1340$
First repetition	-0.078	-0.045	0.018	-0.020
	(0.0155)	(0.0187)	(0.0250)	(0.0267)
	$N = 5934$	$N = 2878$	$N = 4782$	$N = 2103$

Note: In the case of FRD-Matching there are $N_{bt}/2$ students with their second lowest score equal to 3.8 and $N_{bt}/2$ students with that score equal to 4.1. Standard errors in parentheses.

Table 12: Effect of grade retention on dropping out (sharp design and placebo)

	Sharp Design		Placebo	
	(1)	(2)	(3)	(4)
	Donut-Hole RD	RD-Matching	Donut-Hole RD	RD-Matching
Sample				
All	0.017	0.020	-0.009	-0.013
	(0.0066)	(0.0079)	(0.0095)	(0.0092)
	$N = 8694$	$N = 4421$	$N = 7054$	$N = 3236$
Low SES	0.033	0.040	-0.016	-0.025
	(0.0121)	(0.0146)	(0.0185)	(0.0181)
	$N = 4096$	$N = 2040$	$N = 3259$	$N = 1435$
Males	0.010	0.020	0.001	-0.006
	(0.0100)	(0.0117)	(0.0169)	(0.0161)
	$N = 3783$	$N = 1910$	$N = 2891$	$N = 1340$
First repetition	0.006	0.008	0.001	-0.014
	(0.0065)	(0.0077)	(0.0085)	(0.0081)
	$N = 5934$	$N = 2878$	$N = 4782$	$N = 2103$

Note: In the case of FRD-Matching there are $N_{bt}/2$ students with their second lowest score equal to 3.8 and $N_{bt}/2$ students with that score equal to 4.1. Standard errors in parentheses.

References

Anderson, D. M. (2014): "In school and out of trouble? The minimum dropout age and juvenile crime," *Review of Economics and Statistics*, 96(2), 318–331.

Barreca, A. I., M. Guldi, J. M. Lindo, and G. R. Waddell (2011): "Saving Babies? Revisiting the effect of very low birth weight classification," *The Quarterly Journal of Economics*, 126(4), 2117–2123.

Barrington, B. L., and B. Hendricks (1989): "Differentiating characteristics of high school graduates, dropouts, and nongraduates," *The Journal of Educational Research*, 82(6), 309–319.

Cook, P. J., and S. Kang (2013): "Birthdays, Schooling, and Crime: New Evidence on the Dropout-Crime Nexus," Working Paper 18791, National Bureau of Economic Research.

Depew, B., and O. Eren (2015): "Test-Based Promotion Policies, Dropping Out, and Juvenile Crime," Departmental working papers, Department of Economics, Louisiana State University.

Fagan, J., and E. Pabon (1990): "Contributions of delinquency and substance use to school dropout among inner-city youths," *Youth and Society*, 21(3), 306.

Hahn, J., P. Todd, and W. V. der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

Hanushek, E. A., J. F. Kain, and S. G. Rivkin (2004): "Disruption versus Tiebout improvement: the costs and benefits of switching schools," *Journal of Public Economics*, 88(9-10), 1721–1746.

Holmes, C. T., et al. (1989): "Grade level retention effects: A meta-analysis of research studies," *Flunking grades: Research and policies on retention*, 16, 33.

Imbens, G. W., and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142(2), 615–635.

Imbens, G. W., and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge Books. Cambridge University Press.

Jacob, B. A., and L. Lefgren (2009): "The Effect of Grade Retention on High School Completion," *American Economic Journal: Applied Economics*, 1(3), 33–58.

Jimerson, S. R. (2001): "Meta-analysis of grade retention research: Implications for practice in the 21st century," *School psychology review*, 30(3), 420.

Keele, L., R. Titiunik, and J. R. Zubizarreta (2015): "Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout," *Journal of the Royal Statistical Society Series A*, 178(1), 223–239.

King, E. M., P. F. Orazem, and E. M. Paterno (2015): "Promotion with and without learning: Effects on student enrollment and dropout behavior," *The World Bank Economic Review*, p. lhv049.

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(2), 281–355.

Lochner, L. (2004): "Education, Work, And Crime: A Human Capital Approach," *International Economic Review*, 45(3), 811–843.

Manacorda, M. (2012): "The Cost of Grade Retention," *The Review of Economics and Statistics*, 94(2), 596–606.

Reschly, A. L., and S. L. Christenson (2013): "Grade retention: Historical perspectives and new research," *Journal of school psychology*, 51(3), 319–322.

Roderick, M. (1994): "Grade retention and school dropout: Investigating the association," *American Educational Research Journal*, 31(4), 729–759.

Rose, J. S., F. J. Medway, V. Cantrell, and S. H. Marus (1983): "A fresh look at the retention-promotion controversy," *Journal of school psychology*, 21(3), 201–211.

Thornberry, T. P., M. Moore, and R. Christenson (1985): "THE EFFECT OF DROPPING OUT OF HIGH SCHOOL ON SUBSEQUENT CRIMINAL BEHAVIOR*," *Criminology*, 23(1), 3–18.

Zubizarreta, J. (2012): "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery," *Journal of the American Statistical Association*, 107, 1360–1371.

Zubizarreta, J., and C. Kilcioglu (2016): "designmatch: Construction of Optimally Matched Samples for Randomized Experiments and Observational Studies that are Balanced by Design," Discussion paper, R package version 0.1.1.