I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11604

The Effect of Grade Retention on Secondary School Performance: Evidence from a Natural Experiment

Maria Ferreira
Bart H.H. Golsteyn
Sergio Parra-Cely

JUNE 2018

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11604

The Effect of Grade Retention on Secondary School Performance: Evidence from a Natural Experiment

Maria Ferreira
Maastricht University

Bart H.H. Golsteyn
Maastricht University and IZA

Sergio Parra-Cely
Maastricht University

JUNE 2018

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

The Effect of Grade Retention on Secondary School Performance: Evidence from a Natural Experiment*

We study the effects of grade retention on secondary school performance by considering a change in Colombia's educative legislation. In 2010, the rule that forced schools to retain up to a 5% of students was abolished. Exploiting variation in schools' retention rates in a difference-in-differences framework, we find that retained (marginally non-retained) students improve (decline) their performance on language but not on math test scores. We suggest the school's position in the retention distribution, and the proportion of inexperienced teachers in the classroom, can be the mechanisms by which the marginally decreasing returns of grade retention are determined.

JEL Classification: I20, I24, J24

Keywords: retention, Colombia, difference-in-differences

Corresponding author:
Bart H. H. Golsteyn
Department of Economics
Maastricht University
P.O. Box 616
6200 MD
Maastricht
The Netherlands
E-mail: b.golsteyn@maastrichtuniversity.nl

* This research is partly financed by a VIDI grant from the Netherlands Organization for Scientific Research (NWO). The authors are grateful to Lex Borghans, Giorgio Brunello, Andries de Grip, Caroline Hoxby, Olivier Marie, Annemarie Künn-Nelen, Steffen Künn and participants at the Maastricht University's Workshop in Economics, the 32nd conference of the European Society for Population Economics in Antwerp, and the 9th IWAEE workshop in Catanzaro for invaluable comments and suggestions. We also thank the Colombian Inspectorate of Education (ICFES), the National Bureau of Statistics (DANE) and the CEDE institute at Universidad de los Andes, Colombia, for sharing their datasets and school-identifiers on the SABER11 test and the C-600 census.

I. Introduction

Retention in school is common and widespread,[1] but its consequences for school performance are theoretically unclear and empirically diverse. Effects can be expected both for retained and non-retained students. For the retained, there may be positive effects as repeating a grade can help to acquire basic knowledge needed to perform well later on. But retention may instead also have negative effects on school performance if, for instance, self-esteem and motivation decrease as a result. For non-retained students, the relationship between grade retention and performance works via different mechanisms. In principle, students at the upper end of the ability distribution may learn more as the level of teaching adjusts accordingly if weaker students in class are retained. A positive effect of retention at the lower end of the ability distribution may be that the threat of being held back can stimulate children to work harder in school.[2] But this threat may also have negative consequences, as there is a negative correlation between mental stress and academic performance. Taken together, empirical research on the effect of retention on school performance is needed as its expected effects are ambiguous from a theoretical point of view.

An important empirical challenge in studying the relationship between grade retention and school performance is that omitted variables may drive the relationship. For instance, high ability children may be less likely to be retained and may also obtain higher school grades. This implies that a naïve estimation of the effect of retention on academic achievement may be negatively (positively) biased for retained (non-retained) students. In this paper, we propose a framework to recover the causal effect of grade retention on secondary school performance of retained and non-retained students, which combines both administrative data and students' academic records. Specifically, we analyze the effect of retention in 10th grade on performance in 11th grade, the last year in secondary school (nominal age: 16-17), using two administrative datasets from Colombia. The first, provided by the Inspectorate of Education, includes data for all students in the country on scores from a centralized exam in the last year of secondary education. The second dataset, from the central statistics office, contains information on retention rates across all schools and grades in the education system. We are able to link the two datasets using unique school identifiers.

To overcome the endogeneity problem aforementioned, we exploit a policy change with respect to retention. From 2002 to 2009, under the automatic promotion policy regime, schools were by law not allowed to retain more than 5 percent of their students. In 2010, this directive was abolished and since then, schools are free to decide how many pupils should repeat a grade. The abolishment of the law increased retention rates dramatically in some schools, while in oth-

[1] In the United States, around 10 percent of all students are retained between kindergarten and eighth grade. In Germany and France, respectively 9 and 18 percent of all students are retained in primary school (Fruehwirth et al., 2016).

[2] One may also argue that this threat leads to other effects. Belot and Vandenberghe (2014) exploit a law reform to find that an enhanced threat of grade retention does not lead to better medium-term outcomes.

ers it had no effect. We use this information in a difference-in-differences approach, in which treatment and control groups are defined by the above-median historical increase in retention attributed to the law change, analogue to the method used in recent papers in a different context than ours (Havnes and Mogstad 2011; Bauernschuster et al. 2016). Schools in which retention rates increased more than the median change are labeled the treatment group and those that responded less than the median, the control group. Several placebo and falsification tests show that trends in math and language test scores are similar in control and treatment schools before the law was repealed, indicating that we can use a difference-in-differences model to estimate the causal effects of increased retention on school performance.

Our findings indicate that students who have been exposed to higher retention rates one (two) year(s) before taking the high-school exit exam obtained lower (higher) scores on the language test. In contrast, we observe no significant effects on math scores. We attribute the effects of 1-year and 2-years prior increased retention exposure to marginally non-retained students, and retained students, respectively. Additional estimations, where we classify students between those who were plausibly retained (age at exam: 18 years old or older) and non-retained (age at exam: 17 years old or younger) corroborate this interpretation of the results.

When analyzing the results across the distribution of language scores we find that especially low performing, retained students benefited from increased retention. These results suggest that by repeating a class, students at the lower end of the ability distribution get a more thorough understanding of the material which enables them to perform better later on. Distinguishing between low, middle and highly treated schools reveals the non-linearity of the effect of retention: in middle treated schools, i.e. schools that moderately increased retention, the improvement is more pronounced than in highly treated schools. These results imply that increased retention not only has marginal decreasing returns, but also that some schools may be retaining students at nearly optimal levels.

For marginally non-retained students, the effects of increased retention are negative on language scores, especially at the lower end of the test performance distribution. These results remain both qualitatively and quantitatively robust to the inclusion of potentially important control variables, clustering of standard errors at the school level, and when performing other robustness checks. Theoretically, there are several reasons for this negative effect. Firstly, there might be a positive selection effect for non-retained students that is dominated by the plausibly negative influence of less able peers in the classroom. Secondly, students may strategically substitute effort between stem and non-stem subjects as the probability of repeating a grade rises. Decreasing marginal productivity in both courses implies that math scores are not expected to increase as much as language scores decrease.

While the above explanations are plausible, we cannot test them as our sources of information do not contain data on individual retention outcomes, intrinsic/extrinsic motivation, and/or individual effort across academic subjects. Instead, our analysis is focused on school-driven mechanisms that might explain the impacts we identify for retained and marginal students

alike. We provide evidence that neither average class size nor teachers' average educational achievements are relevant to explain these findings. In contrast, increasing participation of positively selected, inexperienced teachers in the classroom seems to be the amplifying force behind the benefits and costs of increased retention. Newly hired teachers are highly motivated but their effort may be (strategically or not) directed to fulfilling the educational needs of good students, at the expense of the less able pupils.

Our analysis contributes to the literature in various disciplines that have studied the effects of retention on school performance. Several articles in School Psychology and Sociology of Education analyze the relationship between grade retention and later school performance, mostly reporting this relationship to be negative. McCoy and Reynolds (1999) report that retention has a negative relationship with reading achievement. Jimerson et al. (1997) find no evidence that retention is related to school performance. Jimerson (1999) follows students for 21 years in a longitudinal study to show that retained students have worse educational and employment outcomes in late adolescence. Silberglitt et al. (2006) find that retained students made less educational progress compared to a random group of other students. Stearns et al. (2007) report that students who repeat a grade prior to high school have a higher risk of dropping out of high school than students who are continuously promoted. An important caveat is that these articles report correlations and not causal estimates. Although correlations are informative, important confounders may bias such estimates. As previously explained, we expect a downward (upward) bias for retained (non-retained) students.

There is a small but growing literature that estimates the causal effect of grade retention on subsequent educational outcomes.[3] The literature reveals that results are mixed, documenting positive as well as negative estimates. The results depend on the context and age of students. Firstly, some papers study the effects of retention at young ages. Koppensteiner (2014) examines the effect of automatic grade promotion on academic achievement (math scores) at primary school in Brazil. Applying a difference-in-differences approach that exploits variation over time and across schools in the grade promotion regime, the author finds a negative and significant effect of about seven percent of a standard deviation on math test scores. Fruehwirth et al. (2016) evaluate the effect of retention on achievement using data from children in kindergarten. Accounting for dynamic selection into retention, they find that children who are retained in kindergarten would have performed as much as 27 percent higher on math and reading tests in the next year if they had not been retained. Jacob and Lefgren (2004) instead find positive effects of retention at an early age. They assess the effects of retention in the Chicago Public School system using variation in retention generated by a test-based promotion

[3]There may also be peer effects of retention. Hill (2014) investigates the extent to which course repeaters in high school mathematics courses exert negative externalities on their course-mates. Using individual and school-specific course fixed effects to control for ability and course selection, the study shows that increasing the share of repeaters in each course results in a moderate, significant increase in the probability of course failure for first-time course-takers. Results suggest that the negative effect is only evident when the share of repeaters reaches a threshold of 5 to 10 percent of the total number of course-takers.

policy, and find that retention has a modest but positive net impact on test scores for third grade students, while it increases academic achievement for low-achieving third graders. However, they also find that retention appears to have little or no effects for sixth-grade students.

Secondly, some studies have assessed the effects of retention on achievement in high school. A first set of papers reports negative effects. Jacob and Lefgren (2009) show that retention among younger students (sixth grade) does not affect the likelihood of high school completion, but retaining low-achieving eighth grade students in elementary school increases the probability that these students will drop out of high school. Manacorda (2012) studies the effects of retention in secondary junior high school (grades 7 to 9) in Uruguay on dropout rates and school attainment, exploiting a discontinuity established by a rule of automatic grade failure for pupils with more than three failed subjects at the end of the school year. The analysis reveals that retention increases school dropout and reduces school attainment. While analyses in secondary education focus mostly on dropout rates or completion of school, García-Pérez et al. (2014) measure the effect of grade retention on Spanish students' PISA math scores at age 15, using the student's quarter of birth as an instrumental variable. They find that grade retention has a negative impact on educational outcomes. Those who are retained during primary education suffer more than those retained in secondary school.

Contrary to these findings, a second set of papers provides estimates of positive effects of retention in high school. Mahjoub (2017) finds large positive effects of retention on test scores: around 1.6 times the standard deviation of the achievement gain, using quarter of birth as an instrument. The average effect of the treatment on the treated (ATT) ranges between one and one-quarter of a standard deviation of the test scores. Grade repetition in junior high school is also shown to increase the probability of graduation by 2.5 percentage points. Eide and Showalter (2001) use an instrumental variable for retention based on exogenous variation across states in kindergarten entry dates to find tentative evidence that retention may benefit students by both lowering dropout rates and raising labor market earnings. They find these effects to be relevant for white students, but not for black students.

A common approach in these studies is that the benefits of retention are evaluated at the margin where retention was increased by the natural experiment. An important issue with this approach is that the estimated benefits may differ at other moments of the distribution of students. For low performing students, the benefits of repeating a class may be positive while for high performing students there are probably negative effects. Schools are aware of this and aim to retain students until the marginal student does not benefit from retention.

The main contribution of our study is that we analyze the non-linearity of the effect of retention on test scores at various moments of the ability distribution. We show that modest increases in retention lead to higher scores in language for the retained students, but when many students are retained, such gains decrease.

This analysis further contributes to the literature studying the effects of retention on educational outcomes. Firstly, by separately analyzing the effects of retention for retained and

marginally non-retained students. As indicated earlier on, the expected effects of retention are different for these groups. To evaluate the costs or benefits of retention for society, it is important to take the effects for both groups into consideration. Second, we also test transmission mechanisms at the school level, highlighting the role of teachers' staff composition on determining the differentials in test scores we observe as an outcome of increased retention. Third, our empirical approach to elicit causal effects departs from most other papers in this literature. We exploit the effects of a law change, which enabled schools to retain more children. Finally, we provide evidence on the effect of retention for a developing economy using a large administrative dataset, representative of the Colombian educational system.

Closest to our approach is the analysis developed by Koppensteiner (2014), who examines the effect of automatic grade promotion on academic achievement (math scores) in primary schools in Brazil using a difference-in-differences approach. Besides that we evaluate effects of retention separately for retained and non-retained students, and that we study the effects in secondary education and not in primary education, our study differs from his in the sense that we can show with placebo tests that the pre-treatment trends in school performance are common in treatment and control groups, i.e. the key underlying assumption of the difference-in-differences framework. Koppensteiner (2014) shows instead that school and student characteristics of treatment and control groups tend to be similar before the treatment occurred.

The setup of this paper is as follows. Section II summarizes the Colombian context and the educational reform we exploit. Section III discusses the empirical strategy in detail. Section IV describes our main sources of information, and the final dataset. Our central findings, including relevant robustness checks, are presented in Section V. Section VI offers a discussion on the transmission mechanisms. Finally, Section VII concludes.

II. Institutional Background

A. The Colombian Educational System

Colombia has an eleven-year system of elementary and secondary education, consisting of five years of primary school (1st to 5th grade), four years of lower secondary education (6th to 9th grade) and two years of upper secondary education (10th to 11th grade).[4] The expected age of entry to 1st grade is six years.[5] Therefore, if children are not retained, they are expected to complete their secondary education at ages 16-17.

[4]Elementary and secondary education in Colombia is offered in two school calendars: A calendar labeled "A" that runs from February until November, and a calendar "B" from September to June. Most schools (92%) in the country operate in calendar A. Formal education is also offered by schools in three different class-schedules: a morning schedule, an afternoon schedule, and a full-day schedule. Students opt or are allocated by the school to attend either one of these. Most students in secondary education attend school either in the morning or the afternoon schedule (78%).

[5]This age is suggested but not mandatory as in Colombia there are no compulsory age-at-school entry laws.

The educational system in Colombia is a comprehensive school system with no academic tracking at any grade.[6] However, at the start of upper secondary education, schools differentiate in the provision of additional courses to complement the compulsory curriculum set by the Ministry of Education. These additional courses are organized in two specialization programs: one is more academic and the other more technical in nature. The academic program provides general education in arts, sciences and humanities, whereas the technical program provides vocational knowledge and practice in technology, craft industry, business, pedagogics, or agriculture.

Upon completion of the 11th grade of secondary school, all students, regardless of the chosen program, participate in a national standardized exam ("SABER11," in Spanish), an achievement and competency test that is administered every year by the National Institute for the Assessment of Education ("ICFES," in Spanish).[7] This exam is a high-stakes evaluation, required not only for admission to tertiary education, but also to receive the high-school diploma. This test is also widely considered as the reference examination to evaluate the quality of secondary education across the country. In line with previous literature on grade retention, we focus on students' performance on the math and language parts of the test as the main outcome of our analyses.

B. The Automatic Promotion Policy Rule (AUP)

In 2002, by mandate of the Ministry of Education (Decree 230 of 2002), schools were each year permitted to retain up to a maximum of five percent of their students. This retention policy was implemented to reduce costs attributed to higher retention rates (i.e. low performance, low motivation, dropouts, etc.) without compromising the quality of education provided by the system (Martínez and Herrera, 2002). According to the policy mandate, a student should have been retained if at least one of the following three circumstances occurred: i) the student received an unsatisfactory performance evaluation in three or more academic subjects in the current academic year, ii) the student received an unsatisfactory performance evaluation in math and/or language courses during the current and two previous grades, or iii) the student failed to attend at least 25% of all academic activities during the current academic year. However, schools were required to adjust their evaluation standards to comply with the law, which forced them to promote at least 95% of all their students.

While the reform was marketed as moderately successful in terms of reducing school dropouts, the incentives to underperform at school as perceived by schools, teachers, and parents, led the Ministry of Education to revoke the Automatic Promotion Policy Rule.[8] In February 2009, the 5% retention restriction was replaced by the Ministry of Education through a

[6] The Ministry of Education regulates all levels of education and national exams for both publicly and privately funded schools.

[7] Hereafter, we will refer to this institute as the Inspectorate of Education.

[8] Ministry of Education, Press Release April 17, 2009.

new regulation mandate (Decree 1290 of 2009), allowing schools from 2010 onwards to retain as many students as they considered necessary, and thereby giving them more discretion in their evaluation and promotion procedures. We use the terms Automatic Promotion Policy (AUP) to indicate the period until 2009 and Free Retention Policy (FRP) from 2010 onwards. Overall, the abolition of the AUP regime increased students' retention rates across all grades of secondary education from 4.3 percent to 7.7 percent, on average, in all schools in the country.

III. Empirical Strategy

We evaluate the effects of retention in 10th grade on math and language performance in the secondary school exit exam conducted at 11th grade. The empirical challenge in studying this question is that omitted variables may drive the relationship. A naïve estimation using OLS may be negatively biased for retained students if the lower scores they obtain are not due to retention but to their lower ability. As students' ability increases, we might expect the benefits of grade retention to be decreasing and, for the upper end of the ability distribution, to negatively impact academic performance. Nonetheless, such counterproductive effects may be veiled by, for instance, the positive sorting of skilled students in subsequent grades as a byproduct of increased retention.

We exploit a policy change in Colombia that occurred in 2010 and implied that schools facing constraints in their retention requirements (AUP regime) were allowed to retain as many students as they considered appropriate (FRP regime). To identify the effect of grade retention on test scores, we implement a difference-in-differences framework which exploits the school-year variation on retention rates.

Schools reacted differently to the new policy, suggesting that the grade retention effect is heterogeneous across schools with similar characteristics. We classify schools in two groups: the treated group, consisting of schools that increase their retention rates after the law change, and the control group, composed by schools in which retention rates remained relatively constant. We classify schools into the treated or control categories using the difference between the schools' average retention rate at 10th grade between both policy regimes, the AUP regime (2007-2009) and the FRP regime (2010-2012). Sorting schools on this difference, we define the treated group as the pool of schools with an above-median increase in their retention rates, and the remaining schools are labeled the control group.[9] Panel (a) in Figure 1 shows the retention rates for treated and control schools across years. In control schools, on average, such rates decreased slightly after the law change by approximately 1.4 percentage points. In contrast, treated schools increased retention rates by 7.6 percentage points, implying that the latter retained 9.0 percentage points more than control schools.

[9]This treatment-control classification is increasingly implemented in the economics literature. Examples of this strategy are provided by Havnes and Mogstad (2011) and Bauernschuster et al. (2016). These authors analyze the effects of increased child care coverage on parental economic outcomes.

Additionally, We classify treated schools based on quintiles of the difference in retention rates. Panel (b) in Figure 1 shows the retention rates across time for the four quintile groups in which schools raise retention during the FRP regime. First, we observe that retention rates among schools in the second quintile barely change. These schools can be considered as an alternative "control" group. Second, we observe three groups of schools (quintiles 3 to 5) that are affected differently by the policy change. Since retention rates in these groups increase on average by 2, 6, and 12 percentage points, we label these schools as low treated, medium treated, and highly treated, respectively. Furthermore, we observe schools in the first quintile as a group of defiant schools since they *decrease* retention rates by 4 percentage points. As we consider these schools not being fully comparable with the universe of compliant schools, we decided to exclude them only from this specific analysis.[10] With the exception of this latter group, all remaining schools retained students in the AUP regime as required, with an average retention rate of 3.6 percent.

The baseline difference-in-differences specification we implement is:

$$Y_{st} = \alpha_s + \delta_t + \sum_{h=1}^{2} \gamma_h [Group_s \times FRP_{t-h}] + \beta X_{st} + \varepsilon_{st}, \tag{1}$$

in which Y_{st} denotes standardized test scores for school s in exam year t. α_s and δ_t are fixed effects by school and exam year, respectively. $Group_s$ is a dummy variable that takes value 1 for schools in the treated group, and zero for schools in the control group. In our basic specification, treated schools ($Treated_s$) are those with above median changes in retention rates and control schools with a below-median change in retention rates. In our more elaborate specification, we include three treatment dummies corresponding to low treated ($LowTreated_s$), medium treated ($MiddleTreated_s$), and high treated schools ($HighTreated_s$). FRP_{t-h} is an indicator variable with value 1 if the FRP regime was in place h years before exam year t, and zero otherwise. The interaction term $Group_s \times FRP_{t-h}$ therefore measures the variation in tests scores that can be causally attributed to the shift in the retention policy from year $t-h$ onwards. For each regression, we run two specifications. First we run our regressions without covariates. Second, for each exam year we include covariates for the first two lags of school-specific attributes that change over time. In this way, we account for pre-FRP variation in characteristics among schools. This set of control variables is denoted in the equation as X_{st}. Finally, ε are standard errors clustered at the school level.

We aim to analyze the effects of increased retention separately for retained and non-retained students. However, it is worth explaining how we obtain these effects as we do not observe individual retention outcomes (see more on this point in the data overview section). The cohort of students taking the exam in year t is largely composed of two types of students: i) 10th grade

[10] Namely, we keep all observations in our estimations, but we refrain from interpreting effects for schools in the defiant group. This restriction in the our analyses is only relevant when we account for the heterogeneous effects of retention. In contrast, results involving treated and control schools classified by the use of the above-median increase in retention correspond to all schools in our sample.

students in year $t-1$ that were promoted to 11th grade at year t, and ii) 10th grade students that were retained in year $t-2$, repeated and passed 10th grade in year $t-1$, and finally enrolled to 11th grade in year t. Hence, our parameters of interest are γ_1 and γ_2.

The first parameter measures the effect of being exposed one year to the FRP policy in 10th grade on schools' tests scores in 11th grade the next year. The expected direction of this effect is ambiguous. On the one hand, the sign of γ_1 reveals whether non-retained students benefited from higher retention rates because of a positive sorting effect. In such a case we expect the effect to be positive. On the other hand, if we interpret this coefficient as the effect of increased retention on the marginal student (i.e. students that should have been retained but were promoted by a very small margin), we might expect the impact to have the opposite sign relative to the effect of retention on the retained students, For instance, if the latter effect is positive, marginal students are worse off when promoted to 11th grade because they will miss the chance to receive further training on the academic subjects they struggled with the most. The second parameter γ_2 measures the impact of FRP regime's exposure in the previous two consecutive years on schools' test scores. Assuming that students are retained in 10th grade only once, this impact can be mostly attributed to retained students. Because of our treatment-control classification, γ_1 and γ_2 are best interpreted as intention-to-treat effects (ITT). To obtain the average treatment effects on the treated (ATT), we will rescale these coefficients by the difference in retention rates between treated and control schools implied by the law change.

The main identification assumption in this setting is that the variation in retention rates is orthogonal to expected changes in test scores. This assumption is equivalent to claim that treatment and control schools would have shared similar trends in test scores if the retention policy had remained the same. We formulate an alternative specification to test this assumption:

$$\begin{aligned} Y_{st} &= \alpha_s + \delta_t + \sum_{k=2008}^{2010} \mu_k [Group_s \times (Year = k)] \\ &\quad + \sum_{k=2011}^{2013} \theta_k [Group_s \times (Year = k)] + \beta X_{st} + \varepsilon_{st}. \end{aligned} \tag{2}$$

In equation (2), the null hypothesis of interest is that pre-FRP differences in trends between treated and control units are not significantly different from zero (i.e. $\mu_{2008} = \mu_{2009} = \mu_{2010} = 0$). Namely, we control for the interaction between the treatment status and those exam years where test takers, by construction, were not exposed to increased retention rates because of the policy change. As we will elaborate further on, we are not able to reject such hypothesis at conventional significance levels.

In addition to the above specification, we also perform placebo tests to account for artificial policy changes that should not have any effect on test scores:

$$Y_{st} = \alpha_s + \delta_t + \sum_{h=1}^{2} \pi_h [Group_s \times FakeFRP_{k,t-h}] + \beta X_{st} + \varepsilon_{st}. \quad (3)$$

In equation (3), $FakeFRP_{k,t-h}$ is an indicator variable that takes the value 1 if the FRP regime was in place during year $t-h$, assuming it (artificially) started either in $k = 2008$ or $k = 2009$. By not being able to reject the null hypothesis of non-significant effects (i.e. $\pi_1 = \pi_2 = 0$), we are confirming that the changes in test scores can be attributed to the elimination of the AUP regime only.

IV. Data Overview

Ideally, a suitable dataset to identify the causal effect of grade retention on school performance should meet two conditions: First, information on individual retention needs to be available. Second, variation on retention outcomes needs to be as good as random. For example, suppose that the Ministry of Education randomizes the obligation to retain no more than 5% of students across all schools at an specific date. Then, if schools' attributes and outcomes have been followed through time, it is straightforward to adopt a difference-in-differences framework to recover the effect of such a policy intervention. A variant of this experimental setting, in a panel data structure, would be that the compliance to the retention law in time changes randomly across schools with similar characteristics.

Our dataset does not contain individual level retention data. However, this does not pose a threat to our study as we are able to recover school-level retention rates from administrative sources. As we use retention rates at 10th grade, the pool of test-takers in our sample consists of non-retained students and students who were retained in 10th grade only once. Controlling for the first two lags of our "FRP regime exposure" variable, we can differentiate the effect for each type of pupil. The second requirement is fulfilled since, conditional to the schools' treatment-control classification discussed before, variation in retention rates is attributed to the policy change alone.

A. Sources of Information and Sample Selection

The sample we use in this paper is taken from two main sources. The first is a dataset from the Colombian Inspectorate of Education. The Inspectorate provides freely downloadable micro-level data on the centralized exam conducted among 2.7 million pupils in their last year of secondary education (11th grade).[11] This exam, known as SABER11, is a standardized test that evaluates every year a range of seven school subjects.[12] Test scores range from 0 to 100 in each

[11] This exam takes place every year in the month of September, three months prior to the official end of the school calendar A.

[12] These subjects are: Math, Language, Physics, Biology, Chemistry, History, and Philosophy.

subject and they are standardized by subject at the national level, so that each student's score is informative about his/her position relative to the national average in that subject. According to the Inspectorate of Education, the tests are comparable for the period 2000-2013.

We use available data from 2007 to 2013 that include math and language scores; student and school identifiers; some schools' attributes such as the academic calendar, daily class schedule, public or private status, specialization programs offered; and information on several individual characteristics such as age, gender, mother's education, and other socio-demographic indicators. We collapse these data at the school-year level, and focus only on our outcomes of interest (i.e. math and language test scores) at several moments of the distribution: mean, 10th percentile, 25th percentile, median, 75th percentile and 90th percentile of the schools' test scores. These scores are all standardized over the entire sample to interpret the effects in terms of standard deviations (SD) at the school level. While representative of the student body that is assessed at the last year of secondary school, this dataset does not collect information on pupil's retention at any stage of the education process.

To obtain retention rates at the school level we rely on the schools' official census, which the Ministry of Education releases each year for public use through the National Bureau of Statistics (DANE, in Spanish). Known as the C-600 census, this dataset contains information on academic indicators that all schools in the country are compelled to report on a yearly basis. We use information from this dataset on retention rates at 10th grade, as well as other school characteristics, such as the number of groups per grade, the number of students enrolled at 10th grade, the number of teachers with a professional degree, the number of teachers hired under the old and new pay scales regulated by the central government, and the number of non-academic staff (managerial, support, health) per school. We use this universal census information for the period 2005-2012. Using unique school identifiers, we are able to match 88.2% of all schools' test scores to the respective school retention data for the entire period 2007-2013. These matched data correspond to 85.6% (N=2,363,997) of all students that took the exam during the same period.

Our unit of observation is a school-exam year combination. The estimation sample consists of first-time SABER 11 test takers[13] from schools that i) offered education exclusively in Calendar A (February to November), ii) did not change this calendar during the period 2007-2013, iii) had no missing values on tests scores, retention rates, and schools' covariates, and iv) reported information on retention rates for at least 3 years, with at least one year before and after the retention law changed. The resulting dataset consist of an unbalanced panel of 6,248 schools, which in total across the 2007-2013 period contains 35,693 observations.

[13] We leave out of the sample the top 1% and bottom 1% of students in terms of their age reported at the exam. This selection criterion excludes extreme outliers who reported ages below 12 or above 40.

B. Common Trends

Figure 2 in panels (a) and (b) shows the average test scores across time for treated and control schools in math and language, respectively. On average, control schools performed better at both subjects during the AUP regime. For instance, students in control schools scored 0.1 of a standard deviation more in the math exam than students enrolled in treated schools. The same patterns are observed in the language exam with students scoring around 0.12 of a standard deviation higher than students in treated schools. Appendix Figure A.1 shows the same results when we plot the residuals from a regression including exam-year and school fixed effects, as well as time-variant school attributes. Figure 3 reveals that such trends are also common when distinguishing the treatment groups for high, middle, and low treated schools. These results remain robust to the inclusion of school-specific covariates (Appendix Figure A.2).

The main conclusion from these graphs is that there is a common trend in test scores between treated and control schools. This allows us to use a difference-in-differences strategy. The difference-in-differences estimator will isolate time invariant confounding factors, leaving the remaining variation to be attributed to the effect of increasing retention in schools. In the results section, we will provide robust statistical evidence that the common trend assumption holds.

C. Summary Statistics

Table 1 reports summary statistics for the sample used to estimate our main results. We present information on schools' characteristics during the AUP regime. Columns (1) and (3) report the number of schools per treatment status, and columns (2) and (4) present the averages of each control variable for both treated and control groups, respectively. Columns (5) and (6) report differences in means and standard errors between treated and control schools.

Schools differ systematically in their attributes during the AUP regime. Considering socio-demographic attributes of students, control schools present a more favorable composition of students from highly educated households (measured as the proportion of mothers with tertiary education), and few students with poverty status. There is also a larger proportion of public schools in this group, relative to the treatment group. However, treated schools operate under shorter working spells relative to schools in the control group. Moreover, treated schools also seem to present some academic differentiation as they also provide other types of training (e.g. pedagogical, technical vocational training).

With regards to school-related characteristics, treated schools, on average, have more groups per grade and more qualified teachers employed at school. Regarding teachers' compensation and renewal of personnel, we observe that treated schools hire slightly more staff under the new pay scale than control schools, but the overall proportion of teachers under the new pay scale had increased during the last three years of the AUP regime for all schools. In contrast, control schools seem to employ more health professionals (e.g. dentists, physicians) than

treated schools. Conversely, treated schools seem to hire more staff for managerial purposes than control schools.

While time-invariant differences between treated and control schools are controlled for by the inclusion of school-specific fixed effects, a potential concern for the identification strategy implemented in this paper is that time-varying school's attributes change at the same time the policy intervention does. To address this issue, in some specifications we include school-specific, time-varying attributes one and two years prior the exam date. As we will elaborate further in our results section, our estimates remain invariant to the inclusion of these controls, suggesting that we identify the effect of retention, net of other elements affecting test scores across time.

V. Results

A. The Effect of the FRP Regime on Schools' Test Scores

Table 2 presents our baseline estimates on the effects of higher retention due to the law change on average math and language standardized test scores. As implied by equation (1), in some specifications we include a set of time variant school-specific attributes to obtain the net impact of the FRP regime. All standard errors are clustered at the school level to ensure we account for potential serial correlation, as indicated in the difference-in-differences literature (Bertrand et al., 2004).

Our findings show that, across all years, the increase in retention does not have a meaningful economic effect on average math school performance. In contrast, we obtain a positive and strongly significant effect on language scores of 6.5% of a SD for a consecutive 2-year prior exposure to the FRP regime, and an average negative effect of 5.5% of a SD (Columns (4)-(5), first row) for being exposed to higher retention rates one year before the exam is taken. As treated schools increased retention 9 percentage points more than control schools, the effect implies an increase in language scores of 7% of a SD for a 10 percentage point rise in retention rates at 10th grade.

A plausible explanation for the positive effect attributed to retained students is that repeating a grade allows them to get a more thorough understanding of the material. Conversely, there may be several reasons for the negative effect on the non-retained students. First, the positive selection effect may be dominated by the negative influence of being in a group with a large fraction of lower performing peers after retention rates increased. Secondly, it may be that students started to allocate strategically more time and effort to study stem subjects when retention rates increased. Decreasing marginal productivity in both subjects implies that math scores do not increase as much as language scores decrease. This latter effect might be particularly relevant for students at the margin of repeating a grade.

B. Differences Between Retained and Non-Retained Students

To provide further evidence on these effects, we use individual level data on the age of the students to classify them as retained or non-retained. In the Colombian educational system, the nominal age of graduation is 17 years. Differences in age at the exam date can be explained mostly by having experienced grade retention (not necessarily at 10th grade only). Hence, we label students as "non-retained" if at the date of the exam they are 17 years old or younger. In contrast, we label students as "potentially retained" if they are 18 years old or older.[14] Then, we collapse the student and school datasets to create a unbalanced panel where the unit of observation is a school-exam year-retention status combination.[15] We estimate the baseline specification for both samples separately.

Table 3 presents our findings from the above exercise. Increased retention rates do not have any strongly significant impact on math scores for both types of students (columns (1)-(4)). For language scores the conclusions are different. In particular, while outcomes for non-retained students are still unaffected, potentially retained students' scores are affected. Estimates in columns (5) and (6) imply that being exposed to higher retention rates under the FRP regime one year (two years) before the exam causes a reduction (increase) of 5% (6%) of a SD on language scores. In general, this evidence supports the notion that i) the effect captured by the two-year prior exposure can be fully attributed to retained students and measures the impact of increased retention on test scores and, ii) the one-year prior exposure coefficient accounts for the effect of increased retention on marginal students that should have been retained but managed to "survive" the higher retention rates in the FRP regime.

C. On the Non-Linear Effects of Grade Retention

Table 4 documents the effects using the baseline specification implied by equation (1) with three dummy variables for all treatment groups of interest, i.e., low, middle and highly treated schools. In line with our basic specification, results from columns (1)-(2) suggest no significant effects of being exposed to the FRP period in years $t-1$ and $t-2$ on average math scores at year t. In contrast, we obtain significant effects for average language test scores which vary depending on the school's treatment classification (columns (3)-(4)).

Coefficients displayed in the first, third, and fifth rows suggest that middle treated schools score 13% of a SD lower relative to control schools because of a 6 percentage point rise in retention. The effects for all other treatment groups, even those not statistically significant, suggest a non-linear quadratic pattern. As all coefficients displayed have a negative sign, we believe there are confounding factors that are positively correlated with the probability of repeating a grade. These factors are masking the true detrimental impact of increased retention

[14] In Colombia, individuals are legally considered as adults if they are 18 years old or older.

[15] This process leads to have two different datasets. A dataset from non-retained students comprising of 35,103 observations corresponding to 6,235 schools, and a second dataset with information on potentially retained students with 34,534 observations from 6,237 schools.

on marginal students, especially of those enrolled in schools that retain at nearly optimal levels. By using the difference-in-differences estimator, we isolate the true effects from the bias attributed to such unobserved variation.

The increase in language scores of the retained students is significant both for the middle and highly treated groups. Interestingly, the gains on test scores appear to be similar for these groups. Dividing these results by the percentage points jump in retention rates corresponding to each group implies that the ATT coefficient of the highly treated group is smaller than that of the medium treated group. Considering that highly treated schools experienced a 12 percentage point increase in retention because of the FRP regime, our findings imply that a one percentage point rise in retention rates at 10th grade explains a 1.08% of a SD increase in test scores. The aforementioned impacts are larger for schools in the "medium treated" group. Namely, a 6 percentage points increase in retention rates at year $t-2$ implies a rise in test scores of about 13% of a SD in year t, so the effect is around 2.16% of a SD per one percentage point jump in retention. Hence, the same percentage point change in middle treated schools is two times as effective as it is in highly treated schools.

Our estimates suggest that retaining students is a strategy that exhibits decreasing marginal returns because there is a non-linear effect on language scores. At some point, higher retention is not expected to increase language performance. These students may, for instance, become demotivated because they must take the same classes again. For the marginal students, the results are only highly significant for the middle treated group which also shows the non-linear nature of the effects obtained.

The main conclusion from these findings is that there is a non-linear effect of retention, as schools that actively retain students do not necessarily benefit more from such a strategy, relative to other schools promoting more students. For the same percentage point increase in retention, middle treated schools obtain larger gains relative to schools that retain more students.

D. Effects along the Test Scores' Distribution

In Figure 4 we plot our difference-in-differences coefficients, this time considering test scores' percentiles by school as dependent variables, and using the above-median change in retention as the treatment classification criteria. In all these estimations we include school-specific covariates, although results barely change when the latter are excluded.

For math scores, we again obtain no effects for retained and marginal students across the entire test score distribution. On the contrary, results on language scores attributed to marginal pupils displayed in panel (c) show that the negative effect discussed earlier is strongest at the lower end of the distribution. For example, students performing at the 25th percentile in treated schools score up to 7% of a SD lower than comparable students at control schools. As average language scores for all students in the upper segment of the distribution were not affected by

retention, these findings suggest that the negative effect of being exposed to higher retention is more severe for lower performing but still non-retained students. We conjecture that this effect can be explained by the fact that students at the bottom of the ability distribution faced a higher threat of retention under the FRP regime, therefore compromising their test scores in the future.

Regarding language scores of retained students, the effect of higher retention is decreasing in students' test performance. Relative to students at the 10th percentile, students in treated schools scored 11% of a SD higher because of the FRP Regime. The effect is still significant but reduces in magnitude as performance increases. For instance, students in the 25th percentile score 8% of a SD more than students with comparable performance at control schools. The FRP regime appears to have no effect for those students performing at the median or above. Overall, these findings suggest that pupils at the lower end of the distribution benefited more from being in schools with increased retention rates due to repeating their coursework. This may also indicate that the benefits of retention are not linear. That is, being retained can be highly beneficial for underperforming students.

We perform the above exercise also for the specification in which we classify three treatment groups as already discussed. Table 5 presents the corresponding difference-in-differences coefficients for math and language scores. The table shows the results across the test scores' distributions distinguishing between low, medium and high treated schools. For math scores we obtain positive impacts at the upper end of the distribution that can be attributed to marginal students, although these effects are only significant at the 10% level. We do observe a significant negative effect for retained students on math scores of nine percent of a SD, suggesting that retained students' performance at the top of the distribution might be compromised as retention increases. In contrast, the positive effects on language scores for retained students, and the negative effects for the marginal students, are mostly relevant for middle treated schools. These effects are smaller in magnitude for the high treated group. This result supports our conjecture about the non-linear nature of the retention effects on test scores.

Several conclusions can be obtained from the estimates presented in this section. First, higher retention does not affect math scores, neither at the average nor at any below-median percentiles of the distribution. Second, higher retention positively (negatively) affects language scores for retained students (marginal students). Third, the fact that the results are stronger for middle treated schools suggests that the effect of retention is non-linear, as these schools obtain larger returns for the same percentage point increase in retention. Higher retention at some point no longer leads to higher scores for the retained students or lower scores for the students at the margin of repeating 10th grade. This also shows that our results are not driven by alternative reasons, such as selection. Finally, results obtained for both retained and marginal students are strongest at the lower end of the test score distribution. This indicates that low ability, retained students might benefit more from increased retention due to repeating classes or because the stigma of retention becomes of lower importance. Marginal students with similar ability score lower because they might face a higher threat of being retained, underperforming at the test

later on.

E. Testing the Common Trend Assumption

In this subsection, we present several robustness checks that support the empirical strategy implemented in this study. Table 6 presents difference-in-differences estimates of the common trends specification implied by equation (2), using exam year 2007 as a baseline. Appendix Table A.1 reports the same analysis for higher, middle, and low treated schools. In both tables we report the F-statistics of the joint test that the AUP-period coefficients are not statistically different from zero. As our estimates suggest, we can conclude that treated and control schools share a common trend that do not determine tests scores during the FRP regime. Regarding the placebo tests implied by equation (3), Table 7 report difference-in-differences coefficients assuming the FRP regime started in 2008 (Columns (1)-(4)) or 2007 (Columns (5)-(8)). The same estimates considering highly, middle, and low treated schools are reported in Appendix Table A.2. As the F-statistics indicate, we do not find evidence showing that pre-existing trends have a direct impact on the variation in test scores we observe after the retention policy changed.

Finally, we perform a falsification test using the subsample of control schools. From roughly 3000 schools in the control group, we select at random 1500 schools and assign them to the treatment group. Then, we estimate the model implied by our baseline specification (1). If we replicate this process say, 1000 times, we should expect to obtain significant results in no more than 50 replications using a 95% confidence level. Otherwise, results from this exercise will cast doubts on our treatment-control classification. Figure 5 displays the absolute t-statistic of each of these replications for our coefficients of interest, where the vertical red line denotes the 5% critical value of a t-student distribution (i.e. 1.96). We also present in Appendix Table A.3 the summary statistics of all parameters recovered from this falsification test. As observed, less than 5% of the replications turn out to be significant as only up to 28 replications are statistically different from zero. In addition, all mean coefficients are virtually zero, with standard errors at least 27 times higher than the reported effect. Overall, these results support the treatment-control categorization used in this paper.

F. Additional Robustness Checks

As indicated in the data section, there is attrition in our data for schools in which either retention rates or pre-FRP regime's characteristics are not completely observed throughout the period of interest. To analyze whether this attrition is selective, we report estimates in Table 9 using only schools from the seven-years balanced panel. As expected, attrition increases dramatically, leaving only 3,281 schools left to consider in the estimation. However, we observe that the signs of our estimates do not change. The effects become strongly significant and slightly higher in magnitude, giving strong support to our baseline findings. In fact, given the magnitudes

obtained from this robustness check, we can consider the coefficients provided in our baseline results as lower bound estimate of the true effect of the FRP regime on test scores.

Another concern in our empirical strategy is the timing between the announcement of the policy change and the time the new regime was officially in place. As discussed before, schools were informed in 2009 that from 2010 onwards they will be allowed to retain as many students as they prefer. It is plausible then that some schools reacted to this announcement by increasing retention rates in 2009. To check whether our results are robust to this behavior we repeat the estimations of our baseline specification, but excluding observations from exam year 2010. Table 10 reports difference-in-differences coefficients from this exercise. All coefficients are virtually the same as we obtain in our central findings, suggesting that schools' incentives to anticipate the policy change are not the main source of variation driving the effects we are documenting in this paper.

VI. Potential Mechanisms

In this section we explore propagation channels that may drive the effects we obtain. As implied by the FRP regime, treated schools significantly increased their retention rates, relative to schools in the control group. Are there any school characteristics that induce some institutions to retain more students? Are some school attributes amplifying the impacts of increased retention? To answer these questions, we assess the extent to which average class size at 10th grade, teachers' qualifications, and changes in the way teachers are remunerated play a role in disseminating the effects of grade retention.

There is a large consensus in the economics of education literature about the negative effects of large class sizes on students' academic performance (Angrist and Lavy, 1999; Fredriksson et al., 2012). Nonetheless, to our knowledge there is no discussion on whether grade retention and class size at school exhibit some complementarities. Assuming everything else constant, increased retention may have a positive impact on class size. We can also reverse the direction of the relationship. Schools with more students per group might have fewer incentives to retain students as large classrooms are more difficult to manage. Hence, we might expect the positive (negative) effects of retention to be weaker (stronger) on retained (marginal) students as the number of pupils per group rises.

Regarding our second transmission channel, recent papers highlight the empirical challenges of identifying the effects of teacher quality in the classroom (Rivkin et al., 2005; Gerritsen et al., 2017). We may expect the benefits (costs) of retention to be amplified (reduced) as teachers' education improves.

We exploit a regulation change in the way public school teachers are remunerated. From 2002 onwards, under Decree 1278, the remuneration, probation period, and screening process for newly hired teachers changed substantially. Under the new system, prospective teachers need to participate in a public entry contest which, after completion, will determine their start-

ing rank and wage. In addition, teachers hired under this new scheme will be subject to a probation period up to 12 months, to then receive tenure that can be revoked if subsequent performance evaluations are not satisfactory. In contrast, teachers hired before June 2002 were subject to the old 1979's, more lenient regulation (Decree 2277). This innovation in the employment relationship of teachers created a mixture of academic staff paid with the old and new pay scales. As it is expected that newly hired teachers will replace those about to retire, the proportion of teachers under the new pay scale at school is a key amplifying mechanism to study. Likewise, the direction of the effect is unclear. On the one hand, there is empirical evidence suggesting that teachers under the new regulation are positively selected, implying positive but moderate effects on school performance (Brutti and Sánchez, 2017). On the other hand, newly hired teachers may be inexperienced on identifying students with different educational needs. As career concerns are prevalent in their probation period, teachers might have incentives to focus their effort on students that are more likely to succeed. Thus, increasing participation of teachers under the new pay scale might have detrimental effects on students, especially those at the lower end of the ability distribution.

To test these mechanisms, we modify our baseline specification as follows:

$$\begin{aligned} Y_{st} &= \alpha_s + \delta_t + \sum_{h=1}^{2} \gamma_h [Treated_s \times FRP_{t-h}] + \sum_{h=1}^{2} \beta_h Attribute_{t-h} \\ &+ \sum_{h=1}^{2} \rho_h [Treated_s \times FRP_{t-h} \times Attribute_{t-h}] + \varepsilon_{st}, \end{aligned} \tag{4}$$

where the variable $Attribute_{t-h}$ denotes each mechanism we intend to test, one and two years before the exam takes place. The coefficients of interest in this specification are ρ_1 and ρ_2, which measure how each attribute in question propagates the effects of increased retention for marginal and retained students, respectively. We present our findings from this analysis in Table 11. Panel A reports the difference-in-differences coefficients. Panels B, C, and D show estimates on the interaction of the difference-in-differences effects with average school's class size at 10th grade, the proportion of teachers with a post-secondary education degree, and the proportion of teachers under the new, government regulated pay scale, respectively.

As observed in Columns (1)-(3) none of the mechanisms considered is masking the null effect of grade retention on math scores. We obtain a marginally significant effect of class size for retained students, but we claim that this can be ignored as it is only significant at the 10% level and very small in magnitude. In contrast, results for language scores indicate two propagation mechanisms worth to be discussed. First, we observe a negative effect of increased teachers' qualifications on test scores, as both coefficients of interest exhibit a negative sign. However, it seems that the effect in question is relevant (at the 10% significance level) only for retained students. These findings support the idea that policy interventions aimed to foster the human capital acquisiton of teachers may not be as effective as other measures to extract the

largest gains from grade retention.

Second, we observe the teachers' composition at treated schools to play a key role in propagating the effects of increased retention. In particular, a 10% jump in the proportion of newly hired teachers implies a rise (drop) in language scores of 2.5% (3.6%) of a SD for retained (marginally non-retained) students. To present these effects in more detail, in Figure 6 we plot the marginal effects of a 10 percentage points increment in the proportion of teachers hired under the new pay scale, one and two years before the exam takes place (panels (a) and (b), respectively). As observed, it is clear that benefits and costs of increased retention for language scores are monotonically increasing. This fact suggests that retained students benefit more from a young academic workforce that is willing to invest time and effort in their education, while marginally non-retained students are harmed since inexperienced teachers may be unable to target them accordingly.

VII. Concluding Remarks

This paper analyzes the effect of retention in 10th grade on school performance in grade 11. We exploit a law change in Colombia with respect to retention. Until 2010, schools were allowed to retain a maximum of 5 percent of their total number of students. After the abolishment of the law in 2010, schools were free to retain as many students as they considered appropriate. This led to a large increase in retention, with considerable heterogeneity across schools. We use a difference-in-differences analysis to study the effect of retention on test performance. Placebo tests suggest that there are common trends in scores among schools that responded in various degrees to the law change.

Our estimates reveal that there are positive effects of more retention on language test performance for retained students. These effects are non-linear, as modest increases in retention have positive effects but larger increases in retention do not necessarily lead to better performance. In addition, our findings suggest that non-retained students at the lower end of the ability distribution perform worse in language tests. Potential explanations for this effect include the negative spillover impacts from formerly retained students, the strategic substitution of effort between stem and non-stem subjects in order to avoid grade retention, and the changes in teachers' workforce composition. We provide evidence of the latter channel, with career concerned, inexperienced teachers failing to identify students at the margin of being retained as the main transmission mechanism. In contrast, we do not find any effects on math scores that can be attributed either to retained or non-retained pupils.

This research shows the importance of analyzing effects of retention at different margins of the ability distribution. Although data restrictions do not allow to recover information on individual retention, we feel confident that the empirical strategy and data construction implemented in this paper aids to solve this limitation by decomposing the effect of retention among different types of students. More research is needed to investigate whether the gains of

retention we identify can be outweighed by other costs of retention, such as school dropouts, career choice regret, delayed (or sudden) labor market participation, forgone income, and the formation of undesirable personality traits, preferences and risk attitudes across the life cycle.

References

Angrist, J. D. and V. Lavy (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics 114*(2), 533–575.

Bauernschuster, S., T. Hener, and H. Rainer (2016). Children of a (policy) revolution: The introduction of universal child care and its effect on fertility. *Journal of the European Economic Association 14*(4), 975–1005.

Belot, M. and V. Vandenberghe (2014). Evaluating the "threat" effects of grade repetition: Exploiting the 2001 reform by the french-speaking community of Belgium. *Education Economics 22*(1), 73–89.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust Differences-in-Differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Brutti, Z. and F. Sánchez (2017). Does better teacher selection lead to better students? evidence from a large scale reform in colombia. *Documento de trabajo CEDE* (11).

Eide, E. R. and M. H. Showalter (2001). The effect of grade retention on educational and labor market outcomes. *Economics of Education Review 20*(6), 563–576.

Fredriksson, P., B. Öckert, and H. Oosterbeek (2012). Long-term effects of class size. *The Quarterly Journal of Economics 128*(1), 249–285.

Fruehwirth, J. C., S. Navarro, and Y. Takahashi (2016). How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects. *Journal of Labor Economics 34*(4), 979–1021.

García-Pérez, J. I., M. Hidalgo-Hidalgo, and J. A. Robles-Zurita (2014). Does grade retention affect students' achievement? some evidence from spain. *Applied Economics 46*(12), 1373–1392.

Gerritsen, S., E. Plug, and D. Webbink (2017). Teacher quality and student achievement: Evidence from a sample of dutch twins. *Journal of Applied Econometrics 32*(3), 643–660.

Havnes, T. and M. Mogstad (2011). Money for nothing? Universal child care and maternal employment. *Journal of Public Economics 95*(11), 1455–1465.

Hill, A. J. (2014). The costs of failure: Negative externalities in high school course repetition. *Economics of Education Review 43*, 91–105.

Iregui, A. M., L. Melo, and J. Ramos (2006). La educación en colombia: Análisis del marco normativo y de los indicadores sectoriales. *Revista de Economía del Rosario 9*(2), 175–238.

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.

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., E. Carlson, M. Rotert, B. Egeland, and L. A. Sroufe (1997). A prospective, longitudinal study of the correlates and consequences of early grade retention. *Journal of School Psychology 35*(1), 3–25.

Jimerson, S. R. (1999). On the failure of failure: Examining the association between early grade retention and education and employment outcomes during late adolescence. *Journal of School Psychology 37*(3), 243–272.

Koppensteiner, M. F. (2014). Automatic grade promotion and student performance: Evidence from Brazil. *Journal of Development Economics 107*, 277–290.

Mahjoub, M.-B. (2017). The treatment effect of grade repetitions. *Education Economics 25*(4), 418–432.

Manacorda, M. (2012). The cost of grade retention. *Review of Economics and Statistics 94*(2), 596–606.

Martínez, G. and B. Herrera (2002). Finalidades y alcances del decreto 230 del 11 de febrero de 2002. *Ministerio de Educación Nacional. Bogotá*, 89–90.

McCoy, A. R. and A. J. Reynolds (1999). Grade retention and school performance:: An extended investigation. *Journal of School Psychology 37*(3), 273–298.

Namen, O. (2017). Does encouraging social promotion affect educational outcomes? Mimeo, Harris School of Public Policy, University of Chicago.

Pinzón, D. (2018). Reprobación y desempeño académico: Evidencia de la implementación de la promoción automática en colombia. Documento CEDE 016198, Universidad de los Andes, CEDE.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

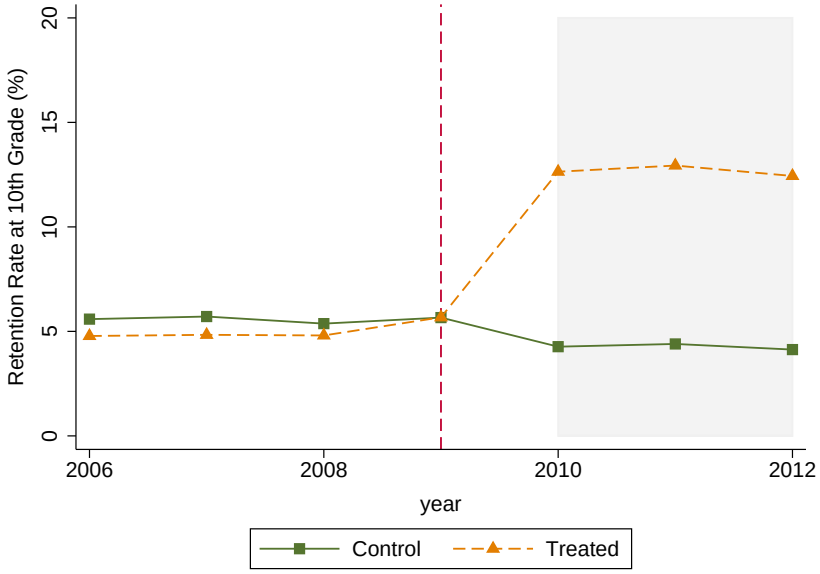
Silberglitt, B., J. J. Appleton, M. K. Burns, and S. R. Jimerson (2006). Examining the effects of grade retention on student reading performance: A longitudinal study. *Journal of School Psychology 44*(4), 255–270.

Stearns, E., S. Moller, J. Blau, and S. Potochnick (2007). Staying back and dropping out: The relationship between grade retention and school dropout. *Sociology of Education 80*(3), 210–240.

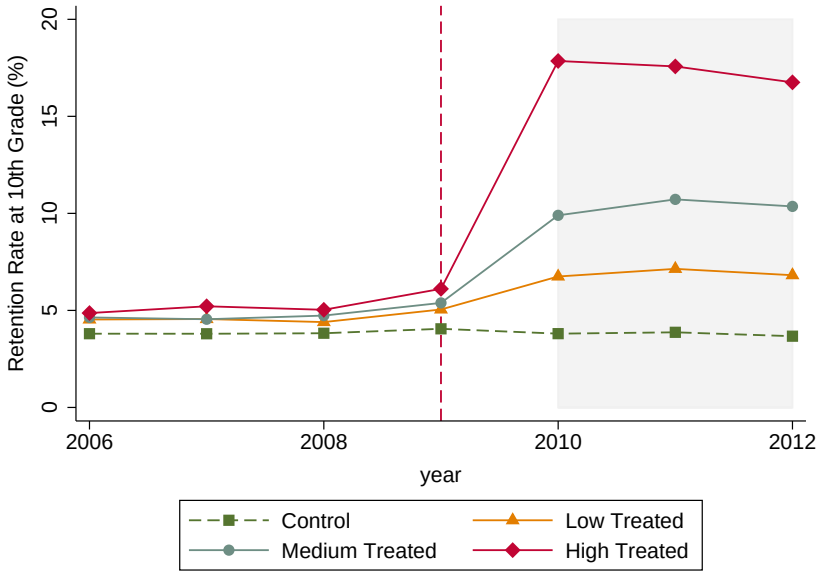
VIII. Tables and Figures

Figure 1: Retention Rates per Treatment Status

(a) Control vs. Treated Schools



(b) Control vs Treated Schools - Heterogeneous Treatment Status



Notes: Panel (a) displays retention rates at 10th grade (in percentage points) for treated and control schools. Treatment schools are defined as those with above -median increase in their retention rates at 10th grade from the automatic promotion years to the free retention years. Control schools are defined as those with below-median increase. Panel (b) shows retention rates at 10th grade for control and treated schools, for the different treatment definitions explained in the main text: Highly Treated ($HighTreated_s$), Medium Treated ($MiddleTreated_s$), and Low Treated ($LowTreated_s$). The dashed vertical line denotes year 2009 where schools were notified that the AUP regime will no longer hold. The gray area denotes the years where the FRP regime was in place.

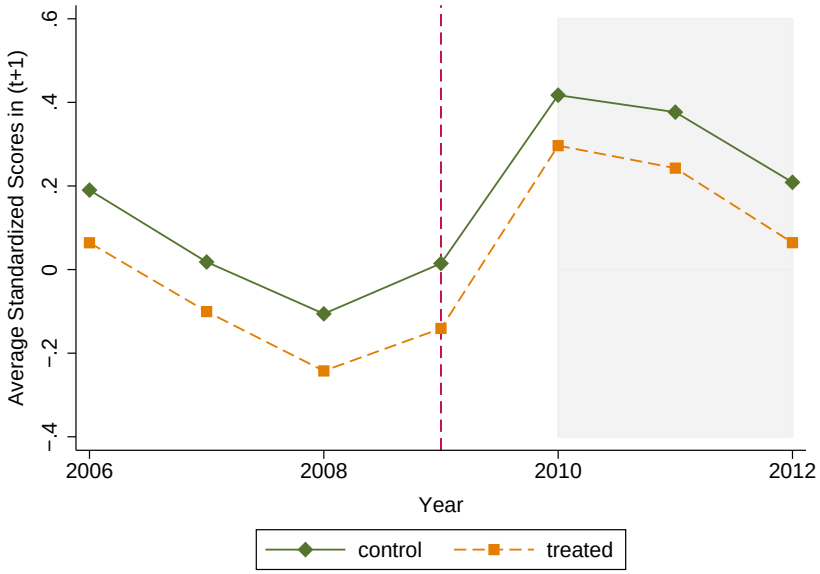
Table 1: Summary Statistics: Schools' Characteristics during AUP Regime

	Treated		Control		Both	
	# Schools	Mean	# Schools	Mean	Difference	s.e.
Average characteristics during AUP years (2007-2009)	(1)	(2)	(3)	(4)	(2)-(4)	
Socio-Demographic School's Composition						
Proportion of female students	3,142	0.528	3,106	0.527	0.001	0.004
Proportion of students from rural areas	3,142	0.342	3,106	0.328	0.014	0.010
Proportion of students from ethnic minorities	3,142	0.078	3,106	0.080	-0.002	0.005
Average age at exam date	3,142	17.628	3,106	17.598	0.030*	0.016
Proportion of students with an educated mother	3,142	0.149	3,106	0.224	-0.075***	0.006
Proportion of students above poverty classification	3,142	0.253	3,106	0.343	-0.090***	0.008
Public school	3,142	0.272	3,106	0.388	-0.117***	0.012
Working spell: 7:00 to 12:00	3,142	0.558	3,106	0.502	0.056***	0.012
Working spell: 13:00 to 18:00	3,142	0.170	3,106	0.110	0.061***	0.008
School type: academic and technical	3,142	0.625	3,106	0.644	-0.019	0.012
School type: pedagogical training	3,142	0.154	3,106	0.128	0.026***	0.009
School type: technical	3,142	0.020	3,106	0.014	0.006*	0.003
School-Related Attributes						
Average class size	3,142	0.201	3,106	0.214	-0.013	0.010
Average number of groups at 10th grade	3,142	32.272	3,106	29.491	2.781***	0.302
# of teachers with qualifications	3,142	2.367	3,106	1.839	0.528***	0.047
Proportion of teachers under the new pay scale	3,142	0.913	3,106	0.881	0.032***	0.003
Average number of managerial personnel	3,142	0.159	3,106	0.129	0.030***	0.005
Average number of support staff	3,142	3.204	3,106	2.837	0.367***	0.049
Average number of health personnel	3,142	0.768	3,106	0.806	-0.038	0.029
	3,142	0.166	3,106	0.239	-0.072***	0.015
Total Schools			6248			

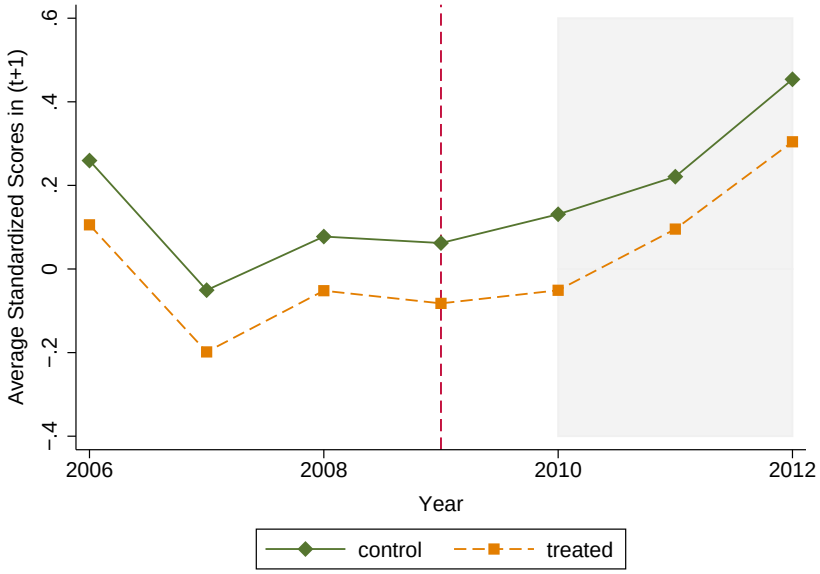
Notes: Data on socio-demographic composition of schools comes from the ICFES SABER11 dataset. Data on schools' attributes come from the administrative records of the C600 school made by the national statistics office (DANE). Treated and control schools are defined as in the text. *** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Figure 2: Test Scores by Treatment Status

(a) Math Scores

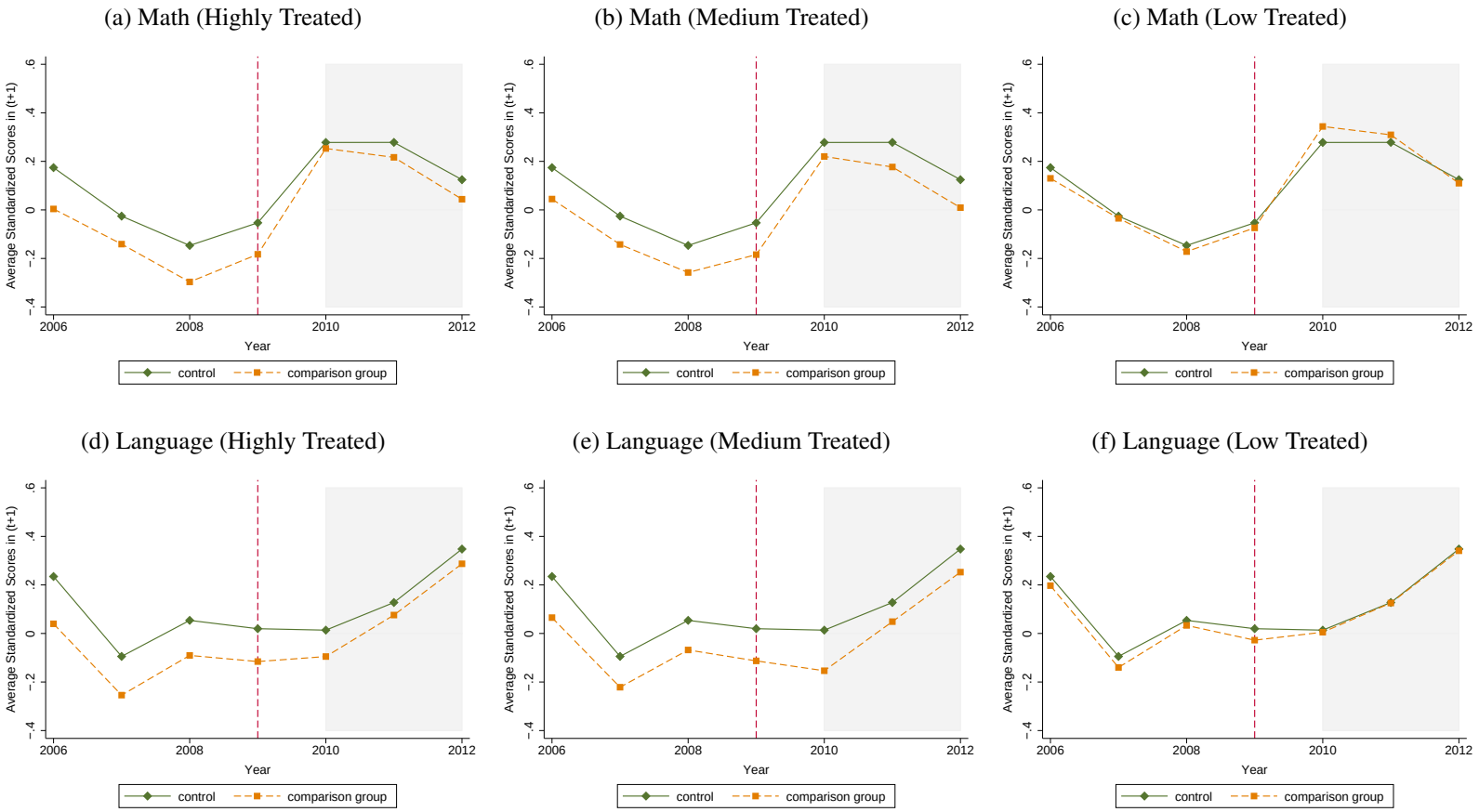


(b) Language Scores



Notes: This figure displays average tests scores in year $(t+1)$, for AUP and FRP years. Panel (a) presents common trends on average math scores between treated (dashed lines) and control (solid lines) schools. Panel (b) shows similar trends on average language test scores.

Figure 3: Test Scores by Treatment Status - Multiple Treatment Groups



Notes: This figure displays average test scores at year $(t+1)$ in the AUP and FRP years for the multiple treatment groups described in the text. Panels (a)-(c) show average math scores for controls schools (solid lines) and highly treated ($HighTreated_s$) medium treated ($MiddleTreated_s$) and low treated ($LowTreated_s$) schools. Panels (d)-(f) present the same figures for average language scores. The dashed vertical line denotes year 2009 where schools were notified that the AUP regime will no longer hold. The gray area denotes the years where the FRP regime was in place.

Table 2: Effect of the FRP Regime on Average Test Scores

	Math Scores		Language Scores	
	(1)	(2)	(4)	(5)
$Treated_s \times FRP_{t-1}$	0.003	0.015	-0.061**	-0.051*
	(0.020)	(0.020)	(0.027)	(0.027)
$Treated_s \times FRP_{t-2}$	0.002	0.006	0.064***	0.066***
	(0.018)	(0.018)	(0.023)	(0.023)
Observations	35,693	35,693	35,693	35,693
R-squared	0.149	0.155	0.094	0.101
# Schools	6,248	6,248	6,248	6,248
Covariates	No	Yes	No	Yes

Notes: Robust standard errors clustered at the school level reported in parentheses. All specifications include fixed effects by school and exam year. Treated (Control) schools are defined as those with above median (below median) increase in their retention rates at 10th grade from the AUP years to the FRP years. The outcome variables are average standardized SABER11 test scores for math and language subjects at year t. The coefficients of interest are the interaction of an indicator of treatment status with a set of dummy variables FRP_{t-1} and FRP_{t-2}, measuring the exposure to the FRP regime one and two years before the SABER11 exam is taken, respectively. Covariates considered in these estimations include the first two lags of: Average class size at 10th grade, number of health, support, and managerial (non-academic) staff per school, number of teachers with a professional degree, number of teachers under the new and old government-regulated pay scales, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.
*** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table 3: Effects of the FRP Regime on SABER11 Test Scores by Plausible Retention Status

	Math Scores				Language Scores			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Treated_s \times FRP_{t-1}$	0.002	0.006	0.017	0.023*	-0.049**	-0.044**	-0.005	0.002
	(0.016)	(0.016)	(0.013)	(0.013)	(0.021)	(0.021)	(0.017)	(0.017)
$Treated_s \times FRP_{t-2}$	-0.010	-0.009	0.002	0.005	0.057***	0.056***	0.015	0.016
	(0.016)	(0.017)	(0.013)	(0.013)	(0.021)	(0.021)	(0.016)	(0.016)
Observations	34,534	34,534	35,103	35,103	34,534	34,534	35,103	35,103
R-squared	0.039	0.041	0.128	0.132	0.050	0.053	0.069	0.073
# Schools	6,237	6,237	6,235	6,235	6,237	6,237	6,235	6,235
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
Potentially Retained	Yes	Yes	No	No	Yes	Yes	No	No

Notes: Robust standard errors clustered at the school level reported in parentheses. This table presents difference-in-differences estimates by partitioning the sample between potentially retained (aged 18 years or older) and non-retained (aged 17 years or younger). All specifications include fixed effects by school and exam year. Treated (Control) schools are defined as those with above median (below median) increase in their retention rates at 10th grade from the AUP years to the FRP years. The outcome variables are average standardized SABER11 test scores for math and language subjects at year t. The coefficients of interest are the interaction of an indicator of treatment status with a set of dummy variables FRP_{t-1} and FRP_{t-2}, measuring the exposure to the FRP regime one and two years before the SABER11 exam is taken, respectively. Covariates considered in these estimations include the first two lags of: Average class size at 10th grade, number of health, support, and managerial non-academic staff per school, number of teachers with a professional degree, number of teachers under the new and old government-regulated pay scales, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.
*** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

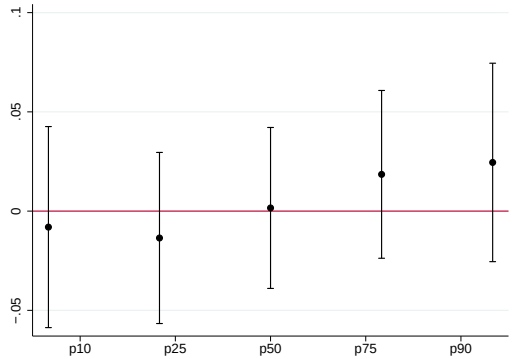
Table 4: Effect of the FRP Regime on Average Test Scores Per Multiple Treatment Status

	Math Scores		Language Scores	
	(1)	(2)	(4)	(5)
$HighTreated_s \times FRP_{t-1}$	0.037	0.053	-0.093**	-0.081*
	(0.033)	(0.033)	(0.043)	(0.042)
$HighTreated_s \times FRP_{t-2}$	-0.007	-0.003	0.124***	0.127***
	(0.030)	(0.031)	(0.037)	(0.037)
$MiddleTreated_s \times FRP_{t-1}$	-0.015	-0.003	-0.138***	-0.127***
	(0.032)	(0.032)	(0.043)	(0.043)
$MiddleTreated_s \times FRP_{t-2}$	-0.017	-0.013	0.127***	0.128***
	(0.030)	(0.030)	(0.037)	(0.037)
$LowTreated_s \times FRP_{t-1}$	0.020	0.027	-0.077*	-0.072
	(0.033)	(0.032)	(0.046)	(0.045)
$LowTreated_s \times FRP_{t-2}$	-0.014	-0.012	0.061	0.062
	(0.030)	(0.031)	(0.039)	(0.039)
Observations	35,693	35,693	35,693	35,693
R-squared	0.149	0.156	0.095	0.102
# Schools	6,248	6,248	6,248	6,248
Covariates	No	Yes	No	Yes

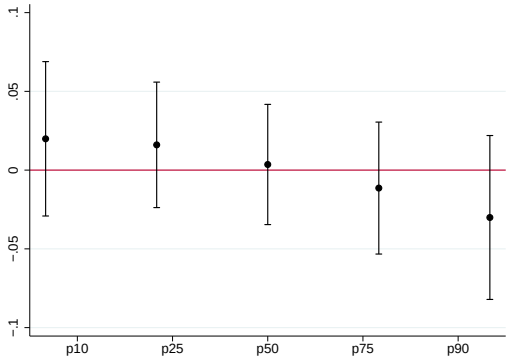
Notes: Robust standard errors clustered at the school level reported in parentheses. Treated schools are defined as in the main text. The outcome variables are average standardized SABER11 test scores for math and language subjects at year t. The coefficients of interest are the interaction of an indicator of treatment status with a set of dummy variables FRP_{t-1} and FRP_{t-2}, measuring the exposure to the FRP regime one and two years before the SABER11 exam is taken, respectively. Covariates considered in these estimations include the first two lags of: Average class size at 10th grade, number of health, support, and managerial non-academic staff per school, number of teachers with a professional degree, number of teachers under the new and old government-regulated pay scales, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.
*** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Figure 4: Effect of the FRP Regime on SABER11 Test Scores' Distribution

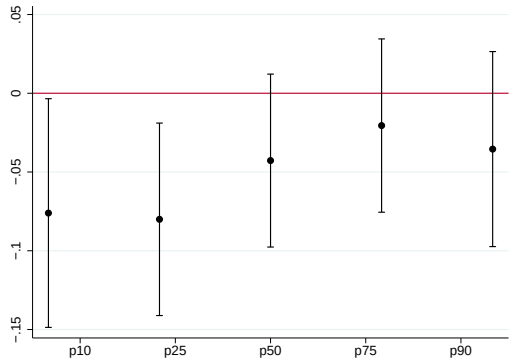
(a) 1-year prior exposure (Math Scores)



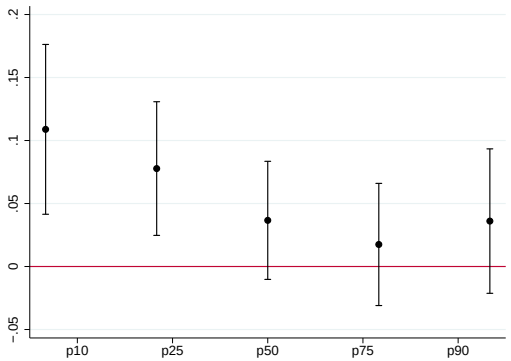
(b) 2-years prior exposure (Math Scores)



(c) 1-year prior exposure (Language Scores)



(d) 2-years prior exposure (Language Scores)



Notes: This figure displays the difference-in-differences effect of the FRP regime on the percentiles of the SABER11 test scores' distribution for both math and language subjects. Panels (a)-(b) present the effect of one and two years prior exposure to the FRP regime for math scores. Panels (c)-(d) plots the same effects for language scores. The caps denote confidence intervals at the 95% significance level.

Table 5: Effect of the FRP Regime on SABER11 Test Scores' Distribution by Multiple Treatment Status

	Math Scores					Language Scores				
	(2) 10th Percentile	(3) 25th Percentile	(4) 50th Percentile	(5) 75th Percentile	(6) 90th Percentile	(8) 10th Percentile	(9) 25th Percentile	(10) 50th Percentile	(11) 75th Percentile	(12) 90th Percentile
$HighTreated_s \times FRP_{t-1}$	0.060	0.039	0.022	0.069*	0.082*	-0.075	-0.087*	-0.095**	-0.049	-0.046
	(0.045)	(0.038)	(0.036)	(0.038)	(0.045)	(0.063)	(0.051)	(0.045)	(0.046)	(0.053)
$HighTreated_s \times FRP_{t-2}$	0.000	0.012	0.011	-0.006	-0.032	0.140**	0.118**	0.132***	0.095**	0.096*
	(0.044)	(0.036)	(0.034)	(0.037)	(0.046)	(0.059)	(0.046)	(0.041)	(0.042)	(0.051)
$MiddleTreated_s \times FRP_{t-1}$	-0.017	-0.017	-0.015	-0.007	0.056	-0.153**	-0.152***	-0.119***	-0.060	-0.089*
	(0.045)	(0.037)	(0.035)	(0.037)	(0.046)	(0.064)	(0.052)	(0.046)	(0.046)	(0.054)
$MiddleTreated_s \times FRP_{t-2}$	0.016	0.015	-0.000	-0.013	-0.093**	0.187***	0.135***	0.113***	0.074*	0.091*
	(0.044)	(0.036)	(0.034)	(0.037)	(0.046)	(0.060)	(0.047)	(0.041)	(0.042)	(0.051)
$LowTreated_s \times FRP_{t-1}$	0.037	0.020	0.028	0.025	0.050	-0.061	-0.044	-0.085*	-0.046	-0.112**
	(0.046)	(0.038)	(0.036)	(0.038)	(0.046)	(0.065)	(0.054)	(0.047)	(0.048)	(0.055)
$LowTreated_s \times FRP_{t-2}$	-0.001	-0.003	-0.019	-0.009	-0.032	0.089	0.019	0.074*	0.050	0.090*
	(0.046)	(0.036)	(0.035)	(0.039)	(0.048)	(0.061)	(0.048)	(0.042)	(0.043)	(0.052)
Observations	35,693	35,693	35,693	35,693	35,693	35,693	35,693	35,693	35,693	35,693
R-squared	0.022	0.052	0.127	0.223	0.253	0.270	0.163	0.075	0.174	0.299
# Schools	6,248	6,248	6,248	6,248	6,248	6,248	6,248	6,248	6,248	6,248
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Robust standard errors clustered at the school level reported in parentheses. All specifications include fixed effects by school and exam year. Treated schools are defined as in the main text. The outcome variables are schools' different percentiles of the standardized SABER11 test scores for math and language subjects at year t. The coefficients of interest are the interaction of an indicator of treatment status with a set of dummy variables FRP_{t-1} and FRP_{t-2}, measuring the exposure to the FRP regime one and two years before the SABER11 exam is taken, respectively.
*** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table 6: Common Trends Assumption Test

	Math Scores		Language Scores	
	(1)	(2)	(3)	(4)
Pre-FRP trends				
$Treated_s \times$1[year=2008]	0.011	0.017	0.025	0.029
	(0.019)	(0.019)	(0.018)	(0.018)
$Treated_s \times$1[year=2009]	-0.021	-0.013	0.030	0.036*
	(0.021)	(0.021)	(0.020)	(0.020)
$Treated_s \times$1[year=2010]	-0.050**	-0.038	0.008	0.015
	(0.024)	(0.024)	(0.019)	(0.020)
FRP trends				
$Treated_s \times$1[year=2011]	-0.014	0.004	-0.045	-0.031
	(0.026)	(0.026)	(0.028)	(0.028)
$Treated_s \times$1[year=2012]	-0.017	0.004	0.021	0.037*
	(0.024)	(0.023)	(0.019)	(0.019)
$Treated_s \times$1[year=2013]	-0.005	0.018	0.016	0.033*
	(0.022)	(0.022)	(0.019)	(0.019)
Observations	35,693	35,693	35,693	35,693
R-squared	0.149	0.155	0.094	0.101
# Schools	6,248	6,248	6,248	6,248
Covariates	No	Yes	No	Yes
F-stat (3; 6,248)	2.541	2.043	1.087	1.368
p-value	0.0546	0.106	0.353	0.251

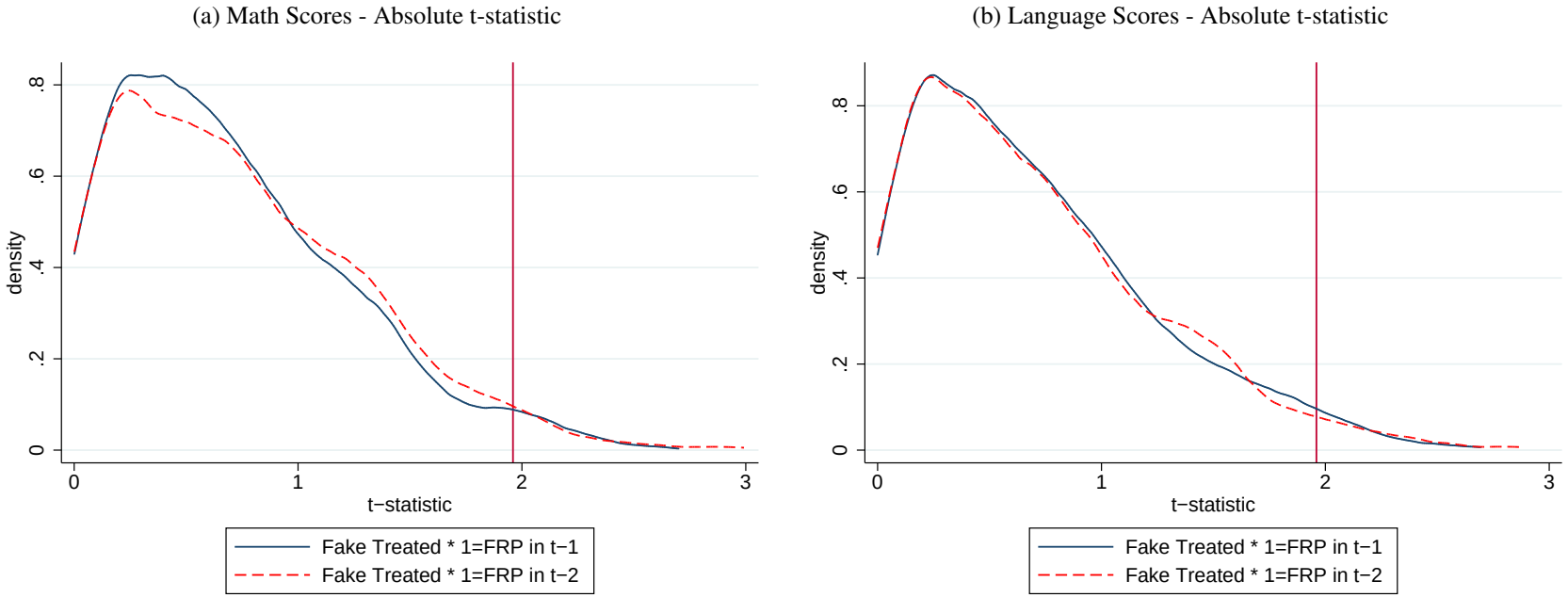
Notes: Robust standard errors clustered at the school level are reported in parentheses. This table shows results for the common trend assumption test. The outcome variables are average standardized test scores for math and language subjects measured for year t. Covariates include the first two lags of: average class size at 10th grade, average managerial, health, and support staff per school, average number of teachers under the old and new pay scales, average number of teachers with a professional degree, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree. F-statistics reported correspond to the null hypothesis that pre-FRP regime trends differences between control and treated schools are not statistically significant. Treated and controls schools are defined as in the main text.
*** p-value < 0.01, ** p-value < 0.05. * p-value < 0.1.

Table 7: Placebo Tests: FRP Regime Started Before Original Date

	Math Scores		Language Scores		Math Scores		Language Scores	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Treated_s \times FakeFRP_{2008,t-1}$	-0.026	-0.021	0.018	0.021				
	(0.018)	(0.018)	(0.017)	(0.017)				
$Treated_s \times FakeFRP_{2008,t-2}$	-0.001	0.009	-0.030	-0.023				
	(0.017)	(0.017)	(0.019)	(0.018)				
$Treated_s \times FakeFRP_{2007,t-1}$					0.011	0.018	0.024	0.028
					(0.019)	(0.019)	(0.018)	(0.018)
$Treated_s \times FakeFRP_{2007,t-1}$					-0.032*	-0.022	-0.018	-0.011
					(0.019)	(0.018)	(0.017)	(0.016)
Observations	35,693	35,693	35,693	35,693	35,693	35,693	35,693	35,693
R-squared	0.149	0.155	0.094	0.101	0.149	0.155	0.094	0.101
# Schools	6,248	6,248	6,248	6,248	6,248	6,248	6,248	6,248
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
F-stats (2; 6,248)	1.390	0.692	1.290	0.972	1.520	0.817	0.960	1.233
p-value	0.249	0.500	0.275	0.378	0.219	0.442	0.383	0.291

Notes: Robust standard errors clustered at the school level are reported in parentheses. All specifications include fixed effects by school and exam year. Treated (Control) schools are defined as those with above median (below median) increase in their retention rates at 10th grade from the AUP years to the FRP years. The dependent variables are average standardized math and language SABER11's test scores. Columns (1)-(4) report results on the placebo test assuming the FRP regime started in 2008. Columns (5)-(8) show estimates on the placebo test assuming the FRP regime started in 2007. F-statistics reported correspond to the joint test of the null hypothesis that placebo effects are not different from zero. Covariates include the first two lags of: average class size at 10th grade, average managerial, health, and support staff per school, average number of teachers under the old and new pay scales, average number of teachers with a professional degree, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.
*** p-value < 0.01, ** p-value < 0.05. * p-value < 0.1.

Figure 5: Falsification Test - 1500 Control Schools as Treated



Notes: Based on 1000 replications. This figure displays absolute t-statistics of the falsification test. We estimate the baseline specification (1) using the control group sample only, but we randomly allocate the treatment status to 1500 control schools. All estimations include fixed effects by school and exam year. The outcome variables are mean standardized SABER11 test scores for math and language subjects at year t. The coefficients of interest are the interaction of an indicator of (false) treatment status with a set of dummy variables FRP_{t-1} and FRP_{t-2}, measuring the exposure to the FRP regime one and two years before the SABER11 exam is taken, respectively. The red vertical line denotes the critical value by which the null hypothesis of non-significance is rejected at the 5% level.

Table 8: Effect of the FRP Regime on Average Test Scores - Balanced Panel Estimations

	Math Scores		Language Scores	
	(1)	(2)	(4)	(5)
$Treated_s \times FRP_{t-1}$	-0.021	-0.007	-0.091***	-0.079**
	(0.023)	(0.022)	(0.032)	(0.031)
$Treated_s \times FRP_{t-2}$	0.019	0.024	0.076***	0.079***
	(0.021)	(0.021)	(0.027)	(0.027)
Observations	22,967	22,967	22,967	22,967
R-squared	0.176	0.184	0.117	0.126
# Schools	3,281	3,281	3,281	3,281
Covariates	No	yes	No	Yes

Notes: Robust standard errors clustered at the school level reported in parentheses. This table reports difference-in-difference estimates when we consider only those schools from a balanced panel dataset of seven years. All specifications include fixed effects by school and exam year. Treated (Control) schools are defined as those with above median (below median) increase in their retention rates at 10th grade from the AUP years to the FRP years. The outcome variables are average standardized SABER11 test scores for math and language subjects at year t. The coefficients of interest are the interaction of an indicator of treatment status with a set of dummy variables FRP_{t-1} and FRP_{t-2}, measuring the exposure to the FRP regime one and two years before the SABER11 exam is taken, respectively. Covariates considered in these estimations include the first two lags of: Average class size at 10th grade, number of health, support, and managerial non-academic staff per school, number of teachers with a professional degree, number of teachers under the new and old government-regulated pay scales, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.
*** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Table 9: Effect of the FRP Regime on Average Test Scores - Excluding Anticipatory Effects

	Math Scores		Language Scores	
	(1)	(2)	(4)	(5)
$Treated_s \times FRP_{t-1}$	-0.012	0.002	-0.068**	-0.056**
	(0.022)	(0.022)	(0.028)	(0.028)
$Treated_s \times FRP_{t-2}$	0.006	0.010	0.066***	0.068***
	(0.019)	(0.019)	(0.023)	(0.023)
Observations	30,576	30,576	30,576	30,576
R-squared	0.158	0.166	0.105	0.112
# Schools	6,234	6,234	6,234	6,234
Covariates	No	Yes	No	Yes

Notes: Robust standard errors clustered at the school level reported in parentheses. This table reports difference-in-differences regressions excluding observations from exam year 2010. All specifications include fixed effects by school and exam year. Treated (Control) schools are defined as those with above median (below median) increase in their retention rates at 10th grade from the AUP years to the FRP years. The outcome variables are average standardized SABER11 test scores for math and language subjects at year t. The coefficients of interest are the interaction of an indicator of treatment status with a set of dummy variables FRP_{t-1} and FRP_{t-2}, measuring the exposure to the FRP regime one and two years before the SABER11 exam is taken, respectively. Covariates considered in these estimations include the first two lags of: Average class size at 10th grade, number of health, support, and managerial non-academic staff per school, number of teachers with a professional degree, number of teachers under the new and old government-regulated pay scales, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.
*** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

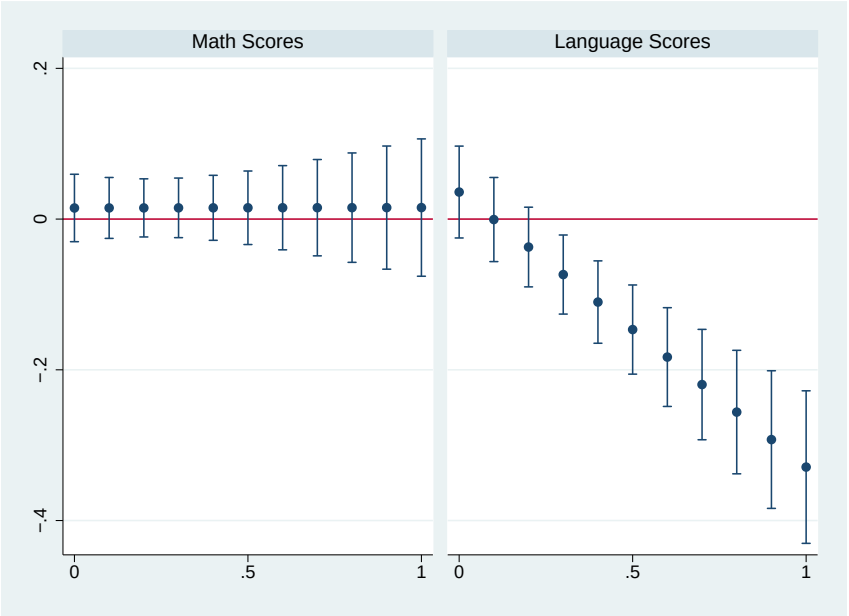
Table 10: Effect of the FRP Regimen on SABER11 Test Scores: Mechanisms

	Math Scores			Language Scores		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Difference in Differences Effect						
$DID_{s,t-1} = Treated_s \times FRP_{t-1}$	0.006	0.073	0.015	-0.079**	0.063	0.036
	(0.029)	(0.067)	(0.023)	(0.036)	(0.078)	(0.031)
$DID_{s,t-2} = Treated_s \times FRP_{t-2}$	0.036	-0.006	0.002	0.051*	0.155***	0.017
	(0.027)	(0.058)	(0.021)	(0.030)	(0.056)	(0.025)
Panel B: Effects by Class Size						
$DID_{s,t-1} \times ClassSize_{s,t-1}$	0.000			0.001		
	(0.001)			(0.001)		
$DID_{s,t-2} \times ClassSize_{s,t-2}$	-0.001*			0.000		
	(0.001)			(0.000)		
Panel C: Effects by Teacher's Qualifications						
$DID_{s,t-1} \times TeachersQual_{s,t-1}$		-0.064			-0.125	
		(0.068)			(0.077)	
$DID_{s,t-2} \times TeachersQual_{s,t-2}$		0.015			-0.093*	
		(0.058)			(0.054)	
Panel D: Effects by Teacher's Pay Scale						
$DID_{s,t-1} \times TeachersNewpay_{s,t-1}$			0.001			-0.365***
			(0.054)			(0.060)
$DID_{s,t-2} \times TeachersNewpay_{s,t-2}$			0.015			0.249***
			(0.047)			(0.050)
Observations	35,693	35,693	35,693	35,693	35,693	35,693
R-squared	0.155	0.155	0.155	0.101	0.102	0.104
# Schools	6,248	6,248	6,248	6,248	6,248	6,248

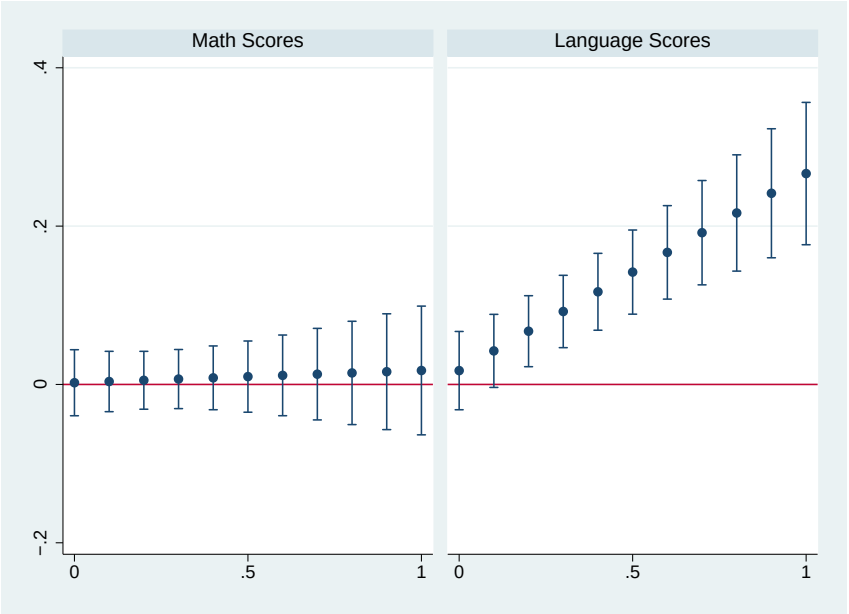
Notes: Robust standard errors clustered at the school level reported in parentheses. This table presents difference-in-differences estimates and their interactions with different school attributes. All specifications include fixed effects by school and exam year. Treated (Control) schools are defined as those with above median (below median) increase in their retention rates at 10th grade from the AUP years to the FRP years. The outcome variables are average standardized SABER11 test scores for math and language subjects at year t. *ClassSize* measures the average number of students per group at 10th grade. *TeachersQual* accounts for the proportion of teachers with a post-secondary education degree per school. *TeachersNewpay* measures the proportion of teachers per school under the new pay scale regulated by the central government.
*** p-value < 0.01, ** p-value < 0.05, * p-value < 0.1.

Figure 6: Mechanisms: Variation in Teachers' Composition

(a) 1-year prior exposure



(b) 2-years prior exposure

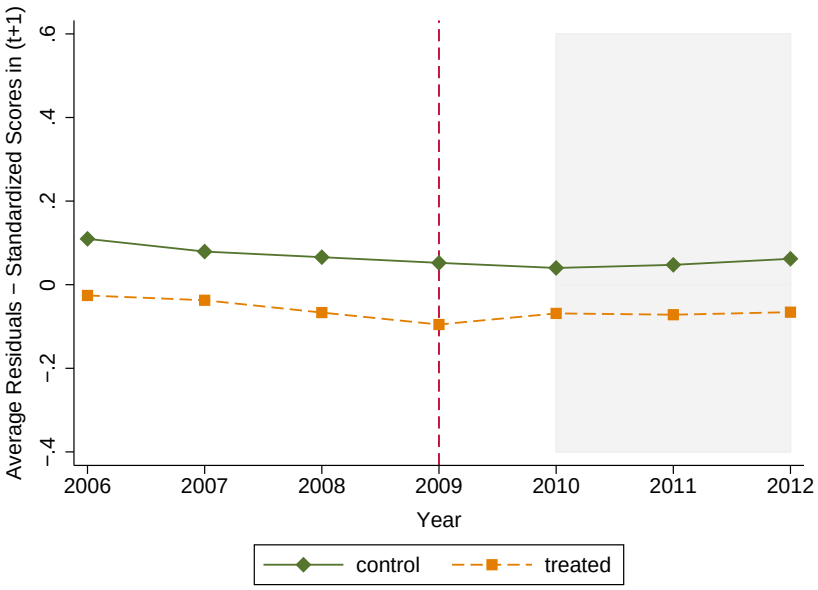


Notes: This figure plots the marginal effects from a difference-in-differences estimation, interacted with the proportion of teachers under the new pay scale. each cap denotes confidence intervals at the 95% level of a 10-percentage points increase in the proportion of teachers under the new pay scale.

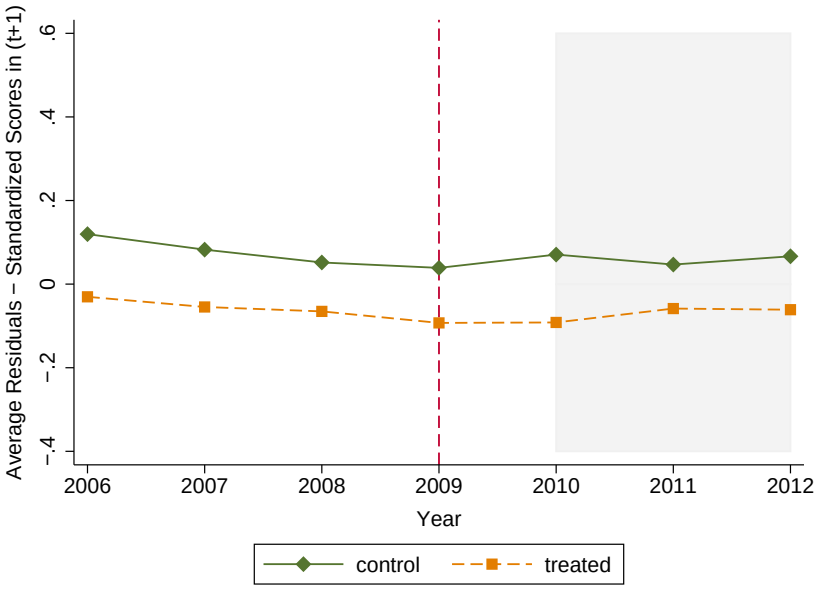
IX. Appendix Tables and Figures (Not for Publication)

Figure A.1: Test Scores Residuals by Treatment Status

(a) Math Scores

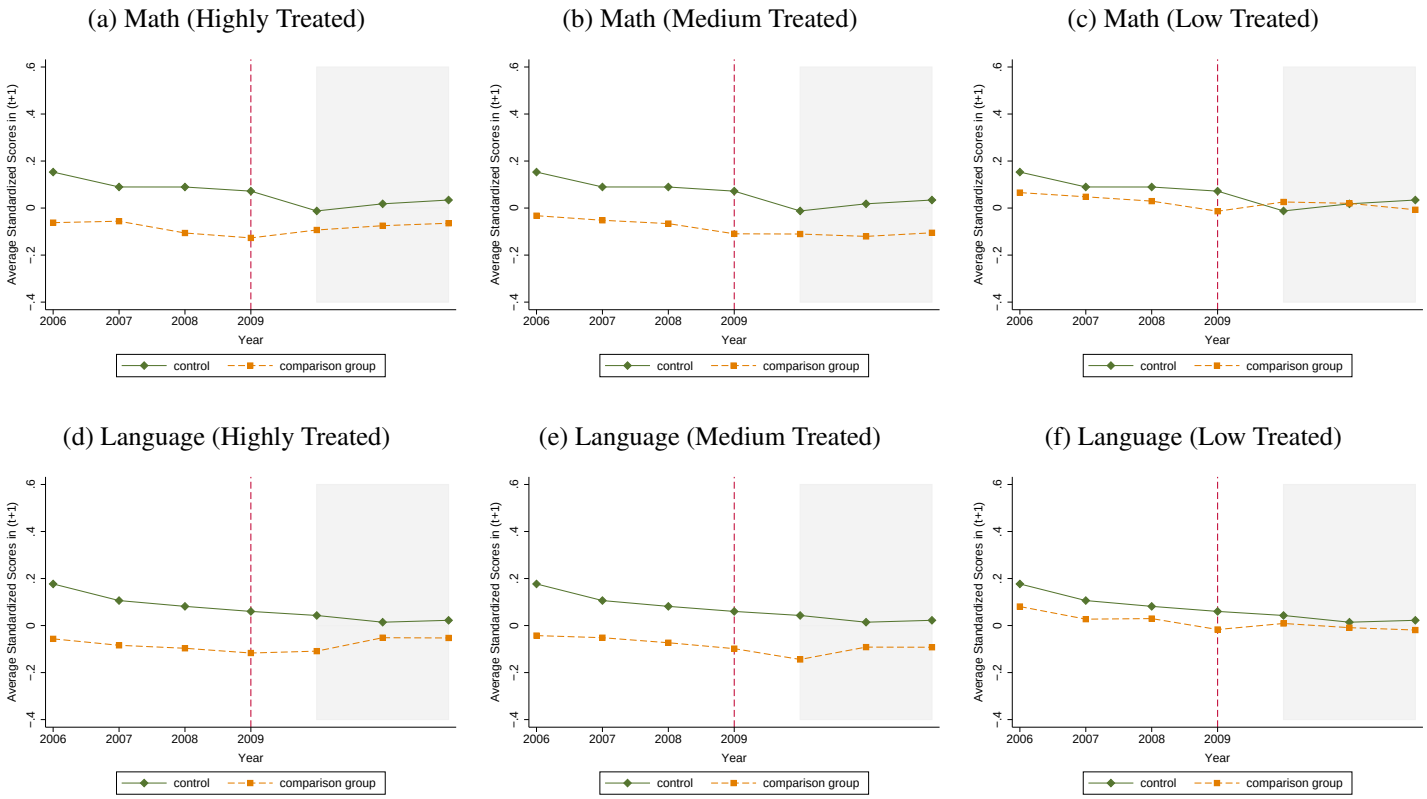


(b) Language Scores



Notes: This figure displays average residuals from an OLS regression where the dependent variable is the average tests scores in year $(t+1)$ as a function of school and exam-year fixed effects, and a set of school specific covariates. Panel (a) presents common trends on math scores' average residuals between treated (dashed lines) and control (solid lines) schools. Panel (b) shows similar trends on language test scores' average residuals. Covariates considered in these estimations include the first two lags of: Average class size at 10th grade, number of health, support, and managerial non-academic staff per school, number of teachers with a professional degree, number of teachers under the new and old government-regulated pay scales, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.

Figure A.2: Test Scores Residuals by Treatment Status - Multiple Treatment Groups



Notes: This figure displays average residuals from an OLS regression where the dependent variable is the average tests scores in year $(t+1)$ as a function of school and exam-year fixed effects, and a set of school's specific covariates. Panels (a)-(c) show math scores' average residuals for controls schools (solid lines) and highly treated ($HighTreated_s$) medium treated ($MiddleTreated_s$) and low treated ($LowTreated_s$) schools. The dashed vertical line denotes year 2009 where schools were notified that the AUP regime will no longer hold. The gray area denotes 2010-2012 years where the FRP regime was in place. Panels (d)-(f) present the same figures for language scores' average residuals. Covariates considered in these estimations include the first two lags of: Average class size at 10th grade, number of health, support, and managerial non-academic staff per school, number of teachers with a professional degree, number of teachers under the new and old government-regulated pay scales, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.

Table A.1: Common Trend Assumption Test by Treatment Heterogeneity Status

	Math Scores		Language Scores	
	(1)	(2)	(3)	(4)
$HighTreated_s \times 1$[year=2008]	0.034	0.044	0.032	0.040
	(0.030)	(0.030)	(0.030)	(0.030)
$HighTreated_s \times 1$[year=2009]	-0.021	-0.008	0.046	0.057*
	(0.033)	(0.033)	(0.031)	(0.031)
$HighTreated_s \times 1$[year=2010]	-0.047	-0.029	0.027	0.039
	(0.037)	(0.037)	(0.031)	(0.031)
$MiddleTreated_s \times 1$[year=2008]	0.003	0.011	0.032	0.038
	(0.032)	(0.032)	(0.031)	(0.031)
$MiddleTreated_s \times 1$[year=2009]	-0.028	-0.018	0.029	0.038
	(0.035)	(0.035)	(0.033)	(0.033)
$MiddleTreated_s \times 1$[year=2010]	-0.080**	-0.066*	-0.002	0.008
	(0.039)	(0.039)	(0.032)	(0.032)
$LowTreated_s \times 1$[year=2008]	0.000	0.006	-0.025	-0.021
	(0.031)	(0.031)	(0.030)	(0.031)
$LowTreated_s \times 1$[year=2009]	-0.013	-0.003	0.024	0.031
	(0.035)	(0.035)	(0.032)	(0.033)
$LowTreated_s \times 1$[year=2010]	-0.047	-0.035	-0.012	-0.002
	(0.037)	(0.037)	(0.032)	(0.033)
Observations	35,693	35,693	35,693	35,693
R-squared	0.150	0.156	0.096	0.103
# Schools	6,248	6,248	6,248	6,248
Covariates	No	Yes	No	Yes
F-stat (9 , 6614)	1.051	0.968	1.064	1.220
p-value	0.397	0.464	0.386	0.277

Notes: Robust standard errors clustered at the school level are reported in parentheses. This table shows results for the common trend assumption test. All specifications include fixed effects by school and exam year. For matters of space, we only report the coefficients from the pre-FRP trends years. The outcome variables are average standardized test scores for math and language subjects measured for year t. Covariates include the first two lags of: average class size at 10th grade, average managerial, health, and support staff per school, average number of teachers under the old and new pay scales, average number of teachers with a professional degree, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree. F-statistics reported correspond to the null hypothesis that pre-FRP regime trends differences between control and treated schools are not statistically significant. Treated and controls schools are defined as in the main text.
*** p-value < 0.01, ** p-value < 0.05. * p-value < 0.1.

Table A.2: Placebo Test Using Multiple Treatment Groups: FRP Regime Started Before 2010

	Math Scores		Language Scores		Math Scores		Language Scores	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: FRP started in 2008								
$HighTreated_s \times FakeFRP_{2008,t-1}$	-0.037	-0.029	0.030	0.037				
	(0.028)	(0.028)	(0.026)	(0.026)				
$HighTreated_s \times FakeFRP_{2008,t-2}$	0.025	0.038	-0.027	-0.018				
	(0.028)	(0.028)	(0.029)	(0.028)				
$MiddleTreated_s \times FakeFRP_{2008,t-1}$	-0.029	-0.023	0.013	0.018				
	(0.029)	(0.029)	(0.027)	(0.027)				
$MiddleTreated_s \times FakeFRP_{2008,t-2}$	-0.034	-0.024	-0.058**	-0.051*				
	(0.027)	(0.027)	(0.029)	(0.029)				
$LowTreated_s \times FakeFRP_{2008,t-1}$	-0.012	-0.006	0.036	0.041				
	(0.030)	(0.030)	(0.027)	(0.027)				
$LowTreated_s \times FakeFRP_{2008,t-2}$	-0.004	0.001	-0.056*	-0.053*				
	(0.028)	(0.028)	(0.030)	(0.030)				
Panel B: FRP started in 2007								
$HighTreated_s \times FakeFRP_{2007,t-1}$					0.035	0.046	0.031	0.040
					(0.030)	(0.030)	(0.030)	(0.030)
$HighTreated_s \times FakeFRP_{2007,t-2}$					-0.035	-0.022	-0.007	0.002
					(0.030)	(0.030)	(0.027)	(0.026)
$MiddleTreated_s \times FakeFRP_{2007,t-1}$					0.003	0.010	0.031	0.037
					(0.032)	(0.032)	(0.031)	(0.031)
$MiddleTreated_s \times FakeFRP_{2007,t-2}$					-0.057*	-0.047	-0.049*	-0.041
					(0.030)	(0.030)	(0.027)	(0.026)
$LowTreated_s \times FakeFRP_{2007,t-1}$					0.001	0.006	-0.026	-0.022
					(0.031)	(0.031)	(0.030)	(0.031)
$LowTreated_s \times FakeFRP_{2007,t-2}$					-0.016	-0.008	0.004	0.010
					(0.031)	(0.031)	(0.027)	(0.027)
Observations	35,693	35,693	35,693	35,693	35,693	35,693	35,693	35,693
R-squared	0.149	0.155	0.094	0.101	0.149	0.155	0.094	0.101
# Schools	6,248	6,248	6,248	6,248	6,248	6,248	6,248	6,248
Covariates	No	Yes	No	Yes	No	Yes	No	Yes
F-stats (6; 6,614)	1.535	1.462	1.456	1.478	1.181	1.146	1.808	2.015
p-value	0.162	0.187	0.189	0.181	0.313	0.333	0.0934	0.0602

Notes: Robust standard errors clustered at the school level are reported in parentheses. This table presents results on placebo tests for the baseline difference-in-differences specification, assuming the FRP started in year 2008. All specifications include fixed effects by school and exam year. Treated (Control) schools are defined as those with above median (below median) increase in their retention rates at 10th grade from the AUP years to the FRP years. The dependent variables are average standardized math and language SABER11's test scores. F-statistics reported correspond to the joint test of the null hypothesis that placebo effects are not different from zero. Covariates include the first two lags of: average class size at 10th grade, average managerial, health, and support staff per school, average number of teachers under the old and new pay scales, average number of teachers with a professional degree, proportion of teachers under the new pay scale, and proportion of teachers with a professional degree.
*** p-value < 0.01, ** p-value < 0.05. * p-value < 0.1.

Table A.3: Falsification Test - 1500 Control Schools as Treated

	Math Scores				Language Scores			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	mean	std. dev.	min	max	mean	std. dev.	min	max
$Treated_s \times FRP_{t-1}$								
Coefficient	0.001	0.026	-0.067	0.079	0.001	0.035	-0.108	0.100
Std. Error	0.029	0.000	0.029	0.029	0.040	0.000	0.040	0.040
t-statistic	0.705	0.516	0.001	2.700	0.696	0.529	0.000	2.694
$P(\|t\| > T_{5\%})$	0.024	0.153	0.000	1.000	0.027	0.162	0.000	1.000
$Treated_s \times FRP_{t-2}$								
Coefficient	-0.001	0.025	-0.081	0.069	-0.000	0.031	-0.100	0.095
Std. Error	0.027	0.000	0.027	0.027	0.035	0.000	0.035	0.035
t-statistic	0.741	0.539	0.001	2.992	0.698	0.536	0.000	2.866
$P(\|t\| > T_{5\%})$	0.026	0.159	0.000	1.000	0.028	0.165	0.000	1.000
# Replications	1000	1000	1000	1000	1000	1000	1000	1000

Notes: Based on 1000 replications. This table reports difference-in-difference estimates on the control group sample when we randomly allocate treatment status to 1500 control schools. All specifications include fixed effects by school and exam year. The outcome variables are average standardized SABER11 test scores for math and language subjects at year t. The coefficients of interest are the interaction of an indicator of (false) treatment status with a set of dummy variables FRP_{t-1} and FRP_{t-2}, measuring the exposure to the FRP regime one and two years before the SABER11 exam is taken, respectively.