The Impact of Grade Retention on Juvenile Crime

Juan Diaz* Nicolas Grau† Tatiana Reyes‡ Jorge Rivera§

This version: April, 2016

Abstract

Using detailed administrative data on schooling and crime records from Chile, we estimate the effect of grade retention between 4th and 8th grade on juvenile crime. We take advantage of a grade retention rule which specifies that students who fail more than one subject must repeat the year. Given that we find some evidence that the forcing variable is – locally – manipulated, we present two empirical strategies to address this potential problem. First, we follow Barreca, Guldi, Lindo, and Waddell (2011) to implement a *donut-hole* FRD. Second, we extend the approach developed by Keele, Titiunik, and Zubizarreta (2015) to implement a method that combines matching with fuzzy regression discontinuity design. These two methodologies deliver consistent results and both show no statistically significant effect in a placebo test. According to our results, grade retention increases the probability of juvenile crime by 1.2 percentage point (pp), an increase of 25%. The effect is higher for males and for low SES students, 1.7 and 2.1 pp, respectively. We also find that grade retention decreases the probability of grade retention in subsequent years, by 12.5 pp (18%), and increases the probability of dropping out by 0.9pp (51%).

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

We thank Guido Imbens, Jose Zubizarreta, and seminar participants at the Department of Economics of the Universidad de Chile 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.

*juadiaz@gmail.com. Department of Statistics, Harvard University
†ngrau@fen.uchile.cl. Department of Economics, University of Chile.
‡treyes@fen.uchile.cl. Department of Economics, University of Chile.
§jrivera@decon.uchile.cl. Department of Economics, University of Chile

1 Introduction

Developing countries show high rates of grade retention in primary and secondary school. As stressed by Manacorda (2012), while Central Asia, eastern and western Europe, and North America present repetition rates that vary between 1% and 2%, Latin America, North Africa, the Middle East, and Southeast Asia have repetition rates between 6% and 9%. Among the arguments supporting this policy, the promoters emphasize the reinforcement of the student's knowledge or discipline, and the potential beneficial effects on subsequent outcomes. Advocates also point out that grade retention policies could be an efficient mechanism to reallocate students to classes, improving the match between students and schools. The critics, on the other hand, highlight that grade repetition does not lead to improvements in school achievement, decreases self-esteem, creates an adaptation cost (to the new class), and increases the probability of dropping out and criminal activity.

Given this debate, and the prevalence of this policy in developing countries, it is relevant to empirically study the effect of grade repetition on different outcomes, especially considering that the retention rates can be directly affected by policy makers.[1] The challenge, however, is how to estimate causal effects, given that the latent school outcomes – the drop-out or crime activity that would be observed in the absence of grade retention – and the propensity to fail a grade are simultaneously determined.

In this paper we estimate the causal effect of primary school grade retention on juvenile crime. We implement our empirical strategy by using a detailed and novel administrative data on schooling and crime records from Chile (years 2007-2014), which tracks students from primary school to any criminal prosecutions they may have during their youth.

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 to implement standard RD methods. However, there is empirical and anecdotal evidence showing that the forcing variable (the second lowest score subject) is manipulated, because it is arguable that teachers' grading decisions at the margin of repetition may not sort

[1] The policy makers, or school managers, cannot control the effort that students exert or their motivation in attending school, but they can define different rules or standards to fail a grade, conditional on student behavior, and, in this way, they can determine the retention rate.

students randomly.

To deal with this problem, we undertake two complementary empirical strategies. In the first approach, we follow Barreca, Guldi, Lindo, and Waddell (2011) 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. 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 the fact of removing observations does not invalidate the RD assumption about the continuity of the outcome variable's expectation around the threshold, conditional on the same treatment status. In the second approach, we addresses 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).

Our more conservative results (the *donut-hole* FRD) show that grade retention between 4th and 8th grade increases the probability of juvenile crime by 1.2 percentage point (pp), an increase of 25%. The effect is higher for males and for students attending low SES schools, 1.7 and 2.1 pp, respectively. The RD-matching method delivers higher values (but lower than OLS).

We also examine the effect of grade retention on dropping out and future grade retention. To do so, we implement the same empirical strategies, but we change the dependent variables. According to our results, grade retention in primary school decreases the probability of grade retention by 7 pp (14%) in subsequent years and increases the probability of dropping out by 0.9pp (51%). 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 impacts on crime. In addition, the effect of grade retention on dropping out also suggests a relevant mechanism through which grade retention may affect juvenile crime: grade retention impacts on dropping out and dropping out impacts on crime.

We implement a placebo test by replicating the "Donut-hole" RD and the RD-matching estimations, but in this case we compare students scoring below and above the thresh-

old, only among those who did not repeat the grade. These two methods deliver no statistically significant effect in this placebo test for the all three outcomes considered.

This paper makes two main contributions. First, together with a recent paper (Depew and Eren (2015)), it is the first paper 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. 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 when there is some evidence of manipulation in the forcing variable.

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

2 Literature Review

Estimating the causal effect of grade retention on juvenile crime has been scarcely studied in the literature. In our closest paper, Depew and Eren (2015) estimate the impact of grade retention (with summer school) on juvenile delinquency (and school dropout) in Louisiana. 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 since a decade, the authors build a RD design, where the forcing variable is the score in a standard test which determines whether a student is promoted or not. To advance to the next grade any student is required to get a minimum score (in both math and language), generating thus the cut off point. Their principal result is that, for students attending eighth grade, the test-based grade retention policy decreases the likelihood of being involved in felony offences during the youth.

Although Depew and Eren (2015) make a remarkably effort for identifying a causal effect from grade retention to juvenile delinquency, they do not correct the latent manipulation that the forcing variable suffers close to the cut off. 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 standard test near the cut off point. If this holds, the variation in score 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 their figures A1 and A2 (page 47) suggest, it seems that there is a sort of strategic allocation of students around the cut off score: there are no students who score marginally below the minimum required. Given this, manipulation is the most likely explanation of this phenomenon, since any student scoring a few points below the cut off is simply promoted to next grade when in theory he should not advance. Now, as stated in the Introduction, despite facing the same manipulation close to the threshold, we solve this issue by improving the RD framework through matching estimators.

This paper is also related with other papers. Cook and Kang (2013) merge administrative data of academic performance with the criminal record of students attending public schools in North Carolina. They exploit the sharp RD design generated by the cut date for starting school and assess its effect on a number of educational outcomes, as well as on committed juvenile crimes. Their main findings are two. First, those students born just after the cut date are more likely to outperform (in math and reading) those born just before in middle school, 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 paper is different from this one in at least two dimensions: we attempt to estimate directly the causal effect of grade retention on juvenile crime and instead of using the date of birth as a forcing variable, we employ the second lowest score subject for generating a fuzzy RD design, which in turn we improve by combining it with a matching method.

There exists a vast literature studying the effect of grade retention on a number of future educational outcomes such as future grade retention and school completion, as we also consider. For instance, Manacorda (2012) uses administrative data of students in Uruguay to examine the effect of grade retention on school dropout and educational attainment. He also estimates an RD where the forcing variable is the number of failed subjects. However, beyond the results, the most remarkable aspect of this work is that, as our work, it deals with the issue that the assignment around the cut off might not be as good as random via worse-case scenario selection. Jacob and Lefgren (2004) and

Jacob and Lefgren (2009) use a RD framework for assessing the effect of grade repetition on some educational attainments among students attending public schools in Chicago. As Depew and Eren (2015), these works also employ a test-based promotion policy as the forcing variable and show that there exists an heterogeneous impact: a positive short-run effect of grade retention on educational performance in children attending third grade, and that grade repetition among students attending eight grade causes an increment in the likelihood of school dropout.

Finally, in terms of methodology, our paper is close to Keele, Titiunik, and Zubizarreta (2015), who develop a method that combines a sharp RD framework with matching estimator. We extend this approach to a fuzzy RD design in order to deal with the manipulation issue around the cut off point.

3 Data and Grade Retention's Rules

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

3.1 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 charge of providing free criminal defence to those accused in Chile. The final data set includes over 1.2 million students and their juvenile criminal records (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 2013 for all students in the country. This panel includes the school attended every year, the grade level (and whether the students repeat the grade), the student's attendance rate, some basic demographic information, and only for 2007 the annual average score for every subject. 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 a standardized test (the SIMCE), which is taken every year by all students in the 4th grade and every other year by all 8th grade students. When the SIMCE is taken, a survey is administered to the parents. From these surveys, we obtain

information about mother's and father's education and family income. We focus our attention on the students who were in 4th to 8th grade in 2007, attending public schools and subsidized private schools.[2] We drop non-subsidized private schools, which represent 8% of the national enrollment.

The DPP's records contain information on all defendants in criminal cases in Chile during the period January 2006 to December 2014. This database includes information on the time of the accusation, the type of offence, and the verdict (including the length of sentence). In this study, we consider only juvenile criminal cases and we omit individuals who committed the most severe crimes, such as murder or rape.[3] Thus, we focus on crimes that can be thought as motivated by a cost-benefit analysis. Given that our "treatment" is grade retention in 2007, we also drop 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*.

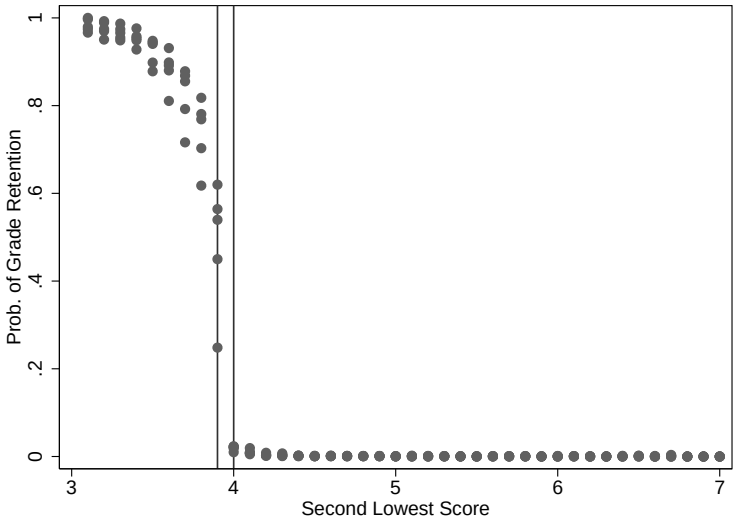
3.2 Rules for Grade Retention

In Chile, grade scores range from 1 to 7, with an increment of 0.1. Although there are other causes of grade retention, the most prevalent cause is scoring below 4 on two or more subjects ($\leq$ 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 in the second -lowest score.

[2]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.

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

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 follow this rule, 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 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 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.

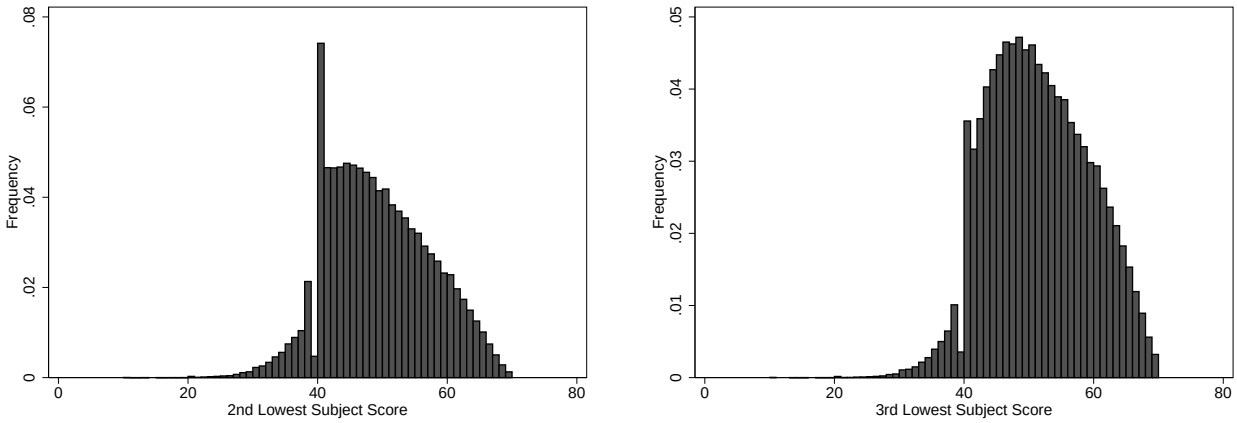
4 Validity of the RD Design

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.

4.1 The Density of the Forcing Variable

Figure 2 shows the histograms for the second and third-lowest scores. There are several lessons from these plots. First, there is a remarkably discontinuity in the histogram for the second-lowest score, around the threshold $(3.9-4)$. Second, this discontinuity is also present in the histogram for the third-lowest score, although to a lesser extent. Third, the discontinuity (and hence the manipulation) seems to be limited to the scores closest to the threshold $(3.9-4)$.

Figure 2: Histograms for the 2nd and 3rd Lowest Score



The first point raises reasonable doubts about the internal validity of a 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 implies that teachers could have other reasons to augment student scores, because the third-lowest score does not impact grade retention.[4]

The last point, which is about local manipulation, is in line with the incentives that teachers face. In fact, even though the anecdotal evidence suggests that school managers promote an upper bound of 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 upgrade that score to a value higher than 4.[5]

[4]We are assuming that manipulation is at the end of the year, when the school manager and teacher can discuss whether to increase the scores of the students close to the threshold.

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

4.2 Tests Involving Covariates

To study the extent to which this manipulation could be a problem and how useful it is to use a 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 to 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 to 0.29.[6] Moreover, all these differences are in the same direction: the repeaters are students with characteristics highly correlated with 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 Group B, where we compare students who score 3.9 in 2007, in their second-lowest subject, with students who score 4 in 2007, in their second-lowest subject. The stories from these two samples are different 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.

[6]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 treatment 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	-65.250	0.000	683972	21293
Attendance 2006	94.6	92.3	0.38	49.705	0.000	683972	21293
Math Simce	0.01	-0.89	1.00	154.636	0.000	683972	21293
Language Simce	0.01	-0.88	0.99	151.548	0.000	683972	21293
Mother Education	10.78	9.50	0.37	52.609	0.000	683972	21293
Father Education	10.86	9.70	0.32	45.761	0.000	683972	21293
Family Income	98960	74412	0.28	43.81	0.000	683972	21293
Male	0.4969	0.6438	-0.30	-42.237	0.000	683972	21293
Crime	0.0235	0.0791	-0.25	-50.967	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	2.654	0.008	2496	885
Attendance 2006	93.5	93.6	-0.02	-0.631	0.528	2496	885
Math Simce	-0.58	-0.65	0.08	2.089	0.037	2496	885
Language Simce	-0.58	-0.62	0.05	1.282	0.200	2496	885
Mother Education	10.43	10.29	0.04	1.074	0.283	2496	885
Father Education	10.54	10.39	0.04	1.077	0.282	2496	885
Family Income	91620	95597	-0.04	-1.08	0.280	2496	885
Male	0.5641	0.5966	-0.07	-1.680	0.093	2496	885
Crime	0.0497	0.0497	-0.00	-0.004	0.996	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	-2.706	0.007	7463	3889
Attendance 2006	93.2	93.1	0.01	0.613	0.540	7463	3889
Math Simce	-0.54	-0.73	0.23	11.888	0.000	7463	3889
Language Simce	-0.57	-0.73	0.19	9.895	0.000	7463	3889
Mother Education	10.36	10.17	0.06	2.822	0.005	7463	3889
Father Education	10.52	10.40	0.04	1.844	0.065	7463	3889
Family Income	88325	88132	0.00	0.11	0.913	7463	3889
Male	0.5748	0.5765	-0.00	-0.170	0.865	7463	3889
Crime	0.0438	0.0550	-0.05	-2.661	0.008	7463	3889

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

The comparison between these two selected samples (Groups A and B) is illustrative about how much we gain by taking advantage of the discontinuity. Without a 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 in observables (*e.g.,* a type of matching) as a credible approach to estimate a causal effect. That said, as it was anticipated in the density analysis, Group B shows some

evidence of manipulation around the threshold, because without manipulation students scoring 3.9 should have –in average– worse performance and lower socioeconomic status than those students scoring 4, which is not always the case in our data. In particular, it is remarkable the difference in the fraction of students who repeated before. A reasonable explanation for this difference is that teachers are more demanding with students who have not failed a grade before, 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 and 4.1 in their second-lowest subject. This selected sample has advantages and disadvantages compared to group B. About the former, the fact all the differences between students bellow and above the threshold have the expected sign is consistent with a situation free of manipulation. Regarding the disadvantages, we loss comparability between the groups bellow and above the threshold, in particular in the performance of the students. 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 be convincing about the need to complement the RD design with another approach to control for the differences in observables.

5 Empirical Approach

Considering the opportunities and problems of 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 Group C sample, we combine a fuzzy regression discontinuity design (FRD) with the matching approach, named *design matching*, developed by Zubizarreta (2012).[7] Finally, our third approach is to implement OLS estimator considering all the students in our sample, and controlling for all the variables used in the other empirical approaches.

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

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),[8] 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.[9]

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

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

[9] 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.

a schools-grade with at least one student at each side of the threshold (Group C, Table 1).[10] 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 unbalance 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 than 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-grade and score 4.1,[11] 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.[12]

We decided to implement this matching approach, as opposed to a more standard type, given that we have a relevant number of school-grade clusters for which there are few students scoring 4.1 to be a match of those scoring 3.8.[13] This situation creates an unbalance in observables that can only be reverted by a method that takes advantage of the school-grade clusters with more options, by not only seeking more similar pairs, but also looking for pairs that compensate this unbalance. For instance, if at the school-grade clusters with only one student scoring 4.1 (*i.e.,* with no option), there is a higher fraction of males bellow than above the threshold, then in the other school-grade clusters the design matching method is going to prefer –in some cases– to match males scoring 4.1 with females scoring 3.8, to compensate the former unbalance.

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

[10] That is, a grade 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.

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

[12] The details of this matching approach are described in the Appendix A.

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

in Group C of Table 1, it is clear that there is an improvement in terms of balance in observables.[14] However, there is an important reduction in the sample size (from 3889 to 2931 individuals bellow 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	-0.551	0.582	2959	2959
Attendance 2006	93.3	93.2	0.02	0.704	0.481	2959	2959
Math Simce	-0.65	-0.68	0.03	1.154	0.248	2959	2959
Language Simce	-0.65	-0.68	0.03	1.277	0.202	2959	2959
Mother Education	10.33	10.26	0.02	0.790	0.430	2959	2959
Father Education	10.51	10.48	0.01	0.343	0.731	2959	2959
Family Income	89012	88476	0.01	0.23	0.821	2959	2959
Male	0.5742	0.5792	-0.01	-0.395	0.693	2959	2959
Crime	0.0412	0.0575	-0.07	-2.881	0.004	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.

6 Results

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

[14] 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-grade clusters for which there are few students scoring 4.1 to be the a match of those scoring 3.8.

6.1 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 1.2 to 3.2 percentage points (pp), and in all specifications the effect is statistically significant.

Table 3: Effect of grade retention on juvenile crime

Sample	(1) Donut-Hole FRD	(2) FRD-Matching	(3) OLS
All	0.012	0.014	0.035
	(0.0060)	(0.0077)	(0.0019)
	$N = 11351$	$N = 5918$	$N = 705083$
Low SES	0.018	0.033	0.044
	(0.0101)	(0.0128)	(0.0027)
	$N = 5345$	$N = 2728$	$N = 358925$
Males	0.021	0.029	0.042
	(0.0090)	(0.0133)	(0.0026)
	$N = 6532$	$N = 2476$	$N = 353425$

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 important. In particular, the impact – measured as percentage points – is larger for male and students from low SES.[15] 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 25% to 67%.[16]

[15]Low SES is defined as the group of students attending schools which are bellow the median in the school average income.

[16]A precautionary note about this range is called for. 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%.

6.2 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.[17] 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.[18] In particular, grade retention in 2007 decreased the probability of future repetitions from 3.8 to 7.8 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 of 7 to 14%.[19]

Table 4: Effect of grade retention on future grade retention

Sample	(1) Donut-Hole FRD	(2) FRD-Matching	(3) OLS
All	-0.070	-0.069	0.116
	(0.0136)	(0.0176)	(0.0035)
	$N = 11351$	$N = 5918$	$N = 705083$
Low SES	-0.039	-0.013	0.108
	(0.0194)	(0.0252)	(0.0047)
	$N = 5345$	$N = 2728$	$N = 358925$
Males	-0.077	-0.099	0.111
	(0.0172)	(0.0256)	(0.0044)
	$N = 6532$	$N = 2476$	$N = 353425$

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 9th and 10th grade. For instance, we say that a student who was attending 4th grade in 2007 dropped out, if she is not attending school in 2012 and 2013.

[17] 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.

[18] Given that there are drop-outs, there is a potential selection bias problem that we have not addressed yet.

[19] The fact that, in the estimation sample, 66% of the students repeat at least one grade after 2007, is a reflection of 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 performance students are overrepresented in the estimation sample.

We follow this definition, to have a comparable measure of drop out, among a group of students who were at different grades in 2007 (4th to 8th grade), and considering that we have schooling data until 2013. Table 5 shows the effects of grade retention on dropping out. Specifically, grade retention in 2007 increases the probability of dropping out, from 0.9 to 1.6 pp (column (1)).[20] Given the measure of drop out used in this paper, 1.7% of the students dropped out after 2007. Thus, these figures represent an increase of 51 to 94%.

Table 5: Effect of grade retention on Dropping out

Sample	(1) Donut-Hole FRD	(2) FRD-Matching	(3) OLS
All	0.009	0.005	0.029
	(0.0039)	(0.0049)	(0.0015)
	$N = 11351$	$N = 5918$	$N = 705083$
Low SES	0.016	0.012	0.040
	(0.0071)	(0.0087)	(0.0023)
	$N = 5345$	$N = 2728$	$N = 358925$
Males	0.012	0.008	0.029
	(0.0050)	(0.0068)	(0.0019)
	$N = 6532$	$N = 2476$	$N = 353425$

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.

Besides the discussion about magnitudes, there are several aspects of these results that are important to highlight. First, as in 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 more 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

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

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.

6.3 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; (2) given this sample, we estimate a standard sharp design RD, by regressing the following equation:[21]

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

Along the same lines, we re-estimate the RD-matching 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 the matched sample, the LATE parameter (τ) is estimating by regressing equation 4.[22]

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 compare students scoring below and above the threshold, only among those who did not repeat the grade.[23] For instance, in the case of the "Donut-

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

[22] 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 bellow the threshold. These samples would be different due to the fact that design matching involves a constraining in the measures of imbalance to be less or equal than certain tolerances.

[23] In principle, we could do the same by comparing those who are bellow and above the threshold and repeated the grade. However, we do not have enough sample size to do that.

hole" RD, we proceed in 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 $E[Y_i|Z_i = 0, W_i = 0, X_i] = E[Y_i|Z_i = 1, W_i = 0, X_i]$.[24]

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

	Sharp Design		Placebo	
	(1)	(2)	(3)	(4)
Sample	Donut-Hole RD	RD-Matching	Donut-Hole RD	RD-Matching
All	0.015	0.016	-0.004	-0.004
	(0.0051)	(0.0063)	(0.0062)	(0.0076)
	$N = 10255$	$N = 5116$	$N = 8468$	$N = 3717$
Low SES	0.020	0.030	-0.003	0.009
	(0.0086)	(0.0108)	(0.0115)	(0.0149)
	$N = 4865$	$N = 2393$	$N = 3956$	$N = 1683$
Males	0.024	0.030	-0.005	0.005
	(0.0078)	(0.0113)	(0.0105)	(0.0154)
	$N = 5968$	$N = 2185$	$N = 4797$	$N = 1513$

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 significance, 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.[25]

[24] See Imbens and Rubin (2015).

[25] That said, it is important to consider that the robustness of our approaches critically depend on the

Finally, in appendix B, we present the robustness analysis for drop out and grade retention after 2007 (Tables 7 and 8). As in the case of crime, in the placebo test all parameters are not statistically significant in the case of future grade retention and drop out. These results confirms the soundness of our empirical strategy to find causal estimates.

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 in the forcing variable, we depart from standard RD methods. First, we follow Barreca, Guldi, Lindo, and Waddell (2011) 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 has two main contributions. First, together with a recent paper (Depew and Eren (2015)), it is the first paper 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.[26] That said, the interpretation of our findings should consider that we are not taking into account other aspects of this policy, *e.g.,* the threat of retention could be an incentive to exert more effort for all the students. 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.

There are at least three interesting extensions that can follow what we have done in this paper. First, it could be more precise, and may be feasible, to find an empirical

level of the imbalance in observables to begin with. For instance, we run the same placebo approaches but comparing students who score 4.1 with students who score 4.4, and we did find 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 used in our estimation.

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

approach to estimate the effect of grade retention on crime, taking into account that some of the students who repeat in earlier grades would instead repeat in a later grade (to go beyond the *lower bound* of the effect). The challenge in doing so is that we only have the score in each subject for the year 2007. Second, it is important to be clearer about the interactions among crime, dropping out and grade retention. Third, there are more rules of grade repetition in the Chilean educational system, beyond the one discussed in the paper. Thus, instead of using the more prevalent one, we could find a way to estimate the effect by exploiting different discontinuities at the same time.

A 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).[27] 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 at past 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 at past 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.

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 obfective function to minimizes is equal to:[28]

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

subject to pair matching and covariate balancing constrains. 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,

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

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

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

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

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

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

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

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 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 repetitions, attendance at past 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 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).[29] 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

[29] λ can be thought 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.

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.

B Robustness Analysis: other outcomes

Table 7: Effect of grade retention on future grade retention (sharp design and placebo)

	Sharp Design		Placebo	
	(1)	(2)	(3)	(4)
Sample	Donut-Hole RD	RD-Matching	Donut-Hole RD	RD-Matching
All	-0.066	-0.068	-0.013	-0.007
	(0.0110)	(0.0140)	(0.0163)	(0.0198)
	$N = 10255$	$N = 5116$	$N = 8468$	$N = 3717$
Low SES	-0.037	-0.017	-0.009	0.006
	(0.0159)	(0.0205)	(0.0246)	(0.0302)
	$N = 4865$	$N = 2393$	$N = 3956$	$N = 1683$
Males	-0.080	-0.106	-0.001	0.008
	(0.0143)	(0.0212)	(0.0224)	(0.0316)
	$N = 5968$	$N = 2185$	$N = 4797$	$N = 1513$

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 8: 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	0.009	0.005	-0.000	-0.000
	(0.0034)	(0.0040)	(0.0041)	(0.0052)
	$N = 10255$	$N = 5116$	$N = 8468$	$N = 3717$
Low SES	0.015	0.012	0.001	-0.000
	(0.0061)	(0.0074)	(0.0082)	(0.0098)
	$N = 4865$	$N = 2393$	$N = 3956$	$N = 1683$
Males	0.011	0.008	-0.001	-0.001
	(0.0044)	(0.0058)	(0.0055)	(0.0078)
	$N = 5968$	$N = 2185$	$N = 4797$	$N = 1513$

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

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.

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.

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.

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 (2004): "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *The Review of Economics and Statistics*, 86(1), 226–244.

——— (2009): "The Effect of Grade Retention on High School Completion," *American Economic Journal: Applied Economics*, 1(3), 33–58.

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.

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

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

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.