



UNIVERSIT  DEGLI STUDI DI MILANO

Performance based incentives in education: an impact evaluation in state run schools of S o Paulo

Thesis presented by
Eduardo Keller da Silva

Supervisor
Catherine Guirkinge (UCL/UNamur)

Supervisor and Reader
Tommaso Frattini (UNIMI)

Academic year 2018

In order to obtain the Joint Degree
Master 120 en Sciences  conomiques, Orientation g n rale, Finalit  sp cialis e (UCL/UNamur)

and

Dottore magistrale in Economics and Political Science (UNIMI)

Acknowledgements

I would like to thank first and foremost my thesis supervisor, professor Catherine Guirking, for all the help and guidance provided during the process. From the early stages of defining a topic to the late challenges presented by the research, her door was always open and her suggestions and critiques always extremely helpful. Her passion and commitment to teaching and research is an inspiration not only for me but for all of her students.

I thank my reader from the University of Milan, professor Tommaso Frattini, for his support and interest in the topic proposed.

I would also like to thank my family and specially my father, Ruy, for all the help during my international adventures and for raising me with a critical view of the world.

I also thank the University of Milan, the University of Namur and the Catholic University of Louvain for the opportunity given and for the great program in economics proposed.

Last but not least, I thank all my friends for making this world a warmer place to be in.

Abstract

The study evaluates the impact of a performance based incentive program implemented in state run schools in São Paulo, Brazil. The program was implemented in 2008 and distributed bonuses to teachers and employees that reached certain targets linked to students test scores in math and Portuguese. The design of the incentive also assigned more weight to improvements made by the worst performing students. We employ difference in differences estimation with propensity score matching to estimate average treatment effects on the treated for the following outcomes for 2009 and 2011: (i) average test scores, (ii) coefficient of variation of test scores, (iii) share of students performing below the basic and at or below the basic threshold established by educational guidelines. The results indicate that the program was only effective on raising student's test scores for the 5th grade in 2009. For the 9th grade in both years and the 5th in 2011 the results indicate that the intervention was ineffective at best. When looking to the impact on the share of low performing students, it seems that the program was successful in decreasing the share of low achievers in 2009 in both the 9th and 5th grades. As an example, the results suggest that the program contributed to a decrease in the share of those performing below basic of around 4 and 2 percentage points in math and Portuguese, respectively. In 2011 the estimations are less conclusive.

List of Figures

Figure 1 - primary education enrollment by provider, 1990 - 2009.....	17
Figure 2 - illustration of the parallel trends assumption.....	24
Figure 3 - percentage standardized bias before and after the matching - 5th grade with municipally run schools as controls	58
Figure 4 - percentage standardized bias before and after the matching - 9th grade with municipally run schools as controls	58
Figure 5 - percentage standardized bias before and after the matching - 5th grade with state run schools as controls	59
Figure 6 - percentage standardized bias before and after the matching - 9th grade with state run schools as controls	59
Figure 7 - average test scores for the 5th grade - municipally run schools as controls.....	60
Figure 8 - average coefficient of variation in test scores for the 5th grade - municipally run schools as controls.....	60
Figure 9 - average share of students below the basic in the 5th grade - municipally run schools as controls	60
Figure 10 - average share of students at or below the basic in the 5th grade - municipally run schools as controls.....	61
Figure 11 - average test scores for the 9th grade - municipally run schools as controls.....	61
Figure 12 - average coefficient of variation in test scores for the 9th grade - municipally run schools as controls.....	61
Figure 13 - average share of students below basic in the 9th grade - municipally run schools as controls.....	62
Figure 14 - average share of students at or below basic in the 9th grade - municipally run schools as controls.....	62
Figure 15 - average test scores for the 5th grade - state run schools of neighboring states as controls ...	62
Figure 16 - average coefficient of variation in test scores for the 5th grade - state run schools of neighboring states as controls	63
Figure 17 - average share of students below basic in the 5th grade - state run schools of neighboring states as controls.....	63
Figure 18 - average share of students at or below basic in the 5th grade - state run schools of neighboring states as controls.....	63
Figure 19 - average test scores for the 9th grade - state run schools of neighboring states as controls ...	64
Figure 20 - average coefficient of variation in test scores for the 9th grade - state run schools of neighboring states as controls	64
Figure 21 - average share of students below basic in the 9th grade - state run schools of neighboring states as controls.....	64
Figure 22 - average share of students at or below basic in the 9th grade - state run schools of neighboring states as controls.....	65

List of Tables

Table 1- Structure of the education system and educational authority	16
Table 2 - Reference scores for the achievement scale.....	19
Table 3 - Number of schools in the sample.....	27
Table 4 - Average share of students below basic or at or below basic in math scores.....	28
Table 5 - Average share of students below basic or at or below basic in Portuguese scores.....	28
Table 6 - variables selected for estimating the propensity score	30
Table 7 - Comparison between state run schools and municipally run schools from São Paulo.....	31
Table 8 - Comparison between state run schools of São Paulo and neighboring states	32
Table 9 - Comparison of covariates before and after matching - 5th grade with municipally run schools as controls.....	35
Table 10 - Comparison of covariates before and after matching - 9th grade with municipally run schools as controls	36
Table 11 - Comparison of covariates before and after matching - 5th grade with state run schools as controls.....	38
Table 12 - Comparison of covariates before and after matching - 9th grade with state run schools as controls.....	39
Table 13 - Placebo regression for the 5th grade using municipally run schools as controls	42
Table 14 - Placebo regression for the 5th grade using state run schools of neighboring as controls	42
Table 15 - Placebo regression for the 9th grade using municipally run schools as controls	43
Table 16 - Placebo regression for the 9th grade using state run schools of neighboring as controls	43
Table 17 - Difference in differences estimation for the 5th grade - municipally run schools as controls ..	44
Table 18 - Difference in differences estimation for the 5th grade - state run schools as controls.....	45
Table 19 - Difference in differences estimation for the 9th grade - municipally run schools as controls ..	46
Table 20 - Difference in differences estimation for the 9th grade - state run schools as controls.....	47
Table 21 - Difference in differences estimations from Oshiro et al. (2015).....	48
Table 22 - Estimated coefficients of probit - 5th grade with municipally run schools as controls	55
Table 23 - Estimated coefficients of probit - 9th grade with municipally run schools as controls	55
Table 24 - Estimated coefficients of probit - 5th grade with state run schools as controls.....	56
Table 25 - Estimated coefficients of probit - 9th grade with state run schools as controls.....	57

Table of Contents

1.	Introduction.....	7
2.	Literature Review	9
3.	The Brazilian Education System.....	15
4.	Design of the Incentive program in São Paulo	18
5.	Methodology	21
5.1.	Propensity Score Matching	21
5.2.	Difference in Differences.....	22
6.	Data	26
7.	Results	34
7.1.	Matching.....	34
7.2.	Regressions.....	41
8.	Conclusion and Discussion	50
	Bibliography.....	53
	Appendix.....	55

1. Introduction

Over the past twenty years Brazil has made significant improvements in its education indicators, especially when considering the provision of universal primary education. In 2012, over 98 percent of the population aged 6 to 14 was enrolled in the primary cycle (UNESCO, 2015). However, the country is still lagging behind when the quality of the service provided is taken into account. In 2009, 60 percent of students aged 15 scored below the basic threshold established by the PISA¹ international assessment, a margin that is well above the OECD average of 14 percent (Bruns et al. 2012).

The literature on education policies frequently draw a line between input versus incentive oriented initiatives to improve school quality. Hanushek (2003) highlights that many of the standard input-based policies, like lowering class sizes or increasing the requirements for teacher's credentials, have not been very effective on achieving this goal. In the Brazilian case, Monteiro (2015) verified that an exogenous increase in the budget of municipally run schools had no effect on indicators related to the quality of education.

In this context, there has been a growing interest on incentive-based policies that link direct measures of teacher's performance to their remuneration. The idea has attracted particular attention since wages are usually the largest component of school's budget (Muralidharan and Sundararaman, 2011).

Following on with this trend, the aim of the present study is to evaluate the impact of a pay for performance program implemented in schools ran by the state of São Paulo, Brazil. The program was created in 2008 and distributed bonuses to teachers and employees of schools that reached certain targets linked to student's test scores in math and Portuguese in the 5th and 9th grades. The design of the incentive also assigned more weight to improvements made by the worst performing students. In this set, average treatment effects on the treated are estimated for the following outcomes for math and Portuguese by grade (5th or 9th): (i) average test scores, (ii) coefficient of variation of test scores, (iii) share of students performing below the basic threshold established by educational guidelines, (iv) share of students performing at or below the basic threshold established by educational guidelines. The study therefore evaluates the impact of the program not only on average scores but also its effect on low achievers. The impact is computed for two post-treatment years, 2009 and 2011.

The strategy employed to estimate treatment effects is to combine propensity score matching with a difference in differences estimation. In the first case, the methodology will allow us to mitigate any potential bias arising from observed differences between treatment and control groups. The second method will address potential biases derived from differences that are not observed by the econometrician. If these unobserved disparities are fixed over time, the difference in differences estimation should produce more accurate estimates of average treatment effects.

The dataset used is made available by the Brazilian System of Basic Education Assessment. The assessment has a census format, evaluating the 5th and 9th grades of all public schools in Brazil every two years since 2005. Information is provided not only on test scores but also on characteristics of students, teachers, principals and schools. With this data, two control groups were defined, one composed by municipally run schools in the state of São Paulo and the other by state run schools in neighboring states.

¹ The Programme for International Student Assessment (PISA) is a worldwide study by the OECD intended to evaluate education systems worldwide.

The results indicate that the program was only effective on raising student's test scores for the 5th grade in 2009. For the 9th grade in both years and the 5th in 2011 the results indicate that the intervention was ineffective at best. When looking to the impact on the share of low performing students, it seems that the program was successful in decreasing the share of low achievers in 2009 in both the 9th and 5th grades. As an example, the results suggest that the program contributed to a decrease in the share of those performing below basic of around 4 and 2 percentage points in math and Portuguese, respectively. In 2011 the estimations are less conclusive.

The main contribution of the study to the literature is to estimate the effects of the program on the share of students performing below or at or below the basic and therefore identify if the design of the program, that puts more weight on improvements made by the worst performers, was successful in its intent. To the best of our knowledge, this is the first research to address this issue in the intervention implemented in the state of São Paulo.

The rest of the study is organized as follows: section two provides a review of the literature related to pay for performance schemes. Section three describes the main features of the Brazilian education system. Section four characterize the design of the incentive implemented in state run schools of São Paulo. Section five describes the methodology employed in the estimations. Section six discuss the characteristics of the dataset used. Section seven present results on the impact of the incentive program on average test scores and share of low performing students. Section eight concludes with a discussion of the potential mechanisms operating behind the findings and point out to directions for future research.

2. Literature Review

The institutional setting of educational systems is usually characterized by a remuneration structure of teachers that have weak links to performance. In most cases salary progression is relatively flat and the promotion policies are strictly linked to seniority rather than individual performance. This setting might create weak incentives for teachers to perform in the best of their abilities (Bruns et al., 2011). Reasons for this of lack of linkage between wages and performance can be partly attributed to difficulties on measuring the individual productivity of teachers as well as concerns that this procedure could divert efforts from educational outcomes that cannot be measured (Baker, 1992; Holmstrom et al., 1991 apud Hanushek et al., 2016).

Critics to merit-based remuneration structures emphasize that for complex activities that are multidimensional, which seems to be the case of education, these types of policy can generate negative impacts. These structures are necessarily based on a measurable outcome that must be met in order for the benefits to be released, the outcomes usually being teacher's attendance or students test scores. The problems are more prominent in the second case, where efforts might be diverted to the subjects being measured, usually math and language tests, to the detriment of other areas such as history and philosophy. Also, there are more subjective dimensions in education that do not seem to be measurable, like group work skills, tolerance to diversity, democratic values etc. which might be desirable but could be negatively affected by such schemes.

Another problem that might arise is the practice of cheating or teaching to the test. For example, the employee might make efforts so that artificially low targets are set, or may allow students to copy from each other. Moreover, students in the lower bound of the tests scores distribution might be left behind and the school might make an effort to enroll only the best performing ones (Bruns et al., 2011). Therefore, a key feature of pay for performance programs is designing an incentive scheme that mitigate these potentially harmful behaviors.

From a principal agent theory perspective, pay for performance schemes could address two common problems of this type of relationship, namely, moral hazard and adverse selection. In the first case, the implementation of bonuses and variable salaries could induce a greater alignment between the interests of the principal and the agent for whom the tasks were delegated. In the second case, the adoption of such policies would tend to attract higher quality teachers that expect to perform well under a system of merit pay, while more traditional salary schemes are seen to attract lower grade applicants (Hasnain et al., 2012).

In this context, Ganimian and Murnane (2016) and Hanushek et al., (2016) conducted a broad review on the literature surrounding education initiatives, focusing on studies that either randomly assigned individuals to treatment or exploited exogenous variation in treatment status from natural experiments. Only four studies that analyzed pay for performance schemes conditional on test scores improvements in developing countries and met these methodological criteria were identified.

Muralidharan and Sundararaman (2011) conducted a randomized evaluation of an incentive policy on primary schools in the Indian state of Andhra Pradesh in 2005. An important methodological contribution of the study is that randomization was done not only on the treatment assignment level but also in the selection of a representative sample of schools. In their words, this is the first evaluation of teacher performance pay that employed this experimental design. The study evaluated two types of performance

pay. One in the form of a group incentive, where the bonus disbursement was conditional on the school as a whole increasing average test scores. The other in the form of individual incentives, where payments were made only to the teachers whose students had an increase in average test scores. The results indicate that after 1 year of treatment both interventions performed equally well, with both incentives increasing test scores in math and Telugu (regional language) by 0,14 and 0,16 standard deviations, respectively. However, after two years of implementation, students studying in individual incentive schools outperformed those at group incentive schools. The outcomes were respectively of 0,33 vs 0,22 standard deviations in math and 0,24 vs 0,09 for the language test. These results were also verified in subjects not linked to the bonus scheme like science and history, suggesting the existence of positive spillover effects. The bonus design was such that disbursements were conditional on average improvements in test scores (math and language) subject to a minimum increase of 5% relative to the previous year. This design was established in order to minimize potentially undesirable behaviors connected to the “threshold effect”, where teachers might focus their efforts on students that are close to a given performance target. Average treatment effects by quantile of test scores were also estimated. It was observed that the results were positive for all quantiles and increasing, suggesting that the intervention had stronger effects on high performing students. The results suggested that the main mechanism thorough which students improved their scores was by a greater effort of teacher’s conditional on them being at the school (absenteeism was not impacted by the policy). Another interesting feature of the study is that other two input-based interventions were implemented in order to compare the cost effectiveness of these two types of policies. In one of the treatment groups the schools were provided with an extra contract teacher and in the other with funds for the supply of school materials. The estimations indicated that the input-based schemes had positive and significant effects on scores, but of a lower magnitude when compared to the incentive structure. On top of that, the interventions based on inputs were 25% more costly when compared to the bonus scheme.

Muralidharan (2012) continue the analysis on a follow up study of the previous experiment. The bonus payment was extended to a total of 5 years in 50 schools (out of the 100 in the previous study) of each treatment group. The results indicated that students who had completed their entire five years of primary education in individual bonus schools scored 0,54 and 0,35 standard deviations higher than those in control schools in math and language assessments respectively. For the group bonus treatment, the estimations are less clear. Over the course of 5 years, group bonus schools scored better than the control group most of the time, with some results being non-significant. However, after the end of the 5 year program students in schools in the group incentive scheme did not scored significantly higher than those in control. There is an indication that the mechanism thorough which scores were increased in treatment schools continues to be an increased teaching activity conditional on the teacher being present. The results also suggest that these score increases represented authentic additions to human capital as opposed to a situation where students just learn how to answer the specific exam linked to the bonus, the so called “teaching to the test”. Looking to the different quantiles of students test scores distributions, the study continues to indicate that positive treatment effects were observed in all quantiles of scores. However, again the impact was higher for the better performing students. An interesting contribution of this paper is that schools that were contemplated with the bonus in the first 2 years of the program but were dropped out of the incentive scheme after this period continued to be analyzed. The estimations indicate that the scores of individual incentive schools that stopped receiving the bonus dropped consistently each year, and by the end of the study their scores were not significantly different from those of the control group.

Muralidharan (2012) note that this behavior is well established in the literature, where usually test score improvements obtained from education interventions do not persist over time.

On this matter Andrabi et al. (2011) point out that most models of learning assume that a student's achievements persist between grades. However, recent studies suggest that treatment effects measured by test scores fade rapidly when looking to a longer term. The paper proceeds with a simple model of learning, $y_{it} = \alpha T_{it} + \beta y_{i,t-1} + \mu_i + \vartheta_{it}$, where y_{it} is a student unobserved achievement in period t , T_{it} is the treatment effect and μ_i is unobserved skills that contribute to learning in each period. β is the parameter that links achievements between periods. The canonical value-added models on learning assume that $\beta = 1$. In this setting, estimates of the treatment effect α will be biased in cases where β is wrongly assumed to be equal to one and baseline achievements of treatment and control groups are not the same. Using data from Pakistani public and private schools Andrabi et al. (2011) analyses the issue empirically and come up with estimations of β ranging from 0,2 to 0,5. These results indicate that any long-term extrapolations of short term effects of educational policies should be made with caution. The results are also in line with the behavior of observed in Muralidharan (2012) for schools that were dropped from the incentive program after 2 years.

The analysis described above should be a cause for concern when discussing policies in education. If treatment effects fade out over the long run why should we address these issues at all? Chetty et al. (2011) portrays a brighter scenario for policy making. Using a large administrative database that followed students from their youth to adulthood, it is showed that students who were taught by high value-added teachers (meaning teachers that increased students test scores by a relatively large amount), were more likely to attend college, receive greater salaries and have children as teenagers.

After this short digression we continue with our review of "high quality"² papers that analyzed pay for performance schemes in developing countries. Glewwe et al. (2010) conducted an experimental evaluation of a teacher incentive program in Kenya that offered prizes to schools conditional on absolute performance and also on improvements relative to the baseline achievement. These group incentive scheme had the form of a tournament, where teachers from the 3 best schools in each category (absolute performance and improvement relative to baseline) received a suit worth \$51. Three prizes for the second, third and fourth best performers were also distributed. In this setting teachers from 24 out of 50 treated schools received a prize. The benefits distributed had values that ranged from 21 to 43 percent of a teacher's monthly salary. The results indicated that the program had a positive impact of 0,1 standard deviations on average scores (math and language) in the first year and of 0,22 standard deviations in the second year. After the program ended in the third year, no significant differences were observed between treatment and control schools, suggesting little persistence of treatment effects. The impacts observed in multiple choice questions were significantly greater than those on "fill-in" questions, indicating that students learned strategies for answering multiple choice sheets rather than attaining real gains in human capital. Another issue is that effects were only significant for the test linked to the incentive scheme. When another exam with no relation to the bonus was taken no significant differences between treatment and control were found. The main mechanism thorough which the effects seem to have operated is by an

² We use the label "high quality" because is a term often used in the literature and also in Hanushek et al., (2016). However, we are aware of the critiques and dangers of portraying randomized control trials as the holy grail of scientific method in social sciences. For a further discussion on the issue see Deaton et al. (2018).

increased effort on test preparation sessions in treated schools. They conclude by highlighting that teacher incentive schemes may not be an effective strategy to support long-term learning.

Barrera-Osorio and Raju (2017) conducted a randomized evaluation of a performance pay program conceived, designed and managed by the government of Punjab, Pakistan. We emphasize here that the intervention analyzed in this study was run by the government, while the programs of the 3 previous evaluations discussed had the influence of NGOs and international organizations. Following on, the study randomized treatment assignment within the lowest performing public primary schools in Punjab. In the first treatment variant, only the school principal is entitled to receive the bonus. In the second variant, all teachers and the principal were contemplated. In the third and last variant, the principal is entitled to receive a bonus that two times greater than the one of the teachers. The disbursement of the benefits was conditional on a composite score with 3 parameters: gains in school enrollment, gains in average test scores and the participation rate of students. The results indicated that none of the treatment variants were successful in raising students test scores in any of the 3 years where the program was run. The intervention seems to have increased school enrolment by 4,1% but only for the first treatment variant. Regarding exam participation, the program increases participation rates on average by 3,4 percentage points in the third year for all treatment variants. The authors consider 3 hypotheses for the lack of effect on average test scores: (i) using only low performing schools might have affected the results (ii) teachers may lack the skills to raise student's achievements and (iii) teachers may optimally direct their efforts on variables (participation rate and enrolment) that maximize payoffs net of effort costs.

We have identified another experiment on a pay for performance scheme in a developing country that was not mentioned by Ganimian and Murnane (2016) and Hanushek et al., (2016). Behrman et al. (2015) conducted a 3 years randomized trial in federal high schools in Mexico, where more than 40 thousand students were randomly assigned to 3 treatment groups. The first treatment variant provided individual incentives to the students themselves, conditional on improvements made in math scores. The second variant gave bonuses to math teachers that improved scores of their own students. Therefore, the incentive is at the individual level (not group based). The third and last treatment group distributed bonuses to students conditional on their own improvements and also in the gains made by their colleagues in class. Payments to math teachers were made according to achievements made by their own class and also on improvements from all other math classes. Disbursements to non-math teachers and school administrators were made conditional on improvements made by all students in the school. In this sense, this last bonus type can be characterized as a group incentive. The design of the scheme also took into account the distribution of student's achievements, putting more weight on improvements made by the worst performers. The results indicate that only the first and third treatment were successful on raising scores. The effects of treatment 2 where only math teachers received bonuses were not significant in any of the years. By the end of the third year, students in the first treatment group had increased their scores by 0,23 standard deviations relative to the control group and the third treatment made improvements of about 0,57 standard deviations. However, a closer examination of the text book answers revealed that part of the gains in scores were due to students copying from each other. No evidence of teachers cheating was found. After accounting for students copying the results remained positive and significant. Moreover, the gains were relatively equally distributed within the range of student's scores, indicating that the design of the bonus was successful on not letting the worst performers behind.

Ganimian and Murnane (2016) also point out that with only one exception, all the studies reviewed that evaluated merit pay plans in the United States came up with null results.

Looking to the Brazilian case, two impact evaluations of a pay for performance program were identified, both relying on propensity score matching and difference in differences estimations. They both analyzed the program implemented in the state of São Paulo in 2008, which will also be the subject of this study. Oshiro et al. (2015) is the work that deals more closely with the analysis presented in this study. It sort of worked as a reference guide while establishing our methodology and running our estimations. We are therefore grateful for their work. The paper was derived from Oshiro (2012), a master's thesis presented at the University of São Paulo. Their research estimated the impact of the performance incentive scheme implemented in state run schools of São Paulo on average test scores of 5th and 9th grade students in math and Portuguese. The results indicate that the program had a positive and significant impact on math and Portuguese scores for the 5th grade on the year right after the implementation (2009), but not 3 years after (2011). The estimations for the 9th grade were inconclusive, with some specifications presenting negative and significant impacts on test scores. This pattern is very similar to the one found in this study. While attempting to explain the results described above, Oshiro et al. (2015) investigate if the bonus scheme promoted changes in school characteristics (student's socio-economic background, pedagogical practices and teacher's and principal's traits). The results indicate that there was no clear variation in characteristics that could explain the behavior described above. We highlight the fact that the socio-economic background of students did not seem to be significantly impacted by the program. In the cases where the results were significant for this variable, the coefficients were negative, indicating that student's background actually worsened after the program. This variable is important because it could indicate a potential selection of best performing students in the day of the exam, a scenario that is not supported by the data. Because of time constraints, we do not test for potential changes in schools' characteristics ourselves, relying on the results of Oshiro et al. (2015) to discuss the issue.

The methodologies employed in this study and the paper described above are very similar. We now focus in some of the differences between them. First, this study makes an original contribution to the evaluation of the program by analyzing treatment effects not only on average test scores but also in the shares of worse performing students. Secondly, there is a data limitation for the year of 2005 that was dealt with differently in the two studies. To summarize the problem, for the year of 2005 there is only data available for a small representative sample of schools. After implementing the matching, our sample of treated and untreated institutions was reduced to a size that we deemed not acceptable³. This represented a problem especially while trying to analyze if the parallel trends assumption required for a difference in differences estimation was supported by the data. Oshiro et al. (2015) do not discuss or mention this issue and use a matched sample for the year of 2005. In this study on the other hand, it was decided that the noise provoked by the sample reduction while implementing propensity score matching would be more detrimental for the results than any potential bias arising from differences in observed characteristics between treatment and control groups. In this context, the issue of parallel trends was discussed here with no matching between treatment and control groups. In order to maintain coherence, we also present the difference in differences estimations with no matching at first and then with matched samples afterwards.

Guimarães et al., (2017) also analyzed the same pay for performance program but focused on the impact that the intervention had on teachers and principal's behaviors. The methods used in the evaluation were the same: difference in difference estimation with propensity score matching. For both teacher's and principal's behaviors the results were inconclusive. Moreover, the study also analyzed the evolution of student test scores. The estimations were similar to the ones of Oshiro et al. (2015), with significant and

³ The issue will be discussed more in depth in the data and results sections.

positive impacts on test scores only for 5th grade students on the year after the implementation of the program.

Another study that was important for this research was conducted by Scorzafave et al. (2015). In this non-published paper, they investigate the effects of pay for performance schemes in the distribution of test scores in Brazil. The strategy was to analyze the differences in distributions of test scores in state run schools in 3 groups of states: those where no incentive scheme was implemented, those where an incentive scheme was implemented with a design that aimed at reducing inequality of achievements and those states that implemented an incentive scheme without an inequality reducing formula. Out of the 27 states in Brazil, 4 implemented policies that rewarded teacher's conditional on student's achievements and added more weight to improvements made by the worse performing ones, 9 states implemented such policies but without considering such inequalities and the other 14 states did not engaged in this type of intervention. Some of the strategies used were to estimate the effects of these different policies on standard deviations of test scores and the ratio between different percentiles of scores. For these outcomes the effects were different from the expectations. The results indicated that designs that induced a reduction in inequality were either ineffective or actually increased the dispersion of scores. On the other hand, when they examined treatment effects on the share of students performing at or below the basic level established by national guidelines the results were more in line with the expectations. States where incentive schemes that induced the reduction of inequalities were implemented observed a reduction in the shares of those performing at or below the basic. Although the effects were rather small, ranging from 0,5 to 0,9 percent in absolute terms.

In this study we follow the contribution made by Scorzafave et al. (2015) and use the share of those at or below the basic as an outcome to be evaluated in the specific case of state run schools from São Paulo.

3. The Brazilian Education System

According to Bruns et al. (2011) apud McGinn et al., (1999) educational systems can usually be characterized by four sources of authority: central or federal government; provincial, state or regional governing bodies; municipal, county or district administrations; and schools. In this setting, the potential decentralization of the educational authority can be made from the central government to lower levels of government and/or from lower level governments to schools. We are going to refer to these four levels of authority as federal or central government, state, municipal and school for the purposes of this study.

With this contextualization, in 1996 the National Education Guidelines and Framework Law (Lei de Diretrizes e Bases da Educação - LDB) was approved in Brazil, structuring what can be defined as the contemporary Brazilian educational system. This law provides for the decentralization of authority, establishing clear responsibilities between the four levels of jurisdiction. The federal government, embodied by the Ministry of Education, defines the guiding principles and coordinates policies for all education levels, in collaboration with states and municipalities. It implements education plans and defines the guidelines of the national curriculum. It is also entitled to provide technical and financial assistance to states and municipalities for the development of their school systems. The central government is also the body responsible for the management of higher education. The states, embodied by the Secretariat of State Education, are responsible for providing primary and secondary education, while the municipalities, embodied by the Secretariat of Municipal Education, are responsible for the early childhood and primary levels. The state authority is also responsible for the regulation and supervision of private schools. As we can see, there is an overlap between states and municipalities and considering the provision of primary education. In this setting, the states and municipalities are responsible for administering these educational levels. They define their own education policy using the national guidelines and are responsible for building schools, providing equipment, meals, transport, hiring teachers and paying their salaries. At the school level, the institutions are responsible for writing and implementing their pedagogical development plan (Bruns et al., 2011; Bruns et al., 2012; OECD, 2015; Unesco, 2010).

Table 1 provides a visual representation of the structure of the education system. Along with the nomenclature of each educational level, it indicates the ideal amount of time required to complete each cycle, the recommended age of each level and also the authority responsible for administering the system.

Table 1- Structure of the education system and educational authority

General programme	Specific programme	International denomination	Grades or years	Theoretical age	Authority
Basic education	Early childhood education	Initial education	4 years	0 - 3	Municipalities
		Pre-school	2 years	4 - 5	
	Primary education	Elementary education	1 st to 5 th grades	6 - 10	Municipalities and States
		Lower secondary school	6 th to 9 th grades	11 - 14	
	Secondary education	Upper secondary education (high school)	1 st to 3 rd form	15 - 17	States
Higher education	Graduate	Bachelor's	2.400 hours (4 - 5 years)	18 or older	Federal
	Postgraduate	Master's	2 years		
		Phd	2 - 4 years		

Source: Ministry of Education, 2008 apud UNESCO, 2010

The 1996 law also brought a new legislation regarding the funding of the educational system. Before this period, large imbalances in expenditure per student existed between regions, and, within states, between school provider (state run or municipally run). As an example,

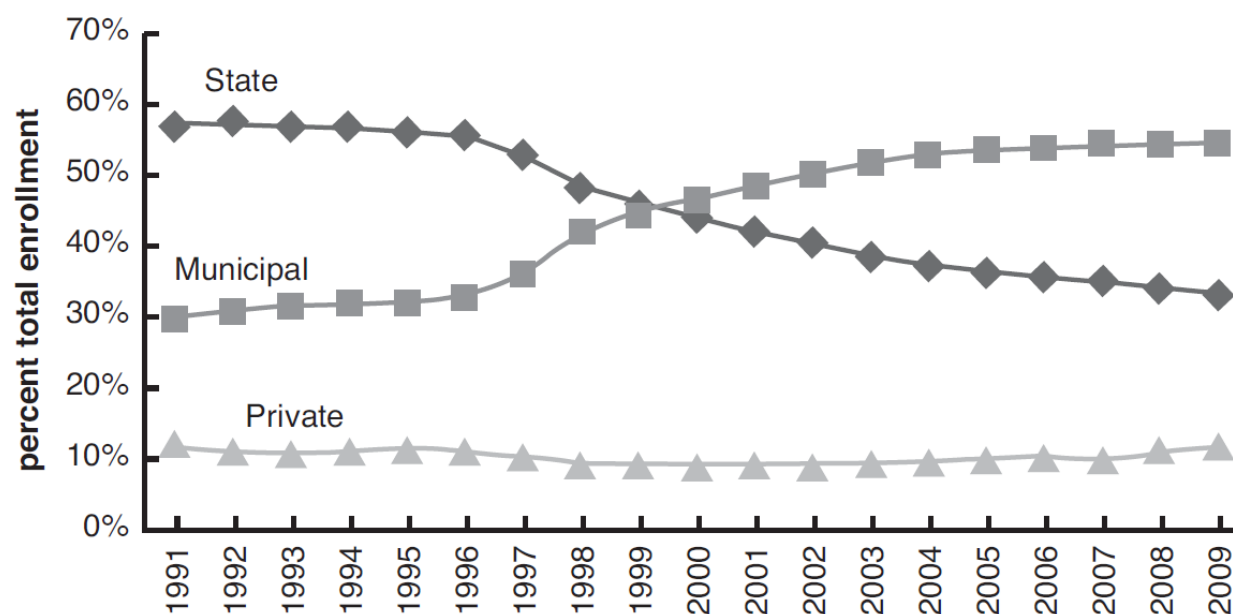
“while spending per primary student in municipal schools in parts of northeast Brazil was less than Rs 100 per year (lower than in Nicaragua and Bolivia), it could be Rs 600 or more in state schools in the same region and Rs 1,500 or more in state and municipal systems in the southeast (on par with schools in Korea and Singapore) (Bruns et al., 2012, p. 4).”

Before 1996, 25% of all state and municipal taxes were required to be spent in education. However, the old legislation did not take into account the vast disparities in tax revenues between regions and jurisdictions. After the new legislation was passed, a minimum expenditure per student was defined. If the municipal school network did not reach the lower limit spending requirement, the state administration should relocate funds to that jurisdiction. Another measure was the creation of the Fundamental Education Development Fund (Fundo de Desenvolvimento do Ensino Fundamental – FUNDEF). Whenever state administrations are not able to meet the minimum level of expenditure required, resources from this federally supported fund were diverted to the jurisdiction in need (Bruns et al., 2012 and OECD, 2015).

The well-known inequality of income and wealth in Brazil is also reflected in the educational system. While the majority of the population uses the free and public network. A small share of the youth is enrolled in private institutions. It can be regarded as common knowledge among Brazilian nationals that private schools are considerably better than the public ones, especially when we consider the case of high fees institutions. On the PISA 2012 international assessment, the gap between private and public Brazilian schools was of 83 points after accounting for socio-economic background (the average gap in OECD countries was of 28 score points) (OECD, 2015).

Figure 1 illustrates the evolution of primary education enrollment by type of administration:

Figure 1 - Primary education enrollment by provider, 1990 - 2009



Source: National Studies for Education Studies and Research apud Bruns et al., 2012

It can be seen that around 10% of Brazilian students in primary education are enrolled in private institutions, while the great majority goes to public schools. Figure 1 also indicates that after the approval of the new legislation in 1996, municipal schools started to play a more prominent role in the provision of primary education. On this matter, Oshiro et al., (2015) indicates that the decentralization process that followed the new law led some schools from the elementary education cycle to migrate from a state administration to a municipal one.

Following on with the study, section 4 proceeds with the description of the incentive scheme implemented in state run schools from São Paulo.

4. Design of the Incentive program in São Paulo

The so called School Quality Program (Programa Qualidade da Escola) was launched in May 2008 by the Secretary of Education of the State of São Paulo. The main goal of the program is to improve the quality of education in state run schools of São Paulo.

In order to attain this objective, the policy evaluates the quality of education in all state run schools each year by means of standardized tests and propose targets to be achieved (São Paulo, 2018 a). The exam, here and after denominated SARESP⁴, is applied for the 3rd and 5th grades, which are part of elementary education in Brazil and also to the 7th and 9th grades, which are part of what is called lower secondary school (São Paulo, 2018). The program also evaluates the 3rd grade of high school, but since this is not the focus of this study we are going to refer only to the grades pertaining to primary education.

The program also contemplated the payment of bonuses to teachers and workers of schools that met, fully or partially, targets proposed each year for the 5th and 9th grades. Although only the targets defined for these are linked to the remuneration scheme, the payments are made to all teachers of all grades and subjects by cycle (elementary or lower secondary). That means that, if for example, a certain amount of the target was met by the 5th grade of a given school, all teachers from elementary education of this school (1st, 2nd, 3rd, 4th and 5th grades) will receive the bonus. The same reasoning applies for lower secondary school (6th, 7th, 8th and 9th grades). The remuneration is also given to other workers of the institutions such as the principal and administrative personnel.

In this setting, the program can be classified as a group-based pay for performance scheme. Lavy (2009) points out that group incentives have been more prevalent than individual incentives in the context of educational policies. It is argued that since education is characterized by complex activities that involve multitasking, teamwork and multiple stakeholders, the measurement of the contribution of a given individual is difficult. On the other hand, the presence of free riders can undermine the effectiveness of the program in the context of group-based incentives.

Another feature of the intervention that is worth noting is that although the people responsible for the application of the exam are the teachers of state run schools of São Paulo, they are distributed between schools where they do not teach in the days of the exam. There is also an external inspector assigned to each school in the day of the exam (São Paulo, 2018). These measures should mitigate potential harmful behaviors that could influence test results, such as cheating and allowing the students to copy from each other.

We continue now giving a more in-depth description of the linkage between targets, schools achievements in test scores and the payment of bonuses, using as reference the information available in São Paulo (2018 a) and Oshiro (2012).

The level of school achievement is based on the so called Education Development Index of the State of São Paulo, here and after denominated IDESP. The IDESP is composed by 2 other indexes:

$$\text{IDESP}_s = \text{ID}_s \times \text{IF}_s \quad (1)$$

ID_s or Achievement Index is an index related to student's achievements measured by test scores and IF_s is the Continuation Index. The idea is that the quality of education should take into account both student's

⁴ System of Evaluation of School Achievement of the State of São Paulo

achievements and the ideal time in which this knowledge should be acquired. The subscription S denotes the cycle for which the quality is being measured (elementary or lower secondary).

The ID_s index corresponds to the relative share of students in each achievement scale established by the program: Below Basic (BB), Basic (B), Adequate (Ad) and Advanced (Av). The first category corresponds to students that demonstrate an insufficient knowledge for their grade and the second one indicate that a student only partially achieved the required competencies. Table (2) presents the reference scores⁵ that comprise each category:

Table 2 - reference scores for the achievement scale

	5th grade		9th grade	
	Portuguese	Math	Portuguese	Math
Below basic	<150	<175	<200	<225
Basic	150 to 200	175 to 225	200 to 275	225 to 300
Adequate	200 to 250	225 to 275	275 to 325	300 to 350
Advanced	>250	>275	>325	>350

Source: School Quality Program technical note, São Paulo (2018)

The ID_{js} is computed with the following formula:

$$ID_{js} = \left(1 - \frac{def_{js}}{3}\right) \cdot 10 \quad (2)$$

Where the subscription j now denotes the subject being evaluated, Portuguese or math. def_{js} represents how much the students are lagging behind from where they should be, which would correspond to the adequate category indicated above. It is defined as:

$$def_{js} = \frac{(3 \cdot BB) + (2 \cdot B) + (1 \cdot Ad) + (0 \cdot Av)}{100} \quad (3)$$

Where BB, B, Ad and Av correspond to the share of students below basic, basic, adequate and advanced levels, respectively. def_{js} ranges from 0 (when all students are at the advanced level) to 3 (when all are at the below basic category). ID_{js} increases with the improvement of the school quality, ranging from 0 when the all students are below the basic and 10 when all are advanced.

Here we can observe an important feature of the program. It puts more weight on the share of worse performing students, creating an incentive for teachers to focus their efforts in these groups. However, there does not seem to be any rationale for the weights indicated above. This could be an interesting line for future research. Investigating which is the optimal weight for the share of students in each category when defining an incentive scheme based on a school quality index.

⁵ The scores run from 0 to 500. It is worth noting that we do not present the scales for the 3rd year of high school, where the advanced category can go until 500.

The ID_s for each school and cycle s is then calculated by making the average between the ID_{js} of each subject j (Portuguese and Math).

The continuation index or IF_s indicates the ratio between students approved and students enrolled by education cycle. It is given by:

$$IF_s = \frac{A_1 + \dots + A_n}{n} = \frac{\sum_{i=1}^n A_i}{n} \quad (4)$$

Where A_i is the ratio between students approved in grade i and the number of students enrolled in this same grade and n is the number of grades in each education cycle (elementary or lower secondary). In this setting, the IF_s or continuation index considers the approval rates of all grades.

Now that we have defined the Education Development Index we give a short description of how the targets for the IDESP were set. First, a long-term target was defined for the 5th and 9th grades. The idea is that by 2030 all the state-run schools of São Paulo should achieve the average level of education quality⁶ experienced by OECD⁷ countries. Once this long-term goal was established, an annual target was set for each school and education cycle so that the IDESP reaches the desired level by 2030.

With both the IDESP and the targets defined, a measure of how the quality of the school evolved in relation to their annual targets can be established. The Compliance Index or IC is defined as:

$$IC_{s,t} = \frac{IDESP_{s,t} - IDESP_{s,t-1}}{Target_{s,t} - IDESP_{s,t-1}} \quad (5)$$

The Compliance Index for the cycle s (elementary or lower secondary) at time t is equal to the ratio between the difference of the IDESP in period t and $t-1$ and the difference between the target at time t and the IDESP of $t-1$. The IC can range from 0% to 120%. If the IC is negative, the lower limit of 0% is considered. If it surpasses the upper threshold, 120% of compliance is considered.

An additional Quality Index or IQ was defined in order to incentivize schools that perform above the state average. The IQ for year t and cycle s is given by:

$$IQ_{s,t} = \frac{IDESP_{s,t} - IDESP(average)_{s,t}}{Target_{s,2030} - IDESP(average)_{s,t}} \quad (6)$$

Where IDESP (average) correspond to the average of the state-run schools of São Paulo in year t and $Target_{s,2030}$ indicate the long-term target set for 2030. If the index is negative, 0% is taken into account.

For the payment of the bonus the sum of IQ and IC is considered, with the percentage being limited again to 0% and 120%. If for example a school reaches 120% of accomplishment, their employees receive a bonus that is equal to 2,9 times their salary. If 100% is met, the bonus goes down to 2,4 salaries. The payments follow this logic in all levels of accomplishment until the limit of 0% is reached, a case in which no bonus is received. A condition for the employee to receive this payment is to have a level of absenteeism lower than 1/3, a measure that can potentially decrease the problem of free riding.

⁶ The education quality of OECD countries was measured using the Programme for International Student Assessment (PISA) as a reference. This assessment is a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. The structure of the PISA and SARESP assessments are similar allowing for comparisons between the 2 exams.

⁷ The Organisation for Economic Co-operation and Development.

5. Methodology

5.1. Propensity Score Matching

One of the fundamental problems of impact evaluations and causal inference in general is that we cannot observe the trajectory of a certain subject or group in two different states, one where it has received treatment and the other where it has not. Borrowing from Angrist and Pischke (2009) notation, the issue can be more formally illustrated in the following scenario:

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

$$= Y_{0i} + (Y_{1i} - Y_{0i})D_i.$$

An observed outcome Y_i can be explained in terms of its potential outcomes given treatment assignment. Y_{1i} corresponds to a world where subject i was subjected to treatment, i.e, $D_i = 1$ and Y_{0i} indicates the outcome for the same individual i in the absence of an intervention. The causal effect resides on the difference between these two potential outcomes $Y_{1i} - Y_{0i}$. Since we cannot observe individual i simultaneously in these two states, one must look to another set of subjects who have not received treatment in order to infer a causal effect. In this sense, it is also not possible or plausible in most empirical settings to estimate individual treatment effects, leading the observer to concentrate on population average treatment effects (Abadie and Cattaneo, 2018; Caliendo and Kopeinig, 2008)⁸.

With this limitation in mind, two parameters related to causal inference are most commonly discussed in the literature, the average treatment effect (ATE) and the average treatment effect on the treated (ATT). In the context of public policy and program evaluation, the ATT estimate seems to be the most relevant method of assessment given that an intervention is usually targeted at a specific population. For example, if we are evaluating a conditional cash transfer program, which is usually targeted at families with a low income, it is more relevant for a policy maker to understand the effect of the intervention on this group rather than the whole population. For this reason, we are going to focus our attention on the ATT indicator.

Following Caliendo and Kopeinig (2008), the ATT is given by:

$$\tau_{att} = E(\tau|D = 1) = E(Y(1)|D = 1) - E(Y(0)|D = 1) \quad (7)$$

Where the average treatment effect on the treated is the difference between the expected outcome of an individual that received treatment, i.e, $E[Y(1)|D=1]$ and the expected outcome of this same subject in a scenario where the intervention have not occurred, $E[Y(0)|D=1]$. As already stated, one major problem of causal inference is that we cannot observe the counterfactual, the second term of this equation. A possible solution is to use the average outcome of untreated individuals as a counterfactual $E[Y(0)|D=0]$.

In this case, a fundamental assumption for the ATT to be efficiently estimated is that:

$$E[Y(0)|D=1] = E[Y(0)|D=0] \quad (8)$$

That is to say, that the expected outcome for the treated and untreated would be the same had they not received the intervention. In the case of an experimental design with random assignment of treatment

⁸ p.5 and p.34, respectively

status this condition is ensured and the ATT can be identified (Caliendo and Kopeinig, 2008). Theoretically, randomization leads to an unbiased estimate of treatment effects because all factors besides treatment status will tend to be on average equally distributed between treatment and control groups. This would be the case not only for the observable characteristics of individuals but also for the unobserved ones. In this context, all competing explanations for the difference in outcome between the treatment and control groups can be discarded and the causal impact of treatment can be assessed (Murnane and Willett, 2011)⁹.

However, in the absence of randomization the relation exemplified in equation (8) is likely to be violated. In this case, the ATT won't be simply the difference between the observed outcomes of the treated and untreated. A selection bias component will need to be added. Using a rearranged equation from Angrist and Pischke (2009), the ATT can now be stated as:

$$E[Y_{1i} - Y_{0i} | D_i = 1] = E[Y_i | D_i = 1] - E[Y_i | D_i = 0] - E[Y_{0i} | D_i = 1] + E[Y_{0i} | D_i = 0] \quad (9)$$

Where $E[Y_{1i} - Y_{0i} | D_i = 1]$ correspond to the ATT, i.e, the difference between individuals i outcome when treated and untreated. On the right-hand side, the difference between the first two components is the observed difference in the outcome for the treatment and control group, which would equate to the ATT in the absence of bias. The selection bias term is the difference between the last two components of equation (9). It indicates that the outcome of the treated group in the absence of treatment would diverge from that of the control. In this set the true ATT is only identified if (8) holds.

One possible identification strategy to tackle this source of bias in non-experimental studies is to assume that potential outcomes are independent of treatment assignment given a set of observable covariates X which are not affected by treatment. Caliendo and Kopeinig (2008) summarize this strategy under the unconfoundedness assumption:

$$Y(0), Y(1) \perp\!\!\!\perp D \mid X \quad (10)$$

The relation implies that potential outcomes are independent of treatment assignment when all variables that affect simultaneously selection to treatment and outcomes are observed by the researcher. Under this assumption the selection bias would "vanish". In the context of a high dimensionality of the vector X , Rosenbaum and Rubin (1983) apud Caliendo and Kopeinig (2008) suggest the use of a balance score. It is showed that if the unconfoundedness assumption, also known as conditional independence assumption (CIA), holds for a set of covariates X , it will also hold for a balancing score $b(X)$. The propensity score is one of this possibilities and is given by $P(D = 1 | X) = P(X)$, where the probability of treatment for an individual is estimated. The unconfoundedness assumption can be re-written substituting X by the propensity score.

Another requirement while tackling the selection bias with (CIA) is that the perfect predictability of treatment assignment given X must be ruled out. More formally the notion can be stated as:

$$\text{Overlap: } 0 < P(D = 1 | X) < 1 \quad (11)$$

It implies that an individual with the same set of covariates X have a positive probability of being both treated and control.

5.2. Difference in Differences

⁹ For a critique of this proposition see Deaton et al. (2018).

Until now we have seen how underlying differences between treatment and control groups can bias the estimation of average treatment effects on the treated. If these differences are observed by the econometrician, it is possible to assume that treatment assignment is independent of a given outcome that we want to evaluate conditional on a set of observed characteristics. This relationship was formally stated in equation (10) under the unconfoundedness or conditional independence assumption.

However, there is also the possibility that there are unobserved characteristics that simultaneously affect treatment assignment and an outcome Y , biasing our results of the ATT. If these unobservables are fixed over time, we can use a difference in differences estimation to mitigate this source of bias.

Following Angrist and Pischke (2009) and bringing the formulation of the problem to the context of this study, the strategy can be summarized in the following way:

$$Y_{ist} = \beta D_{st} + \gamma_s + \delta_t + \varepsilon_{ist} \quad (12)$$

Where i denotes the school, s indexes which is the school administration type, if it is a state run school from São Paulo or one of our 2 control groups (municipally run schools from São Paulo or state run schools from neighboring states) and t denotes the period (2007, before the intervention or 2009 and 2011, after the intervention). D_{st} is a dummy for state run schools of São Paulo and β corresponds to a constant defined as $E(Y_{1ist} - Y_{0ist} | s, t)$. Equation (12) indicates that an outcome Y_{ist} is determined by a period effect δ_t that is common across treatment and control, a time invariant confounder γ_s that is not independent from D_{st} (Abadie and Cattaneo, 2018) and the average treatment effect on the treated βD_{st} . The error term ε_{ist} is assumed to be independent from treatment assignment so that $E(\varepsilon_{ist} | D_{st}) = E(\varepsilon_{ist})$.

The main objective of the difference in difference estimation is to identify β and get rid of any potential bias arising from γ_s , our time invariant component that is common within treatment assignment. Considering only the years of 2007 (pre-intervention) and 2009 (post-intervention) as an example, this can be done with the following procedure:

$$E(Y_{ist} | s = \text{treatment}, t = 2009) - E(Y_{ist} | s = \text{treatment}, t = 2007) = \delta_{2009} - \delta_{2007} + \gamma_{treated} - \gamma_{treated} + \beta$$

And

$$E(Y_{ist} | s = \text{control}, t = 2009) - E(Y_{ist} | s = \text{control}, t = 2007) = \delta_{2009} - \delta_{2007} + \gamma_{control} - \gamma_{control}$$

In this setting we can already see how we can get rid of the time invariant confounder γ_s that is specific to treatment assignment. In order to get an unbiased estimate β we continue with the difference in differences estimation, that is given by:

$$(E(Y_{ist} | s = \text{treatment}, t = 2009) - E(Y_{ist} | s = \text{treatment}, t = 2007)) - (E(Y_{ist} | s = \text{control}, t = 2009) - E(Y_{ist} | s = \text{control}, t = 2007)) = \beta$$

The specification above indicate that if we do the difference between the variation of Y from 2007 to 2009 for the treated and the variation of Y for the same period in the control group, an unbiased estimate of β or ATT can be obtained.

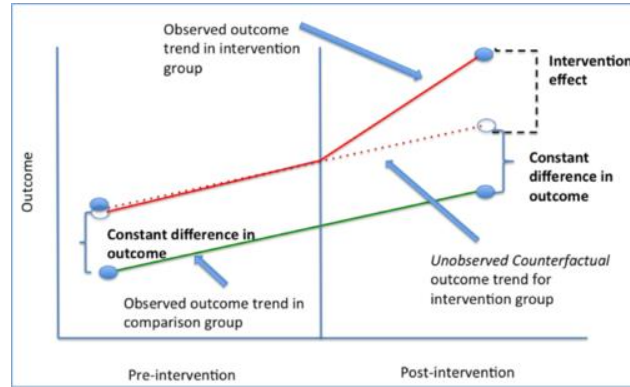
With the framework described above we provide now the regression used to estimate equation (12) in the context of this study:

$$Y_{ist} = a + \gamma \text{ treatment}_s + \delta_1 \text{ year2009}_t + \delta_2 \text{ year2011}_t + \beta_1 \text{ treated2009}_{st} + \beta_2 \text{ treated2011}_{st} + \varepsilon_{ist} \quad (13)$$

Y_{ist} is the outcome of interest. “Treatment” is a dummy variable that assume the value 1 if it is a state-run school from São Paulo and 0 otherwise. “year2009” and “year2011” are dummy variables for the respective periods when the intervention occurred. “treated2009” and “treated2011” are the interaction terms that assume the value 1 only if the observation is from a state-run school from São Paulo in 2009 or 2011. β_1 and β_2 are our coefficients of interest that indicate the average treatment effect on the treated.

The key identifying assumption here is that the trends of an outcome Y would be the same in both treatment and control group in the absence of an intervention. The policy implemented in state run schools of São Paulo would induce a deviation from this trend (Angrist and Pischke, 2009). Figure 2 provides an illustration of the issue:

Figure 2 - Illustration of the parallel trends assumption



Source: Columbia University, Mailman School of Public Health website

We can see that before the intervention the trends of the outcome moved in parallel in the treatment and control groups. After the intervention, we observe that the trend of the treated group changes and the one of the control remains constant. The fundamental assumption is that the trend for the treated group would also remain constant if the intervention had not occurred. The idea is formalized by Abadie and Cattaneo (2018) in the following way:

$$E(\Delta Y_{0is1} | D_{st} = 1) = E(\Delta Y_{0is1} | D_{st} = 0) \quad (14)$$

i.e., the average outcome for the treated had they not received treatment and the average outcome of the untreated would have experienced the same variation over time, a relationship that can be verified when comparing the dotted line and the observed trend of the control group in figure (2).

Although this cannot be proved, the validity of the assumption can be assessed in some ways. First, if we have access to data on multiple periods before the intervention, we can make a visual inspection of trends of the treatment and control groups for these periods. If they moved in tandem before the beginning of the program, it can be argued that is more likely that they would continue to do so had the intervention not occurred (Gertler et al., 2016). Another strategy for analyzing the issue is to conduct a placebo difference in differences for the periods before the intervention only, where it is evaluated the impact of

a non-existent intervention on the treated group in a period before the actual intervention (Abadie and Cattaneo, 2018). Both of these strategies will be discussed in the section where the results are presented.

Following on with our analysis we present now the characteristics of the dataset used in our estimations.

6. Data

To evaluate the impact of the performance-based incentive scheme on student's achievements and test scores distributions we are going to use data from the so called "Prova Brasil" or Brazil Test. This assessment is organized at the federal level by the Brazilian Ministry of Education and the National Institute of Research in Education (INEP). Data from 2005, 2007, 2009 and 2011 is going to be used in our estimations. Its main features are summarized below.

In 2005 the Brazilian System of Basic Education Assessment was restructured and a new tool for evaluation was implemented, the so called "Prova Brasil" or Brazil Test. The exam was implemented in a census format and it is run every 2 years. It consists of an examination of Portuguese and math achievements for 5th and 9th grade students. The universe of the test comprehends all public schools in Brazil, gathering information not only on scores but also on students, teachers, principals and schools' characteristics. The standardized test is based on item response theory which allows for between years and grades comparisons (Fernandes and Gremaud, 2009; DAEB, 2018).

The goals of the test are to evaluate the quality of education at the school and student level and provide systematic information and data aiming at the improvement of the educational system. Regarding the standards of the assessment, Bruns et al. (2012) highlights that it matches and even surpass the current practices in the United States and many OECD countries in the quality, quantity and relevance of the school and student performance information that it provides.

An advantage of using this dataset is that the bonus scheme is not directly linked to the results of "Prova Brasil"¹⁰. As pointed by Oshiro et al., (2015), in this case there is not a direct incentive for unwanted behaviors such as helping students during the test or other types of fraud. We therefore believe that things like cheating or teachers helping students to answer the test will not be a major source of bias in our results. The practice of teaching to the test, where teachers increase their efforts on tasks aimed at raising student's scores on the specific exam linked to the bonus, should not greatly affect our estimations. In this sense, if we assume that our results are not affected by other sources of bias, a positive result on treatment effects on test scores would imply something closer to true gains in human capital.

An important feature is that the "Prova Brasil" test is not compulsory. The municipal and state administrations can choose to participate or not. On this issue, Fernando et al., (2009) indicates that in 2005 and 2007, the first two editions, all state administrations participated and 99% of the municipal ones did so. A posterior research for the years of 2009 and 2011 indicated that this picture remained unchanged. A clarification can be added here. The decision to participate is made by the municipal and state administrations, the schools themselves do not have the authority to decide on this matter.

Some limitations have been observed in the first two years of the assessment. According to the INEP website, in 2005 and 2007 only the results for urban schools were made available at the school level. Therefore, our estimations will be done only for this group, the urban schools.

¹⁰ The bonuses are conditional on the results from an educational assessment run at the state level. As already mentioned, the bonus is linked to an exam implemented by the secretary of education of the state of São Paulo the System of Evaluation of School Achievement of the State of São Paulo or SARESP. The scales, the thresholds delimiting the below basic, basic, adequate and advanced categories and also the skills measured are the same on both the federal and state assessments.

Another problem is that the full data for 2005 has not yet been released. The only publicly available information is of a representative sample of schools. This will prove to be an important limitation of the study when trying to verify the parallel trends assumption required for a difference in difference estimation, especially because of the small sample size at hand. The strategies for dealing with this issue will be discussed later on.

The vast availability of data brought the question on how to define a control group. The municipally run schools of São Paulo came as a natural candidate given that any potential biases arising from confounding factors specific to the state of São Paulo would be mitigated. Moreover, this control group was also used by the previous studies of Oshiro et al., (2015) and Guimarães et al., (2017), allowing for a comparison of results on average test scores. Another control group composed only by state run schools would also be desirable given that there might be factors that influence our results that are specific to state administrations¹¹. First, state schools from the neighboring states of Minas Gerais and Paraná were chosen as potential controls. However, the propensity score matching procedure using only these two states produced poor results. The standardized differences between the covariates of the treatment and control groups were much higher than the ones deemed acceptable in the literature. Therefore, the sample was expanded to include the two states mentioned above plus the states of Santa Catarina, Rio Grande do Sul and Mato Grosso do Sul. This setting improved considerably the quality of the matching, with a substantial decrease in the standardized differences. Table 3 presents the sample sizes for each category by year.

It can be seen that there is a certain variation on the number of schools with data available between years, especially in the case of the municipally run schools of São Paulo. This variation can be partially attributed to the fact that some institutions did not had data on the professor and/or the principal's questionnaire in 2007. As an example, the municipally run schools of São Paulo do not have data on the principal's questionnaire for 339 schools while for the treatment group this happens in 103 cases. A choice was made not to use the schools with this type of missing data given that this would compromise the propensity score matching later on.

Table 3 - Number of schools in the sample

	Treatment: State run schools of São Paulo	Control: Municipally run schools of São Paulo	Control: State run schools of neighboring states
2005	57	48	387
2007	5.382	4.400	10.769
2009	5.417	5.093	10.280
2011	5.198	5.376	9.676

It can also be verified in table 3 the already mentioned issue of a small sample in 2005. While for the difference in differences estimation itself this would not pose as a problem, given that only one pre-treatment period (2007) would be enough, the situation gets more complex while trying to sustain the parallel trends assumption. After performing the matching on the propensity score for 2005, the treatment

¹¹ Although it is not clear what this potential factors could be given the high discretionary power of the state administrations to conduct their education policy, meaning that there might be no similarities between state run schools from different states conditional on them being state run.

and control groups were reduced to 14 schools each when considering the municipal schools of São Paulo as counterfactuals. This issue is further discussed in the next section.

Following now with the descriptive statistics, tables 4 and 5 present the average share of students below the basic and at or below the basic thresholds for math and Portuguese scores. Table 4 portrays a worrisome picture of the average achievements of students in math. While from 2005 to 2009 some progress has been made, with an overall decrease in the share of the low performing students, in 2011 this trend seems to have stabilized at what can be deemed as a high level of educational deficit. In 2011, 20% of the 5th grade students from state run schools of São Paulo were performing below the basic threshold, meaning that they did not yet had the minimum skills to move forward to the next grade. Looking now to those performing at or below basic, 51% of the students did not achieve what is considered to be the adequate level for their grade. The case of the 9th grade is even more troublesome. More than two thirds of the students were below the adequate threshold in 2011. The scenario for the two control groups is similar, with low levels of student's achievements.

Comparing the 5th and 9th grades, it can be seen that for all groups the later has a higher educational deficit. Oshiro et al., (2015) argues that one of the reasons for these discrepancies could be that throughout their school life, many students move forward to the next grade without having achieved an adequate level. The succession of this phenomena could lead to a higher accumulation of educational deficits on the 9th grade and a relative lower performance.

Table 4 - average share of students below basic or at or below basic in math scores

	Treatment: State run schools of São Paulo				Control: Municipally run schools of São Paulo				Control: State run schools of neighboring states			
	5th grade		9th grade		5th grade		9th grade		5th grade		9th grade	
	below basic	at or below basic	below basic	at or below basic	below basic	at or below basic	below basic	at or below basic	below basic	at or below basic	below basic	at or below basic
2005	47%	81%	48%	92%	48%	80%	41%	92%	33%	76%	31%	87%
2007	36%	75%	36%	90%	28%	68%	32%	87%	30%	72%	28%	87%
2009	19%	49%	26%	64%	17%	47%	25%	65%	16%	48%	20%	64%
2011	20%	51%	25%	70%	17%	49%	23%	70%	14%	46%	17%	66%

Table 5 illustrate the situation for achievements in Portuguese. A similar trend can be observed compared to the previous table. There was an overall decrease of those performing below basic or at or below basic from 2005 to 2009 and a relative stabilization of the trend in 2011. The picture portrayed in 2011 is a little bit better when compared to math, especially if we consider the 9th grade. At that period, 15% of the 9th graders from state run schools of São Paulo were performing below the basic threshold and 58% were at or below the basic level, a sizeable difference relative to the average math performance. The same analysis can be extended to our two control groups.

Table 5 - - average share of students below basic or at or below basic in Portuguese scores

	Treatment: State run schools of São Paulo		Control: Municipally run schools of São Paulo		Control: State run schools of neighboring states	
	5th grade	9th grade	5th grade	9th grade	5th grade	9th grade

	below basic	at or below basic	below basic	at or below basic	below basic	at or below basic	below basic	at or below basic	below basic	at or below basic	below basic	at or below basic
2005	30%	66%	32%	82%	31%	68%	24%	80%	23%	67%	20%	79%
2007	29%	70%	26%	82%	22%	64%	24%	80%	23%	68%	21%	81%
2009	18%	50%	15%	53%	15%	49%	14%	54%	14%	48%	10%	54%
2011	18%	50%	15%	58%	14%	47%	14%	59%	12%	43%	11%	57%

Comparing the treatment to our two control groups, it can be seen that the state-run schools of São Paulo have a higher concentration of students in the low performing thresholds in almost all years and grades in both subjects. If these differences are due to an unobserved characteristic of state run schools of São Paulo that is fixed over time, which could be case, the difference in difference estimation would mitigate any potential bias arising from this unobserved heterogeneity between treatment and control.

We continue now having a closer look at the information provided by the students, teachers and principals questionnaires. The amount of data available constituted both an advantage and a curse. As an example, our “cleaned” file containing the information on student’s characteristics had more than 2 million observations just for 2007 when considering our treatment and control groups. The amount of computer processing power and time required to run the estimations and especially the propensity score matching¹² at the student level seemed not feasible given the goals and guidelines of this study. Therefore, a choice was made to work at the school level. All the variables from students and teachers were aggregated by grade (5th or 9th), representing now the share of people with a given characteristic within school.

Table 6 describe the variables used to compare the treatment and control groups and implement the matching. The choice of covariates for estimating the propensity score should be guided by economic theory and previous research. Only variables that are said to affect simultaneously the probability of treatment and the outcome of interest should be included (Caliendo and Kopeinig, 2008).

Variables 1 to 11 are related to the student’s socioeconomic situation and correspond to the proportion of students with a given characteristic within school. There is extensive research showing that the background is an important determinant of performance and test scores. It is worth noting that the wealth index was created using principal component analysis and the information available on student’s assets at home¹³. The idea is that instead of just summing up all assets and therefore assigning an arbitrary weight, constructing an index using the first principal component allow us to summarize the variability among the assets and assign more appropriate weights to each variable (World Bank, technical note 4).

Variables 12 to 17 indicate teacher’s characteristics and their type of contract. There is a certain intuition in the idea that the behavior of teachers with a temporary contract is less impacted by a pay for performance scheme.

Variables 18 to 25 were computed using the head of the school answers to the questionnaire. It contains more general information about the school such as infrastructure and underlying problems. There is an

¹² Some of the matching attempts at the school level took more than an hour in Stata

¹³ Information on the following items was provided in the questionnaire and used in the index: television, radio, dvd, fridge, washing machine, vacuum cleaner, car, computer, freezer and domestic worker

extensive literature on the impacts of school infrastructure on student's performance, although the results seem to be inconclusive.

Table 6 - variables selected for estimating the propensity score

	Variable	Description
	Student level	
1	book 1_20	share of students with 1 to 20 books at home
2	book 21_100	share of students with 21 to 100 books at home
3	book 101	share of students with more than 100 books at home
4	mom - low education	share of students whose mother have not studied or completed 4th grade
5	mom - 8th grade	share of students whose mother completed 8th grade
6	mom - high school	share of students whose mother completed high school
7	wealth index	wealth score index using principal component analysis
8	preschool attendance	share of students that attended preschool
9	work	share of students who work
10	average math	average math score
11	average Portuguese	average Portuguese score
	Teacher level	
12	teachers - no degree	share of teachers with no university degree
13	teachers - in person degree	share of teachers who obtained a degree that is not online
14	teachers - specialization	share of teachers who have some kind of specialization
15	teachers - average wage	teachers average wage from school
16	teachers - stable contract	share of teachers with the most stable type contract
17	teachers - temporary contract	share of teachers with temporary contract
	Principal level	
18	director - no specialization	dummy for director that have no specialization
19	director - wage	director wage from school
20	director - stable contract	share of teachers with a stable contract (difficult to fire) in the whole school
21	director - financial resources	dummy for school not experiencing financial shortages
22	director - turnover teachers	share of schools where teacher's turnover was not a problem
23	director – no computer students	share of schools with no computers available for students
24	director - no library	share of schools without a library
25	director - no laboratory	share of schools without a laboratory

Tables 7 and 8 allow us to compare the underlying differences between treatment and control. It is presented the average outcome for each covariate together with the standardized percentage bias and the t-test for the equality of means between the two groups. A first glance at the differences between municipal and state schools in table 7 indicate that the 9th grade present more similarities between the two groups when compared to the 5th. While for the higher grade we cannot reject the null hypothesis of equality between means for 8 of the 25 variables at a 95% level, for the 5th grade only for the share of students with between 21 and 100 books the null hypothesis is not rejected.

A closer look at the 5th grade indicate that while differences are observed, there does not seem to be a major divergence in terms of characteristics between municipal and state schools. The average performance of students seems to be higher on municipal schools, with a 10 points difference in math and 8 in Portuguese. The share of teachers with some kind of specialization is 13 percentage points higher on the control group while the share of those with a temporary contract is almost 7 points lower. Looking now to the variables computed using the principal's questionnaire it can be seen that on average the share of teachers with a stable contract is 13 percentage points higher in municipal schools. It is worth noting that this variable is a measure for the school as whole and not only for the grades in question. The share of schools which did not experience any kind of financial shortages is considerably higher for the control group, with a 34 percentage point's difference. The problem of teacher's turnover is also less prominent on municipal schools, with a 16 percentage point's divergence. Regarding the infrastructure, state schools seem to have on average more computers available but less libraries when compared to the municipal ones.

The differences pointed out for the 5th grade are similar to those observed for the 9th. Generally, it seems that municipal schools outperform the state ones in a few variables that could propitiate better student's achievements. The propensity score matching implemented later on is one of the tools used to tackle this potential source of bias.

Table 7 - Comparison between state run schools and municipally run schools from São Paulo

Variable	5th Grade				9th Grade			
	Mean		%bias	p> t	Mean		%bias	p> t
	Treated	Control			Treated	Control		
book 1_20	0.523	0.529	-7.9	0.007	0.602	0.603	-2.0	0.569
book 21_100	0.155	0.152	4.4	0.135	0.183	0.179	6.1	0.079
book 101	0.089	0.082	14.3	0.000	0.055	0.050	15.9	0.000
mom 8th grade	0.140	0.136	8.1	0.006	0.184	0.178	10.3	0.004
mom high school	0.117	0.110	13.1	0.000	0.204	0.198	6.3	0.076
preschool	0.338	0.362	-15.6	0.000	0.421	0.399	14.0	0.000
work	0.131	0.115	25.4	0.000	0.183	0.179	5.2	0.136
mom low education	0.071	0.079	-15.4	0.000	0.111	0.112	-2.2	0.538
wealth index	0.082	0.001	16.6	0.000	0.082	0.001	7.8	0.026
average math	193.07	203.55	-59.0	0.000	242.08	247.35	-33.8	0.000
average Portuguese	176.05	184.16	-50.9	0.000	231.15	234.69	-24.3	0.000
teacher - no degree	0.124	0.101	11.2	0.000	0.002	0.002	-0.0	0.995
teacher - in person degree	0.827	0.845	-7.3	0.012	0.982	0.982	1.1	0.742
teacher - specialization	0.301	0.431	-42.1	0.000	0.426	0.561	-42.8	0.000
teacher - average wage	1367.2	1327.8	11.3	0.000	1523.7	1604.9	-17.3	0.000
teacher - temporary contract	0.188	0.120	27.4	0.000	0.148	0.129	8.1	0.024
teacher - stable contract	0.527	0.498	7.7	0.009	0.615	0.557	16.2	0.000
director - no specialization	0.255	0.314	-13.0	0.000	0.231	0.302	-15.9	0.000
director - wage	2419.9	2282.9	24.0	0.000	2450.4	2494.9	-7.7	0.022
director - stable contract	0.576	0.707	-46.7	0.000	0.536	0.685	-57.2	0.000
director - financial resources	0.305	0.637	-70.5	0.000	0.293	0.683	-84.8	0.000

director - turnover teachers	0.611	0.768	-34.5	0.000	0.480	0.718	-49.9	0.000
director - no computer for students	0.091	0.280	-50.1	0.000	0.055	0.148	-31.0	0.000
director - no library	0.231	0.179	13.0	0.000	0.087	0.085	1.0	0.783
director - no laboratory	0.873	0.834	10.9	0.000	0.585	0.654	-14.4	0.000

Table 8 presents the characteristics of the control group composed by the state schools on the vicinity of São Paulo. It can be seen that for the 5th grade, the distribution of the number of books at home is similar between treatment and control, with the majority of students having between 1 and 20 books. The students of the state schools of São Paulo have mothers with a slightly higher education with 14% having completed the 8th grade and 11% high school. For our control group this numbers go down to 12% and 9%, respectively. The average scores are again lower for the treated, although with smaller difference between the two groups when compared to the previous table.

Looking now to teacher's characteristics, it is observed that the control group resorted more to university degrees that have at least some kind of distance learning. On the other hand, this group was the one with a higher share of professors having some kind of specialization, 49% for the control against 30% for the treated. Another difference that can be highlighted is the higher average salary of the teachers in the state of São Paulo, which corresponds to R\$ 1.367 against R\$1.000 in the neighboring states. The control group seems to resort more to more stable contracts. While 70% of teachers enjoy this benefit in the neighboring states, only 52% do so in São Paulo. Regarding the infrastructure, it seems that the treated have a considerable higher amount of schools with computers available for their students. On the other hand, the state schools from São Paulo seems to suffer from a lack of libraries when compared to neighboring states. While in the control only 7% do not have this kind of structure, in the treatment this figure rises to 23%.

For the 9th grade similar divergences are observed between treatment and control. It can be pointed that the average difference in the mother's education is a bit higher for this group. While for the treated the share of mothers who completed the 8th grade and high school are respectively 18% and 20%, for the control group the figures go down to 14% and 16%, respectively. Another difference that can be mentioned is that while 42% of students in the state schools of São Paulo have participated in preschool, only 32% have done so in the control group.

Table 8 - Comparison between state run schools of São Paulo and neighboring states

Variable	5th Grade				9th Grade			
	Mean		%bias	p> t	Mean		%bias	p> t
	Treated	Control			Treated	Control		
book 1_20	0.523	0.519	3.8	0.208	0.602	0.602	-0.7	0.757
book 21_100	0.155	0.164	-12.4	0.000	0.183	0.171	16.1	0.000
book 101	0.089	0.092	-5.6	0.061	0.055	0.049	17.2	0.000
mom 8th grade	0.140	0.121	33.4	0.000	0.184	0.149	55.3	0.000
mom high school	0.117	0.091	42.1	0.000	0.204	0.165	39.3	0.000
preschool	0.338	0.299	24.3	0.000	0.421	0.326	54.6	0.000
work	0.131	0.117	19.2	0.000	0.183	0.221	-40.3	0.000
mom low education	0.071	0.079	-13.6	0.000	0.111	0.126	-19.5	0.000

wealth index	0.082	0.001	43.5	0.000	0.082	0.001	34.5	0.000
average math	193.07	199.19	-35.9	0.000	242.08	250.13	-51.7	0.000
average Portuguese	176.05	180.91	-30.7	0.000	231.15	235.26	-28.6	0.000
teacher - no degree	0.124	0.127	-1.4	0.625	0.002	0.006	-9.6	0.000
teacher - in person degree	0.827	0.638	57.7	0.000	0.982	0.880	57.6	0.000
teacher - specialization	0.301	0.496	-55.0	0.000	0.426	0.655	-71.3	0.000
teacher - average wage	1367.2	1004.1	99.6	0.000	1523.7	1119.6	104.6	0.000
teacher - temporary contract	0.188	0.192	-1.3	0.655	0.148	0.215	-26.1	0.000
teacher - stable contract	0.527	0.707	-51.1	0.000	0.615	0.717	-31.6	0.000
director - no specialization	0.255	0.241	3.4	0.218	0.231	0.185	11.5	0.000
director - wage	2419.9	2073.8	63.2	0.000	2450.4	2169.1	50.5	0.000
director - stable contract	0.576	0.645	-24.4	0.000	0.536	0.623	-32.9	0.000
director - financial resources	0.305	0.197	25.0	0.000	0.293	0.218	17.3	0.000
director - turnover teachers	0.611	0.585	5.2	0.061	0.480	0.522	-8.3	0.000
director - no computer for students	0.091	0.329	-61.2	0.000	0.055	0.256	-57.6	0.000
director - no library	0.231	0.073	45.0	0.000	0.087	0.048	15.7	0.000
director - no laboratory	0.873	0.617	61.3	0.000	0.585	0.490	19.0	0.000

It is unclear whether the net differences between treatment and control would be more conducive of better student's achievements in one of the two groups. In any case, the matching on the propensity score should mitigate any potential bias arising from these underlying differences.

7. Results

7.1. Matching

We proceed now with the estimation of the propensity score followed by the matching. The propensity score was estimated using a probit model, which is the default option in the `psmatch2` command in Stata. According to Caliendo and Kopeinig (2008), probit and logit models should be preferred when estimating the propensity score when compared to linear probability models, where the predictions can move outside of the (0, 1) boundaries. When treatment assignment is binary, which is the case of this study, there is no clear guidelines for which of the 2 models should be preferred. We therefore choose the probit estimator for convenience.

The estimations of the propensity score, i.e., the probability of being treated are reported in the appendix. It can be noted that most of the variables were significant on estimating the probability of being treated. In the cases where they were not significant for a given control group or grade, they were successful in predicting treatment assignment for another.

Following with the procedure, the matching algorithm chosen was by nearest neighbor without replacement. In this specification, each treated school is matched with an institution from the control that has the closest propensity score. The choice of no replacement implies that each untreated school can only be used once for a match. The main reason for choosing this algorithm is that it was also the method employed in the studies of Oshiro et al., (2015) and Guimarães et al., (2017), the 2 works identified that also evaluate the program in question. This will allow for a better comparability of the results.

An important exercise to verify the quality of the matching is to observe the differences between the covariates in the treatment and control groups. An often-used measure to assess these divergences is the standardized differences or standardized bias. Unlike the t-test, this measure is not impacted by sample size, which allows for a better comparison between the matched/unmatched and treated/untreated samples (Austin, 2009). It is given by:

$$d = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

Where $\bar{x}_{treatment}$ and $\bar{x}_{control}$ are respectively the sample mean of the treated and untreated covariate in question and $s^2_{treatment}$ and $s^2_{control}$ denotes the variance of the variable in each of the groups. A shortcoming of this measure is that there does not seem to be a consensus regarding the amount of “differences” that should be tolerated. Austin (2009) points that some researches propose a threshold of 0,1 (10%), while Harder et al., (2010) indicate a limit as high as 25%. In this study we were able to maintain the standardized differences lower than the 25% threshold for almost all variables in each specification. This was mainly done by increasing the amount of states used in our control group of state run schools, an issue explained in the previous section.

Tables 9 to 12 present the means of each variable for the matched/unmatched and treated/untreated groups, together with the corresponding t-tests and percentage standardized differences. A visual representation of the standardized bias between treatment and controls can be seen in the appendix.

Table 9 shows the results for 5th grade using municipally run schools as controls. It is observed that the matching greatly reduced the bias between state run and municipally run schools. Out of 25 covariates 9 remained with a standardized difference above the 10% threshold. This is the case for the average test scores that seem to be consistently lower in the treatment group. Using now the 25% threshold as a reference, the variable director – stable contract, which indicates the share of teachers in the whole school that enjoy the most stable type of contract, remained with an absolute standardized difference of 28% after the matching. This indicates that the municipally run schools of São Paulo resort more too stable contracts than our treatment group. The other variable that surpassed the higher threshold was the indicator of the share of schools who had not experienced some kind of financial difficulties. The municipal schools seem to be much less affected by this problem leading to a high standardized difference after the matching, of 42% in absolute terms.

The matching for the 9th was considerably better when compared to the 5th grade. Table 10 indicates that only for the variables director – stable contract and director – no computer for students the standardized differences surpassed slightly the 10% threshold.

The matching using state run schools of neighboring states was considerably worse when compared to our previous control group. Table 11 shows that 7 variables surpassed the 25% limit when matching fifth grade schools. As already seen in the previous section, the average salary of teachers in this control group is systematically lower than in the state schools of São Paulo. The matching was also not successful in reducing the differences between schools' infrastructures. It is clear that the state-run schools of São Paulo put a greater emphasis on making computers available for students when compared to neighboring states. On the other hand, the share of schools with no libraries or laboratories is much higher for out treatment group. Table 12 portrays a similar scenario for the 9th grade with now 8 covariates going above the 25% limit. It can be noted that differences in terms of infrastructure for this grade are reduced but stay above the 10% threshold.

Table 9 - Comparison of covariates before and after matching - 5th grade with municipally run schools as controls

Variable	Unmatched	Mean		% Reduction		t-test	
	Matched	Treated	Control	%bias	bias	t	p> t
book 1_20	U	0,523	0,529	-7.9		-2.68	0.007
	M	0,523	0,527	-5.0	36.4	-1.57	0.117
book 21_100	U	0,155	0,152	4.4		1.50	0.135
	M	0,155	0,154	1.2	73.1	0.38	0.707
book 101	U	0,089	0,082	14.3		4.86	0.000
	M	0,089	0,086	6.3	56.1	1.97	0.049
mom 8th grade	U	0,140	0,136	8.1		2.77	0.006
	M	0,140	0,139	1.8	77.2	0.58	0.565
mom high school	U	0,117	0,110	13.1		4.50	0.000
	M	0,117	0,115	4.5	65.4	1.37	0.171
preschool	U	0,338	0,362	-15.6		-5.29	0.000
	M	0,338	0,350	-7.8	49.8	-2.50	0.013
work	U	0,131	0,115	25.4		8.67	0.000
	M	0,131	0,121	15.6	38.6	4.75	0.000

mom low education	U	0,071	0,079	-15.4		-5.23	0.000
	M	0,071	0,072	-1.9	87.4	-0.62	0.532
wealth index	U	0,082	0,001	16.6		5.68	0.000
	M	0,082	0,060	4.5	72.5	1.41	0.157
average math	U	193.07	203.55	-59.0		-19.70	0.000
	M	193.07	197.33	-24.0	59.4	-8.11	0.000
average Portuguese	U	176.05	184.16	-50.9		-17.25	0.000
	M	176.05	179.6	-22.2	56.3	-7.04	0.000
teacher - no degree	U	0,124	0,101	11.2		3.87	0.000
	M	0,124	0,109	7.2	35.5	2.16	0.031
teacher - in person degree	U	0,827	0,845	-7.3		-2.51	0.012
	M	0,827	0,841	-5.8	20.8	-1.76	0.079
teacher - specialization	U	0,301	0,431	-42.1		-14.20	0.000
	M	0,301	0,372	-23.1	45.0	-7.36	0.000
teacher - average wage	U	1367,2	1327,8	11.3		3.66	0.000
	M	1367,2	1382,0	-4.2	62.5	-1.23	0.219
teacher - temporary contract	U	0,188	0,120	27.4		9.58	0.000
	M	0,188	0,136	20.8	23.9	6.14	0.000
teacher - stable contract	U	0,527	0,498	7.7		2.61	0.009
	M	0,527	0,542	-4.3	44.6	-1.33	0.185
director - no specialization	U	0,255	0,314	-13.0		-4.46	0.000
	M	0,255	0,280	-5.4	58.9	-1.67	0.096
director - wage	U	2419,9	2282,9	24.0		8.01	0.000
	M	2419,9	2421,8	-0.3	98.6	-0.10	0.918
director - stable contract	U	0,576	0,707	-46.7		-16.36	0.000
	M	0,576	0,655	-28.0	40.0	-8.41	0.000
director - financial resources	U	0,305	0,637	-70.5		-24.13	0.000
	M	0,305	0,503	-42.0	40.3	-12.57	0.000
director - turnover teachers	U	0,611	0,768	-34.5		-12.13	0.000
	M	0,611	0,696	-18.8	45.5	-5.53	0.000
director – no computer for students	U	0,091	0,280	-50.1		-16.36	0.000
	M	0,091	0,142	-13.7	72.7	-4.92	0.000
director - no library	U	0,231	0,179	13.0		4.54	0.000
	M	0,231	0,192	9.9	24.3	2.97	0.003
director - no laboratory	U	0,873	0,834	10.9		3.70	0.000
	M	0,873	0,837	10.0	8.0	3.07	0.002

Table 10 - Comparison of covariates before and after matching - 9th grade with municipally run schools as controls

Variable	Unmatched		Mean		% Reduction	t-test	
	Matched	Treated	Control	%bias		t	p> t

book 1_20	U	0,602	0,603	-2.0		-0.57	0.569
	M	0,603	0,603	-0.1	96.5	-0.02	0.988
book 21_100	U	0,183	0,179	6.1		1.76	0.079
	M	0,178	0,179	-0.3	95.7	-0.06	0.954
book 101	U	0,055	0,050	15.9		4.52	0.000
	M	0,051	0,050	3.6	77.2	0.81	0.417
mom 8th grade	U	0,184	0,178	10.3		2.92	0.004
	M	0,180	0,178	3.4	67.3	0.74	0.459
mom high school	U	0,204	0,198	6.3		1.77	0.076
	M	0,199	0,198	0.6	90.2	0.13	0.893
preschool	U	0,421	0,399	14.0		3.89	0.000
	M	0,403	0,399	2.3	83.4	0.50	0.614
work	U	0,183	0,179	5.2		1.49	0.136
	M	0,182	0,179	4.5	13.6	1.00	0.316
mom low education	U	0,111	0,112	-2.2		-0.62	0.538
	M	0,112	0,112	-0.7	69.8	-0.14	0.886
wealth index	U	0,082	0,001	7.8		2.23	0.026
	M	0,082	0,060	1.5	80.8	0.33	0.743
average math	U	242.08	247.35	-33.8		-9.92	0.000
	M	245.94	247.35	-9.1	73.1	-2.00	0.046
average Portuguese	U	231.15	234.69	-24.3		-7.01	0.000
	M	233.6	234.69	-7.5	69.2	-1.66	0.098
teacher - no degree	U	0,002	0,002	-0.0		-0.01	0.995
	M	0,002	0,002	1.6	-6975.9	0.31	0.757
teacher – in person degree	U	0,982	0,982	1.1		0.33	0.742
	M	0,981	0,982	-1.2	-6.7	-0.26	0.792
teacher - specialization	U	0,426	0,561	-42.8		-12.17	0.000
	M	0,534	0,561	-8.4	80.4	-1.84	0.065
teacher - average wage	U	1523,7	1604,9	-17.3		-5.65	0.000
	M	1572,1	1604,9	-7.0	59.5	-1.55	0.122
teacher - temporary contract	U	0,148	0,129	8.1		2.27	0.024
	M	0,138	0,129	3.7	54.1	0.84	0.402
teacher - stable contract	U	0,615	0,557	16.2		4.73	0.000
	M	0,588	0,557	8.7	46.6	1.91	0.056
director - no specialization	U	0,231	0,302	-15.9		-4.48	0.000
	M	0,296	0,302	-1.4	91.3	-0.30	0.767
director - wage	U	2450,4	2494,9	-7.7		-2.29	0.022
	M	2539,3	2494,9	7.7	0.4	1.70	0.090
director - stable contract	U	0,536	0,685	-57.2		-15.47	0.000
	M	0,650	0,685	-13.3	76.8	-3.14	0.002
	U	0,293	0,683	-84.8		-23.44	0.000

director - financial resources	M	0,703	0,683	4.4	94.8	0.98	0.328
director - turnover teachers	U	0,480	0,718	-49.9		-13.34	0.000
	M	0,690	0,718	-5.8	88.4	-1.34	0.182
director – no computer for students	U	0,055	0,148	-31.0		-9.65	0.000
	M	0,113	0,148	-11.6	62.5	-2.28	0.023
director - no library	U	0,087	0,085	1.0		0.28	0.783
	M	0,088	0,085	1.1	-8.2	0.24	0.809
director - no laboratory	U	0,585	0,654	-14.4		-3.93	0.000
	M	0,630	0,654	-5.1	65.0	-1.13	0.259

Table 11 - Comparison of covariates before and after matching - 5th grade with state run schools as controls

Variable	Unmatched	Mean		% Reduction		t-test	
	Matched	Treated	Control	%bias	bias	t	p> t
book 1_20	U	0,523	0,519	3.8		1.26	0.208
	M	0,523	0,526	-3.1	19.2	-1.02	0.307
book 21_100	U	0,155	0,164	-12.4		-4.13	0.000
	M	0,155	0,162	-9.1	26.6	-2.93	0.003
book 101	U	0,089	0,092	-5.6		-1.87	0.061
	M	0,089	0,092	-5.6	-1.0	-1.79	0.074
mom 8th grade	U	0,140	0,121	33.4		11.35	0.000
	M	0,140	0,133	12.5	62.7	3.83	0.000
mom high school	U	0,117	0,091	42.1		14.85	0.000
	M	0,117	0,107	17.3	58.9	5.28	0.000
preschool	U	0,338	0,299	24.3		8.51	0.000
	M	0,338	0,327	6.5	73.2	2.03	0.042
work	U	0,131	0,117	19.2		6.55	0.000
	M	0,131	0,121	13.6	29.2	4.10	0.000
mom low education	U	0,071	0,079	-13.6		-4.56	0.000
	M	0,071	0,073	-3.8	71.8	-1.26	0.207
wealth index	U	0,082	0,001	43.5		14.03	0.000
	M	0,082	0,060	10.1	76.8	3.85	0.000
average math	U	193.07	199.19	-35.9		-12.57	0.000
	M	193.07	196.74	-21.6	40.0	-6.81	0.000
average Portuguese	U	176.05	180.91	-30.7		-10.86	0.000
	M	176.05	179.1	-19.3	37.2	-5.96	0.000
teacher - no degree	U	0,124	0,127	-1.4		-0.49	0.625
	M	0,124	0,133	-3.6	-155.4	-1.09	0.275
teacher – in person degree	U	0,827	0,638	57.7		19.09	0.000
	M	0,827	0,763	19.5	66.3	6.62	0.000
teacher - specialization	U	0,301	0,496	-55.0		-18.61	0.000

	M	0,301	0,441	-39.6	28.0	-12.05	0.000
teacher - average wage	U	1367,2	1004,1	99.6		32.63	0.000
	M	1367,2	1146,0	60.7	39.1	16.55	0.000
teacher - temporary contract	U	0,188	0,192	-1.3		-0.45	0.655
	M	0,188	0,211	-7.7	-500.5	-2.31	0.021
teacher - stable contract	U	0,527	0,707	-51.1		-18.13	0.000
	M	0,527	0,633	-30.1	41.1	-8.87	0.000
director - no specialization	U	0,255	0,241	3.4		1.23	0.218
	M	0,255	0,227	6.6	-94.1	2.03	0.042
director - wage	U	2419.9	2073.8	63.2		22.08	0.000
	M	2419.9	2227.5	35.1	44.4	10.87	0.000
director - stable contract	U	0,576	0,645	-24.4		-9.01	0.000
	M	0,576	0,620	-15.5	36.4	-4.72	0.000
director - financial resources	U	0,305	0,197	25.0		9.31	0.000
	M	0,305	0,220	19.7	21.1	5.91	0.000
director - turnover teachers	U	0,611	0,585	5.2		1.87	0.061
	M	0,611	0,592	3.7	28.3	1.14	0.256
director – no computer for students	U	0,091	0,329	-61.2		-20.33	0.000
	M	0,091	0,191	-25.8	57.9	-8.90	0.000
director - no library	U	0,231	0,073	45.0		18.00	0.000
	M	0,231	0,101	37.2	17.4	10.86	0.000
director - no laboratory	U	0,873	0,617	61.3		20.72	0.000
	M	0,873	0,703	40.8	33.4	12.98	0.000

Table 12 - Comparison of covariates before and after matching - 9th grade with state run schools as controls

Variable	Unmatched	Mean		% Reduction		t-test	
	Matched	Treated	Control	%bias	bias	t	p> t
book 1_20	U	0,602	0,602	-0.7		-0.31	0.757
	M	0,602	0,601	0.4	43.1	0.17	0.863
book 21_100	U	0,183	0,171	16.1		7.13	0.000
	M	0,183	0,178	5.8	63.8	2.43	0.015
book 101	U	0,055	0,049	17.2		7.67	0.000
	M	0,055	0,053	5.3	69.5	2.10	0.036
mom 8 th grade	U	0,184	0,149	55.3		24.22	0.000
	M	0,184	0,166	28.3	48.8	11.84	0.000
mom high school	U	0,204	0,165	39.3		17.63	0.000
	M	0,204	0,188	16.8	57.3	6.81	0.000
preschool	U	0,421	0,326	54.6		24.33	0.000
	M	0,421	0,371	29.0	46.9	11.88	0.000
work	U	0,183	0,221	-40.3		-17.66	0.000

	M	0,183	0,202	-20.5	49.1	-8.89	0.000
mom low education	U	0,111	0,126	-19.5		-8.47	0.000
	M	0,111	0,115	-5.5	71.5	-2.40	0.017
wealth index	U	0,082	0,001	34.5		14.71	0.000
	M	0,082	0,060	8.9	74.1	4.00	0.000
average math	U	242.08	250.13	-51.7		-22.98	0.000
	M	242.08	247.7	-36.0	30.2	-15.33	0.000
average Portuguese	U	231.15	235.26	-28.6		-12.85	0.000
	M	231.15	234.47	-23.1	19.3	-9.59	0.000
teacher - no degree	U	0,002	0,006	-9.6		-4.04	0.000
	M	0,002	0,004	-4.8	50.2	-2.32	0.021
teacher – in person degree	U	0,982	0,880	57.6		23.73	0.000
	M	0,982	0,954	15.8	72.6	10.60	0.000
teacher - specialization	U	0,426	0,655	-71.3		-32.02	0.000
	M	0,426	0,611	-57.7	19.1	-22.96	0.000
teacher - average wage	U	1523.7	1119.6	104.6		45.94	0.000
	M	1523.7	1253.4	70.0	33.1	27.29	0.000
teacher - temporary contract	U	0,148	0,215	-26.1		-11.56	0.000
	M	0,148	0,198	-19.7	24.6	-8.23	0.000
teacher - stable contract	U	0,615	0,717	-31.6		-14.50	0.000
	M	0,615	0,709	-29.1	7.7	-11.80	0.000
director - no specialization	U	0,231	0,185	11.5		5.35	0.000
	M	0,231	0,176	13.7	-18.6	5.57	0.000
director - wage	U	2450.4	2169.1	50.5		22.54	0.000
	M	2450.4	2296.1	27.7	45.1	11.45	0.000
director - stable contract	U	0,536	0,623	-32.9		-15.20	0.000
	M	0,536	0,625	-33.8	-2.5	-13.76	0.000
director - financial resources	U	0,293	0,218	17.3		8.05	0.000
	M	0,293	0,234	13.4	22.1	5.37	0.000
director - turnover teachers	U	0,480	0,522	-8.3		-3.82	0.000
	M	0,480	0,509	-5.7	31.5	-2.30	0.021
director – no computer for students	U	0,055	0,256	-57.6		-24.50	0.000
	M	0,055	0,140	-24.3	57.8	-11.65	0.000
director - no library	U	0,087	0,048	15.7		7.51	0.000
	M	0,087	0,050	15.0	4.3	6.03	0.000
director - no laboratory	U	0,585	0,490	19.0		8.70	0.000
	M	0,585	0,467	23.8	-25.3	9.62	0.000

7.2. Regressions

As already discussed in the methodological section, the parallel trends between treatment and control groups on a given outcome is a fundamental assumption when running a difference in differences estimation. One of the ways of analyzing the issue is by visually inspecting the trends in both groups. Given the amount of outcomes estimated in this study we decided to attach the graphs of the average outcomes over the years in the appendix. The data limitation in 2005 lead us to produce the charts using the full sample with no matching. As already stated, implementing the matching in 2005 would reduce our sample to 14 schools in each group.

It can be noted that the visual inspection of the graphs was not very reassuring on verifying if the parallel trends assumptions holds. In many cases the trends in treatment and control are not parallel before the intervention. After 2007, when the intervention started, many of the outcomes for the control group do not maintain the trajectory observed in the pre-treatment periods (2005-2007), a behavior that would be desirable for an unbiased difference in differences estimation.

We highlight some of the graphs that presented more reassuring trends. Figure 10 presents the average share of students at or below basic for 5th graders using municipal schools as controls. In the case of math, it can be seen that before the intervention treatment and control moved relatively in tandem. After 2007, municipal schools continued on their trend and the treated observed a sharper decrease in the share of those performing at or below the basic, indicating that the policy had the desired effect. Figure 11 shows that average math test scores of 9th graders in municipal and state schools of São Paulo moved in parallel before the intervention. After 2007 it seems that some kind of shock broke the growth trend in schools from both jurisdictions. However, the break was less pronounced in the treatment group, indicating that the policy might have contributed to attenuate the shock. Figure 14 points to a similar scenario to the one just described above. The average share of 9th grade students at or below basic in math and Portuguese in municipal and state schools of São Paulo moved in tandem before the intervention. After 2007 some kind of shock seem to have decreased this outcome in both groups but the decline was sharper in state schools, indicating that the program might have had the desired effect. Similar behaviors can be observed in figures 17 in both subjects, 18 in the case of math.

In order to shed more light into the parallel trends assumption, we perform a placebo difference in differences. In this setting, the treatment year will be 2007 (placebo) and the pre-treatment period 2005. The idea is that if no significant results are found, the common trends assumption is more likely to hold. On the contrary, if a significant impact is found for the placebo period, it is likely that any effects observed on the estimations are biased from underlying differences in trends between treatment and control groups (Gertler et al., 2016).

The placebo test is performed using our 2 control groups, municipal schools and state run schools from neighboring states, for the 5th and 9th grades. Given our data limitation on 2005 and in order to keep our sample as large as possible, no matching methods were implemented on this exercise. Table 13 reports the results for the 5th grade using municipal schools as controls.

Table 13 - Placebo regression for the 5th grade using municipally run schools as controls

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2007	-11.16** (5.311)	-9.063* (4.709)	0.0140 (0.00861)	0.00522 (0.00941)	8.399** (3.950)	7.644** (3.582)	6.567 (4.184)	8.285** (4.163)
Observations	5,379	5,379	5,379	5,379	5,379	5,379	5,379	5,379
R-squared	0.078	0.056	0.038	0.058	0.084	0.066	0.061	0.040
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								

It can be seen that many of our outcomes of interest have significant coefficients at a 95% level. The results indicate that the municipally run schools of São Paulo might not be the best group to be used as a counterfactual for our estimations on the 5th grade. We proceed with the placebo analysis using now the state run control group.

Table 14 - Placebo regression for the 5th grade using state run schools of neighboring as controls

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2007	7.280* (3.868)	0.0154 (3.566)	-0.0169** (0.00707)	-0.0136* (0.00763)	-7.820** (3.272)	-1.150 (2.898)	-2.152 (3.257)	2.681 (3.318)
Observations	6,747	6,747	6,747	6,747	6,747	6,747	6,747	6,747
R-squared	0.027	0.019	0.125	0.129	0.042	0.041	0.010	0.004
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								

In this case we can observe an improvement in the results. The coefficient for average Portuguese scores becomes non-significant and for math it becomes significant only at a 10% level. For the share of low performing students, the only significant coefficient is now for the share of those performing below the basic threshold in math. The estimation suggests that the state-run control group is a better counterfactual for the 5th grade.

We continue now with the placebo test for the 9th grade with municipal schools as controls. Table 15 shows that significant effects are found only for the share of students performing below the basic threshold in Portuguese and for the coefficient of variation of Portuguese.

Table 15 - Placebo regression for the 9th grade using municipally run schools as controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2007	2.621 (4.357)	4.325 (4.134)	-0.00483 (0.00548)	-0.0130** (0.00610)	-3.304 (3.653)	-6.251** (3.123)	2.713 (2.391)	-0.567 (2.768)
Observations	4,497	4,497	4,497	4,497	4,497	4,497	4,497	4,497
R-squared	0.028	0.012	0.005	0.009	0.025	0.012	0.026	0.009
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								

Looking to the second control group, table 16 indicates that state run schools from neighboring states are not likely to serve as a good counterfactual. Almost all the outcomes of interest present significant results.

Table 16 - Placebo regression for the 9th grade using state run schools of neighboring as controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2007	10.06*** (3.703)	6.416* (3.396)	-0.00726 (0.00581)	-0.0132** (0.00552)	-9.455*** (3.106)	-7.262*** (2.633)	-2.580 (2.170)	-2.775 (2.358)
Observations	9,765	9,765	9,765	9,765	9,765	9,765	9,765	9,765
R-squared	0.058	0.019	0.039	0.106	0.072	0.047	0.022	0.001
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								

The placebo estimations indicated that the parallel trends assumption is more likely to hold depending on the control group and grade analyzed. For the 5th grader's, the best counterfactual group seems to be the state-run schools for the states in the vicinity of São Paulo. On the other hand, in the case of the 9th grade, the group of municipal schools fit better as a counterfactual.

We proceed now with the DD regressions for the actual treatment years, 2009 and 2011. First, the tables containing the results using the whole data available will be presented. In this case, no matching method was implemented. It is understood that since the choice of a placebo test without matching in 2005 was made, it will be more coherent to also present the results with the full data. Afterwards, the estimations with the propensity score matching are shown.

Table 17 indicates that the program had positive impacts on 5th grade students when the municipal schools of São Paulo are used as controls. Considering the regressions without matching first, the average test scores increased by almost 7 points in mathematics and 4 points in Portuguese in 2009. The results for 2011 were also positive and significant, with a slight decrease.

More impressive were the outcomes for the share of students below the basic and at or below the basic threshold. For the former, there is an indication that the program decreased the share of students performing below basic by almost 6% in math and 4% in Portuguese in 2009. The results for the latter for the same year were slightly higher in absolute terms and also significant. In 2011 the scenario was similar with a slight decrease on the impact.

Table 17 - Difference in differences estimation for the 5th grade - municipally run schools as controls

5th grade - control: municipal schools not matched								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2009	6.921*** (0.775)	4.692*** (0.654)	-0.00191* (0.00107)	-0.00778*** (0.00112)	-5.962*** (0.461)	-4.368*** (0.401)	-6.142*** (0.572)	-4.838*** (0.528)
treated2011	5.562*** (0.787)	2.944*** (0.664)	0.000963 (0.00109)	-0.00416*** (0.00114)	-4.967*** (0.468)	-3.502*** (0.407)	-5.250*** (0.581)	-3.253*** (0.536)
Observations	16,629	16,629	16,628	16,629	16,639	16,639	16,639	16,639
R-squared	0.165	0.143	0.081	0.052	0.258	0.174	0.347	0.295
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								
5th grade - control: municipal schools matched								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2009	5.061*** (0.831)	2.904*** (0.734)	7.44e-05 (0.00122)	-0.00269** (0.00128)	-3.754*** (0.532)	-2.217*** (0.476)	-4.231*** (0.610)	-2.991*** (0.588)
treated2011	2.921*** (0.849)	0.452 (0.750)	0.00491*** (0.00124)	0.00156 (0.00131)	-2.537*** (0.544)	-1.141** (0.487)	-2.921*** (0.623)	-0.911 (0.601)
Observations	10,386	10,386	10,385	10,386	10,390	10,390	10,390	10,390
R-squared	0.220	0.174	0.083	0.040	0.313	0.205	0.441	0.362
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								

When looking to the matched samples the results are similar but of a lower magnitude. As an example, the impact on the share of students performing below basic decreased by nearly 2 percentage points in both math and Portuguese. A similar trend can be observed for outcomes 7 and 8. It is also worth noting that the coefficients on average Portuguese scores and the share of those performing at or below basic for the same subject become insignificant.

It is important to point out that the placebo test indicated that for the 5th grade, the municipal schools of São Paulo did not seem to be a good counterfactual for our estimations, given that the parallel trends assumption is likely to be violated. We continue now with the regressions that used schools of neighboring states as controls. The earlier examination pointed that this group fits better as a counterfactual.

When considering the non-matched sample, Table 18 shows that most of the results from the previous specification hold for 2009, although with a substantial decrease on the effect of the program. The positive impact on the average math and Portuguese scores decrease to 4 points and 1 point respectively. A similar development is observed for outcomes 5, 6 and 7, with the coefficients maintaining their sign and significance and decreasing in magnitude in absolute terms. The variable for the share of students performing at or below the basic in Portuguese now becomes positive and insignificant.

Table 18 - Difference in differences estimation for the 5th grade - state run schools as controls

5th grade - control: state run schools not matched								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2009	4.151*** (0.740)	1.295** (0.650)	-0.00565*** (0.00115)	-0.00936*** (0.00118)	-3.581*** (0.466)	-2.031*** (0.405)	-2.252*** (0.570)	0.199 (0.543)
treated2011	-0.679 (0.762)	-3.952*** (0.669)	-0.00382*** (0.00118)	-0.00688*** (0.00122)	-0.952** (0.480)	0.0293 (0.417)	1.867*** (0.587)	4.984*** (0.560)
Observations	17,987	17,987	17,986	17,986	18,008	18,008	18,008	18,008
R-squared	0.207	0.197	0.127	0.117	0.293	0.220	0.406	0.372
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								
5th grade - control: state run schools matched								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2009	5.515*** (0.855)	1.917** (0.764)	-0.00635*** (0.00131)	-0.00770*** (0.00137)	-3.550*** (0.553)	-1.610*** (0.488)	-2.665*** (0.650)	0.278 (0.628)
treated2011	-0.856 (0.884)	-4.475*** (0.790)	-0.00315** (0.00136)	-0.00532*** (0.00142)	0.00890 (0.572)	0.912* (0.505)	2.912*** (0.672)	6.601*** (0.650)
Observations	9,998	9,998	9,997	9,997	10,002	10,002	10,002	10,002
R-squared	0.235	0.207	0.176	0.161	0.332	0.245	0.455	0.405
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								

The results for 2011 cast much more doubts on the effectiveness of the program. The impact on test scores are now negative, and significant in the case of Portuguese. For the share of students performing below basic, the effect of the program seems to have faded out. For Portuguese, the coefficients are now non-significant and in math the effect decreased considerably, with a negative impact of less than 1%. More alarming are the outcomes for those performing at or below the basic. The coefficients are now positive and significant for both subjects, indicating that the program actually contributed to an increase in the share of the lowest performing students.

A similar analysis holds for the regressions where a matching on the propensity score was implemented. The impact of the program in 2011 was even more adverse than when the non-matched sample was used.

There is an indication that the policy contributed to an increase in the share of those performing at or below the basic of about 6% in Portuguese and 3% in mathematics.

We proceed now with the estimations for the 9th grade. Table 19 presents the results with the municipal schools as controls. In our placebo analysis, this was the group that seemed to fit better as a counterfactual. Focusing first on the non-matched sample, the effects on the average scores are much less pronounced when compared to the 5th grade. Only the coefficient for math in 2009 is significant. On the other hand, the program seems to have had the desired impact on the lowest performing students. The shares of those below basic and at or below the basic threshold decreased for the two subjects in both years in this specification, with a more expressive outcome in 2009.

Table 19 - Difference in differences estimation for the 9th grade - municipally run schools as controls

9th grade - control: municipal schools not matched								
VARIABLES	(1) average math	(2) average portuguese	(3) coefficient variation math	(4) coefficient variation portuguese	(5) below basic math	(6) below basic portuguese	(7) at or below basic math	(8) at or below basic portuguese
treated2009	1.655** (0.749)	0.647 (0.711)	0.00385*** (0.000890)	0.00490*** (0.00106)	-3.076*** (0.496)	-1.254*** (0.402)	-3.904*** (0.435)	-3.052*** (0.454)
treated2011	-1.140 (0.738)	-1.101 (0.700)	0.00513*** (0.000876)	0.00419*** (0.00104)	-1.972*** (0.488)	-0.987** (0.396)	-3.285*** (0.429)	-2.319*** (0.447)
Observations	14,218	14,218	14,218	14,218	14,227	14,227	14,227	14,227
R-squared	0.036	0.105	0.048	0.039	0.166	0.255	0.574	0.615
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								
9th grade - control: municipal schools matched								
VARIABLES	(1) average math	(2) average portuguese	(3) coefficient variation math	(4) coefficient variation portuguese	(5) below basic math	(6) below basic portuguese	(7) at or below basic math	(8) at or below basic portuguese
treated2009	0.262 (0.797)	-0.506 (0.756)	0.00402*** (0.000952)	0.00615*** (0.00112)	-2.583*** (0.535)	-0.902** (0.436)	-4.016*** (0.464)	-3.019*** (0.485)
treated2011	-2.362*** (0.798)	-2.202*** (0.757)	0.00605*** (0.000953)	0.00544*** (0.00113)	-1.429*** (0.536)	-0.485 (0.436)	-3.131*** (0.465)	-1.942*** (0.486)
Observations	12,429	12,429	12,429	12,429	12,435	12,435	12,435	12,435
R-squared	0.041	0.116	0.052	0.045	0.175	0.265	0.593	0.631
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								

The regressions with the matched samples indicate that the program had a negative and significant impact on average test scores in 2011. In 2009 the policy seems to be ineffective. When looking to the effect on the share of the lowest performing students this specification presents similar results to the previous one, with the biggest difference being that outcome 6 is now non-significant in 2011.

Table 20 shows the regressions using the state-run schools as controls. In the case of no matching, the outcomes for the average scores are all negative and significant in both years, indicating that the program

was at best ineffective in raising student's achievements. The shares of the lowest performing seem to have been negatively impacted in 2009, with the exception of the below basic category in Portuguese. In 2011 the outcome is less clear. A small negative impact is observed for outcome 6, while for variables 7 and 8 there is an indication that the program increased the share of those with the lowest scores.

Table 20 - Difference in differences estimation for the 9th grade - state run schools as controls

9th grade - control: state run schools not matched								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2009	-3.302*** (0.493)	-3.613*** (0.451)	0.00745*** (0.000668)	0.0101*** (0.000727)	-0.952*** (0.331)	-0.177 (0.265)	-2.851*** (0.309)	-1.306*** (0.316)
treated2011	-4.491*** (0.494)	-2.290*** (0.452)	0.0104*** (0.000670)	0.00346*** (0.000729)	0.111 (0.332)	-0.691*** (0.266)	1.033*** (0.310)	1.051*** (0.317)
Observations	28,695	28,694	28,694	28,694	28,714	28,714	28,714	28,714
R-squared	0.113	0.150	0.124	0.163	0.219	0.270	0.517	0.575
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								
9th grade - control: state run schools matched								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	average math	average portuguese	coefficient variation math	coefficient variation portuguese	below basic math	below basic portuguese	at or below basic math	at or below basic portuguese
treated2009	-3.851*** (0.540)	-3.625*** (0.509)	0.00774*** (0.000765)	0.0103*** (0.000821)	-0.154 (0.372)	0.0710 (0.301)	-1.614*** (0.333)	-0.654* (0.348)
treated2011	-4.792*** (0.543)	-1.984*** (0.512)	0.0106*** (0.000769)	0.00342*** (0.000826)	0.874** (0.374)	-0.561* (0.303)	2.449*** (0.334)	1.908*** (0.350)
Observations	18,958	18,958	18,958	18,958	18,966	18,966	18,966	18,966
R-squared	0.097	0.147	0.132	0.165	0.236	0.290	0.583	0.619
Standard errors in parentheses								
*** p<0.01, ** p<0.05, * p<0.1								

The matched sample indicate a similar scenario for the policy, with negative impacts on average scores and ambiguous results for the effects on the share of low performing students. It is worth noting that the placebo test indicated that the state-run schools of neighboring states were not the best control group to be used as a counterfactual when considering the 9th grade.

A more general overview of the results can now be done while attempting to identify what were the effects that the performance-based incentive scheme had on state run schools of São Paulo. There was no clear pattern identified when comparing the regressions using the full sample and the ones with matching on the propensity score. While some differences can be seen in the coefficients, no change of sign or great variation in significance was verified.

In general, the effects observed in 2009 are stronger than in 2011. It is also noted that for this period the impacts were more in line with the design of the policy, with an increase in test scores and a decrease in

the share of low performing students, especially for the 5th grade. The outcomes for 2011 are more misleading. The regressions on average scores suggest that the program might have had negative effects on the 9th grade while for the 5th more contradictory results were found. The outcomes related to the shares of the lowest performers were also inconclusive for this period, with significant and conflicting coefficients depending on the control group and grade.

Although less pronounced, another possible trend is that outcomes related to the math subject seem to be more impacted by the policy when compared to Portuguese related variables, with coefficients being usually smaller in absolute terms for the language subject.

To summarize, the program seemed to be effective on raising average test scores only for the 5th grade on the year right after the implementation of the program (2009). For the 9th grade in both years and the 5th in 2011 the results indicate that the intervention was ineffective at best. There is actually evidence that the program was detrimental to these outcomes. The scenario for the worse performing students is a bit brighter. The policy seems to have consistently reduced the share of those below the basic and at or below the basic for both grades and subjects in 2009. The estimations for 2011 are more misleading. For the 5th and 9th grades the impact on the share of those below the basic is either slightly negative or non-significant. For the share of those at or below the basic conflicting results were found, with some specifications presenting a positive coefficient and others negative.

In order to compare our results with previous research on the topic, it is presented in table 21 the estimations obtained by Oshiro et al. (2015). As mentioned before, only the impact of the program on the average test scores was evaluated. The left side columns are the ones more comparable to this study, containing the regressions that used the municipal schools of São Paulo as controls. The right side columns have estimations using southeast state schools as controls, while this study utilizes this group plus the states of Minas Gerais and Mato Grosso do Sul. Only the matched data will be referred to in this exercise.

The results from the 5th grade for the municipal control group present some similarities. The coefficient for math in 2009 is of a similar magnitude and significant in both cases while in Portuguese our results are significant and the comparison study is not. In 2011 both coefficients in Oshiro et al., (2015) are non-significant while here just the one for Portuguese is not robust. Now referring to the 9th grade, the results are very similar. In 2009 the estimations are not significant and in 2011 they are negative and significant at a 99% level.

Table 21 - Difference in differences estimations from Oshiro et al. (2015)

	Municipal schools from São Paulo as controls				Southeast state run schools as controls			
	5th grade		9th grade		5th grade		9th grade	
	Portuguese	Math	Portuguese	Math	Portuguese	Math	Portuguese	Math
treated2009	1,26	4,00***	-0,35	0,32	-0,16	2,88***	-2,61***	-3,14***
	(1,00)	(1,09)	(1,18)	(1,26)	(0,9)	(0,99)	(0,69)	(0,74)
treated2011	-1,49	1,72	-3,86***	-3,36***	-2,69***	-0,19	-3,64***	-6,09***
	(1,00)	(1,12)	(1,15)	(1,28)	(0,91)	(0,99)	(0,69)	(0,76)
Observations	5.640	5.814	3.960	3.900	7.386	7.596	11.490	11.040
R²	0,1489	0,1983	0,089	0,0154	0,1579	0,2052	0,1152	0,0335

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Although not directly comparable, the exercise is now done for the state schools control group. For the 5th grade, while this study presents positive and significant coefficients for both subjects in 2009, the comparison research indicates a positive and significant outcome only for math. For 2011 the results are quite similar with negative estimations in both subjects and only the impact on Portuguese being significant. The scenario for the 9th grade is also encouraging. All the results are negative and highly significant in both studies, with some minor variations in the magnitude of the coefficients.

It is reassuring to observe a relative convergence of effects in both evaluations. If the estimations hold in the case of average test scores and without diminishing the major limitations of the methodology employed, it is also likely that the original contribution presented here for the impacts of the program on low performing students are consistent and robust.

8. Conclusion and Discussion

We now conclude with a discussion of the potential mechanisms operating behind these results and indicate venues for future research. As a first step, we start by analyzing possible dysfunctional behaviors that might have arisen such as cheating or selecting students in the day of the test or during enrollment periods.

After an investigation in newspapers and magazines we have identified 3 cases where some kind of fraud in the exams might have happened. In 2009 fifteen state run schools of São Paulo were denounced for irregularities in the application of the exam. In some cases, the sealed packages with tests arrived at the schools with missing exams, indicating that some schools might have received the assessment before the application date. In one institution, teachers were reported to have helped the students during the test. In this case the own teachers of the school were the ones giving the assessment, a practice that in theory is not allowed in the program¹⁴. In 2011, all students from a small state school in the city of Sorocaba got the maximum score in math and near the maximum in Portuguese. After an investigation was conducted, evidence was found that the teacher helped students answering the test. Again in this case, it seems that the rule where teachers cannot give the assessment to their own school was not verified¹⁵. In 2014 evidence on another type of fraud surfaced. Several students in a state school in the municipality of Osasco reported that they were not allowed to enter the school in the day of the exam. According to them, the door keeper said that they had missed too many classes and therefore could not take the assessment, indicating that the school was selecting only the best performing students to make the test¹⁶. It is not clear if the 3 cases mentioned above were fully investigated by Secretary of Education of São Paulo and if further action was taken. While these episodes should be cause for concern, our research indicates that this type of actions are not systematically employed in the treated schools. We cannot however rule out this possibility.

On the other hand, even if such actions are taking place, they are probably not a great source of bias in our results. As already stated in a previous section, the assessment used to compute treatment effects is not linked to the bonus. While we used data from the federally run “Prova Brasil” test, the incentive scheme is conditional on results from the state run SARESP exam.

On the other hand, another type of selection might take place that would impact our results even if the assessment used in the estimations is not the same as the one linked to the bonus. If schools were able to select the best students during enrollment this would positively impact scores in both the SARESP and “Prova Brasil” exams and reduce the share of low performers. However, we do not think that this type of practice takes place. Usually the norm is that students enroll in the beginning of elementary school (1st grade) or lower secondary school (6th grade) and remain in the same school until they complete each cycle. In this case, selection in a large scale would only be possible in these initial grades. If an elementary school started selecting students in 2009 a potential positive effect of the selection would only be observed in 2013. On the more unlikely scenario that schools are able to select students in grades closer to the ones linked to the bonus we rely on estimations from Oshiro et al. (2015) to analyze the issue. Due to time constraints we were not able to run these tests ourselves. A widely publicized notion in the education

¹⁴ <https://www1.folha.uol.com.br/fsp/cotidian/ff0112200909.htm>

¹⁵ <https://jornalgn.com.br/blog/luisnassif/escola-e-acusada-de-fraude-no-saresp>

¹⁶ <https://educacao.estadao.com.br/noticias/geral,alunos-denunciam-fraude-no-saresp-em-escola-de-osasco,1619218>

literature is that a student socio economic background is a key determinant of achievements. Therefore, a significant impact in this variable after the implementation of the program might indicate that some kind of selection occurred. An analysis of the socioeconomic index of students between 2007 and 2009 found no significant treatment effects on this outcome. From 2009 to 2011 the impact on the index is actually slightly negative and significant, suggesting that no selection took place.

After ruling out some of the potential mechanisms that might have affected the results we now look to those that seem more plausible. To explain the differential impacts on the 5th and 9th grade is important to consider that the quality of public education in Brazil is relatively low and many students move on to next grade without having achieved the ideal level of achievement for their level. Throughout the school life it is likely that these deficits accumulate over the years, implying that is more difficult to effect change in higher grades. Another issue that might explain these differentials is that 5th grade students are usually taught by one generalist teacher that gives most of the subjects, including math and Portuguese, while the 9th grade is usually characterized by the presence of specialist's teachers that are responsible for only one subject. In this setting, a free rider behavior is more likely to arise in the 9th grade, where the link between effort and bonus is less salient (Oshiro et al. 2015).

The fall in the impact of the program from 2009 to 2011 seems to be the most complex phenomena to rationalize. The issue was not addressed by Oshiro et al. (2015). Our best guess is that teachers might have engaged in the so called practice of "teaching to the test", where efforts are diverted to activities that specifically focus on the exam linked to the bonus but might not provoke real gains in human capital. This behavior was also verified by Glewwe et al. (2010) in the pay for performance program implemented in Kenya. In that setting, students test scores in treated schools improved only in the exam that was linked to the incentive. In other tests of the same subjects no significant improvement was seen. They verified that the amount of time exclusively dedicated to training for the exam increased in class. We posit here that a similar mechanism might have taken place in São Paulo. On the year right after the implementation of the program (2009), teachers were still not fully aware of the characteristics of the assessment and therefore could not focus their attention on it. In that period, an effort with a broader spectrum might have occurred which translated into gains in the exam not linked to the bonus, the "Prova Brasil" test. However, 3 years after implementation teachers might have learned how to promote activities specifically designed for the exam linked to the scheme, the SARESP assessment. In that case, the effect of the program in the test not linked to the incentive could cease or even change signs since less time would be employed to broad spectrum activities.

This is however just a conjecture. We do not test for this hypothesis in the present study. This could be an interesting venue for future research on the topic. We now discuss a potential design for a study that could address the issue and would require access to the data on the exam linked to the bonus.

To the best of our knowledge the dataset of the SARESP assessment (the one linked to the incentive) is not publicly available at the school or student level. However, Brazil has developed a set of transparency laws and mechanisms that facilitate the access to this type of information, especially for academic purposes. Another way to go would be to establish an institutional relationship with Secretariat of Education of the State of São Paulo and propose an impact evaluation of the program. This would however bump into political issues since the validation of the hypothesis described above would not be beneficial for those in charge.

Assuming that the data is available, the difference between the scores in the two exams¹⁷ (SARESP and Prova Brasil) could be computed for the year's pre and post intervention and then the average treatment effects on the treated could be estimated using the same methodology employed in the present study. If treatment effects are significant for this outcome and indicate that the scores in the exam linked to the incentive increased relative to the one not related to the bonus, it could imply that the practice of teaching to the test is taking place. The construction of a control group would only be possible using the municipally run schools of São Paulo since they are the only other jurisdiction that is allowed to take the SARESP exam. This potential control group is not obliged to take the exam so this could be a source of bias that should be carefully examined.

We conclude by acknowledging that pay for performance schemes might have beneficial impacts on education systems, especially if the design of the program puts more weight in the lowest performing students. In the present study, the program seems to have systematically contributed to a reduction in the share of low performing students in both grades in 2009. However, the effects fade out in 2011. This indicate that there is still room for improving the design of the intervention, which could constitute another path for future research on the topic.

¹⁷ They use the same scale

Bibliography

- Abadie, A., Cattaneo, M.D., 2018. Econometric Methods for Program Evaluation. *Annual Review of Economics* 10, 465–503. <https://doi.org/10.1146/annurev-economics-080217-053402>
- Andrabi, T., Das, J., Khwaja, A.I., Zajonc, T., 2011. Do Value-Added Estimates Add Value? Accounting for Learning Dynamics. *American Economic Journal: Applied Economics* 3, 29–54.
- Angrist, J., Pischke, J.-S., 2009. Mostly Harmless Econometrics: An Empiricist's Companion, in: Mostly Harmless Econometrics: An Empiricist's Companion.
- Austin, P.C., 2009. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine* 28, 3083–3107. <https://doi.org/10.1002/sim.3697>
- Barrera-Osorio, F., Raju, D., 2017. Teacher performance pay: Experimental evidence from Pakistan. *Journal of Public Economics* 148, 75–91. <https://doi.org/10.1016/j.jpubeco.2017.02.001>
- Behrman, J.R., Parker, S.W., Todd, P.E., Wolpin, K.I., 2015. Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Journal of Political Economy* 123, 325–364. <https://doi.org/10.1086/675910>
- Bruns, B., Evans, D., Luque, J., 2012. Achieving World-Class Education in Brazil: The Next Agenda. The World Bank. <https://doi.org/10.1596/978-0-8213-8854-9>
- Bruns, B., Filmer, D., Patrinos, H.A., 2011. Making Schools Work: New Evidence on Accountability Reforms. The World Bank. <https://doi.org/10.1596/978-0-8213-8679-8>
- Caliendo, M., Kopeinig, S., 2008. Some Practical Guidance for The Implementation of Propensity Score Matching. *Journal of Economic Surveys* 22, 31–72. <https://doi.org/10.1111/j.1467-6419.2007.00527.x>
- Chetty, R., Friedman, J., Rockoff, J., 2011. The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood (No. w17699). National Bureau of Economic Research, Cambridge, MA. <https://doi.org/10.3386/w17699>
- Deaton, A., Cartwright, N., 2018. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine* 210, 2–21. <https://doi.org/10.1016/j.socscimed.2017.12.005>
- Ganimian, A.J., Murnane, R.J., 2016. Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations. *Review of Educational Research* 86, 719–755. <https://doi.org/10.3102/0034654315627499>
- Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B., Vermeersch, C.M.J., 2016. Impact Evaluation in Practice. The World Bank. <https://doi.org/10.1596/978-0-8213-8541-8>
- Glewwe, P., Ilias, N., Kremer, M., 2010. Teacher Incentives. *American Economic Journal: Applied Economics* 2, 205–227. <https://doi.org/10.1257/app.2.3.205>
- Guimarães, S.C., Marconi, N., 2017. O Impacto da Remuneração Variável no Desempenho do Servidor Público. *Administração Pública e Gestão Social* 1. <https://doi.org/10.21118/apgs.v1i1.1148>
- Hanushek, E.A., 2003. The Failure of Input-Based Schooling Policies. *The Economic Journal* 113, F64–F98.

Hanushek, E.A., Machin, S., Woessmann, L. (Eds.), 2016. Handbook of the economics of education. Volume 5: ..., Handbooks in economics. Elsevier North-Holland, Amsterdam Boston Heidelberg London New York Oxford Paris.

Harder, V.S., Stuart, E.A., Anthony, J.C., 2010. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods* 15, 234–249. <https://doi.org/10.1037/a0019623>

Hasnain, Z., Manning, N., Pierskalla, J.H., 2012. Performance-related Pay in the Public Sector: A Review of Theory and Evidence. World Bank, Policy Research Working Paper 57.

Monteiro, J., 2015. Gasto Público em Educação e Desempenho Escolar. *Revista Brasileira de Economia* 69, 467–488. <https://doi.org/10.5935/0034-7140.20150022>

Muralidharan, K., 2012. Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India 43.

Muralidharan, K., Sundararaman, V., 2011. Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy* 119, 39–77. <https://doi.org/10.1086/659655>

Murnane, R.J., Willett, J.B., 2011. Methods matter: improving causal inference in educational and social science research. Oxford University Press, Oxford; New York.

OECD, 2015. Education Policy Outlook: Brazil.

OSHIRO, C.H., 2012. Efeito do pagamento de bônus aos professores sobre a proficiência escolar. Master's thesis 69.

Oshiro, C.H., Scorzafeve, L.G., 2015. Impacto Sobre o Desempenho Escolar do Pagamento de Bônus aos Docentes do Ensino Fundamental do Estado de São Paulo. *Revista Brasileira de Economia* 69. <https://doi.org/10.5935/0034-7140.20150010>

São Paulo, 2018. Sumário Executivo: SARESP 2017. Secretaria de Educação do Estado de São Paulo.

São Paulo, n.d. Programa Qualidade da Escola. Secretaria de Educação do Estado de São Paulo Technical note, 2018 a.

Scorzafeve, L.G., Dorigan, T.A., Ferreira, J.N. de P., 2015. Efeito das políticas de bonificação aos professores sobre a desigualdade de proficiência no ensino fundamental. Unpublished manuscript.

UNESCO, 2015. Education for All 2015 National Review Report: Brazil.

UNESCO, 2010. World Data on Education: Brazil. 7th edition.

Victor Lavy, 2009. Performance Pay and Teachers' Effort, Productivity, and Grading Ethics. *The American Economic Review* 99, 1979–2011.

World Bank, n.d. Measuring Living Standards: Household Consumption and Wealth Indices. Quantitative Techniques for Health Equity Analysis - Technical Note #4.

Appendix

Table 22 - Estimated coefficients of probit - 5th grade with municipally run schools as controls

Variable	Coefficient	SE	z	P> z	95% confidence interval	
book 1_20	0,004	0,003	1,320	0,188	-0,002	0,010
book 21_100	-0,004	0,004	-0,970	0,332	-0,013	0,004
book 101	-0,008	0,005	-1,400	0,160	-0,018	0,003
mom 8 th grade	0,011	0,004	2,530	0,012	0,002	0,020
mom high school	0,026	0,004	5,860	0,000	0,017	0,035
preschool	-0,001	0,002	-0,920	0,358	-0,005	0,002
work	0,014	0,004	3,430	0,001	0,006	0,021
mom low education	-0,012	0,005	-2,180	0,030	-0,022	-0,001
wealth index	0,779	0,068	11,380	0,000	0,645	0,914
average math	-0,031	0,004	-8,080	0,000	-0,039	-0,024
average Portuguese	0,000	0,004	0,050	0,958	-0,008	0,008
teacher - no degree	0,003	0,001	2,020	0,043	0,000	0,005
teacher - inperson degree	0,002	0,001	1,390	0,165	-0,001	0,004
teacher - specialization	-0,007	0,001	-10,460	0,000	-0,009	-0,006
teacher - average wage	0,000	0,000	-1,170	0,240	0,000	0,000
teacher - temporary contract	0,009	0,001	9,080	0,000	0,007	0,011
teacher - stable contract	0,005	0,001	7,800	0,000	0,004	0,007
director - no specialization	-0,201	0,047	-4,250	0,000	-0,293	-0,108
director - wage	0,000	0,000	1,880	0,060	0,000	0,000
director - stable contract	-0,009	0,001	-10,590	0,000	-0,010	-0,007
director - financial resources	-0,780	0,043	-18,190	0,000	-0,864	-0,696
director - turnover teachers	-0,112	0,047	-2,400	0,017	-0,204	-0,020
director - computer for students	-1,012	0,063	-16,130	0,000	-1,135	-0,889
director - no library	0,141	0,053	2,640	0,008	0,036	0,245
director - no laboratory	0,348	0,060	5,840	0,000	0,231	0,465
_cons	5,679	0,441	12,880	0,000	4,815	6,543

Table 23 - Estimated coefficients of probit - 9th grade with municipally run schools as controls

Variable	Coefficient	SE	z	P> z	95% confidence interval	
book 1_20	0,016	0,005	3,560	0,000	0,007	0,025
book 21_100	0,009	0,005	1,830	0,068	-0,001	0,019
book 101	0,029	0,009	3,330	0,001	0,012	0,046
mom 8 th grade	0,009	0,005	1,860	0,063	0,000	0,019
mom high school	0,006	0,004	1,390	0,164	-0,002	0,015
preschool	0,006	0,002	3,320	0,001	0,002	0,009

work	0,006	0,003	1,790	0,074	-0,001	0,012
mom low education	0,015	0,006	2,710	0,007	0,004	0,027
wealth index	0,368	0,085	4,320	0,000	0,201	0,535
average math	-0,026	0,004	-6,760	0,000	-0,034	-0,019
average Portuguese	0,008	0,004	1,800	0,071	-0,001	0,016
teacher - no degree	-0,004	0,009	-0,450	0,652	-0,022	0,014
teacher - inperson degree	0,002	0,003	0,630	0,531	-0,004	0,008
teacher - specialization	-0,007	0,001	-9,230	0,000	-0,009	-0,006
teacher - average wage	0,000	0,000	-4,840	0,000	0,000	0,000
teacher - temporary contract	0,003	0,001	2,100	0,036	0,000	0,005
teacher - stable contract	0,007	0,001	8,890	0,000	0,006	0,009
director - no specialization	-0,180	0,055	-3,250	0,001	-0,288	-0,071
director - wage	0,000	0,000	-1,350	0,176	0,000	0,000
director - stable contract	-0,010	0,001	-10,240	0,000	-0,012	-0,008
director - financial resources	-0,875	0,050	-17,670	0,000	-0,972	-0,778
director - turnover teachers	-0,285	0,052	-5,470	0,000	-0,387	-0,183
director - computer for students	-0,793	0,085	-9,280	0,000	-0,960	-0,625
director - no library	0,085	0,088	0,960	0,337	-0,088	0,257
director - no laboratory	-0,251	0,053	-4,780	0,000	-0,354	-0,148
_cons	4,923	0,699	7,050	0,000	3,554	6,293

Table 24 - Estimated coefficients of probit - 5th grade with state run schools as controls

Variable	coefficient	SE	z	P> z 	95% confidence interval	
book 1_20	-0,022	0,003	-7,150	0,000	-0,028	-0,016
book 21_100	-0,035	0,004	-7,980	0,000	-0,044	-0,026
book 101	-0,040	0,005	-7,390	0,000	-0,051	-0,030
mom 8 th grade	0,019	0,004	4,410	0,000	0,011	0,028
mom high school	0,052	0,005	10,800	0,000	0,043	0,062
preschool	0,011	0,002	6,010	0,000	0,007	0,014
work	0,026	0,004	6,650	0,000	0,018	0,033
mom low education	-0,006	0,005	-1,120	0,262	-0,016	0,004
wealth index	0,695	0,057	12,260	0,000	0,584	0,806
average math	-0,013	0,004	-3,250	0,001	-0,020	-0,005
average Portuguese	-0,008	0,004	-1,910	0,056	-0,017	0,000
teacher - no degree	0,008	0,001	6,670	0,000	0,006	0,011
teacher - inperson degree	0,016	0,001	15,280	0,000	0,014	0,018
teacher - specialization	-0,014	0,001	-19,240	0,000	-0,016	-0,013
teacher - average wage	0,001	0,000	21,260	0,000	0,001	0,001
teacher - temporary contract	-0,009	0,001	-8,420	0,000	-0,011	-0,007
teacher - stable contract	-0,014	0,001	-15,870	0,000	-0,016	-0,013
director - no specialization	-0,013	0,059	-0,230	0,819	-0,129	0,102

director - wage	0,001	0,000	11,200	0,000	0,000	0,001
director - stable contract	-0,008	0,001	-8,500	0,000	-0,010	-0,006
director - financial resources	0,433	0,059	7,340	0,000	0,318	0,549
director - turnover teachers	0,122	0,052	2,340	0,019	0,020	0,224
director - computer for students	-1,068	0,070	-15,240	0,000	-1,206	-0,931
director - no library	0,682	0,073	9,310	0,000	0,538	0,825
director - no laboratory	1,010	0,061	16,490	0,000	0,890	1,130
_cons	1,368	0,470	2,910	0,004	0,446	2,289

Table 25 - Estimated coefficients of probit - 9th grade with state run schools as controls

Variable	Coefficient	SE	z	P> z 	95% confidence interval	
book 1_20	0,010	0,003	3,320	0,001	0,004	0,017
book 21_100	-0,002	0,004	-0,660	0,511	-0,010	0,005
book 101	0,007	0,006	1,090	0,276	-0,005	0,019
mom 8 th grade	0,042	0,004	11,750	0,000	0,035	0,049
mom high school	0,025	0,003	8,250	0,000	0,019	0,031
preschool	0,021	0,001	15,800	0,000	0,018	0,023
work	-0,013	0,002	-5,460	0,000	-0,018	-0,008
mom low education	0,030	0,004	7,310	0,000	0,022	0,038
wealth index	0,078	0,047	1,670	0,095	-0,014	0,170
average math	-0,031	0,003	-11,710	0,000	-0,036	-0,026
average Portuguese	0,005	0,003	1,800	0,071	0,000	0,011
teacher - no degree	0,003	0,006	0,470	0,641	-0,009	0,014
teacher - inperson degree	0,028	0,002	14,610	0,000	0,024	0,031
teacher - specialization	-0,016	0,001	-25,040	0,000	-0,017	-0,015
teacher - average wage	0,002	0,000	32,330	0,000	0,002	0,002
teacher - temporary contract	-0,022	0,001	-19,070	0,000	-0,024	-0,020
teacher - stable contract	-0,019	0,001	-20,040	0,000	-0,021	-0,017
director - no specialization	0,135	0,051	2,650	0,008	0,035	0,234
director - wage	0,000	0,000	13,160	0,000	0,000	0,001
director - stable contract	-0,010	0,001	-12,540	0,000	-0,012	-0,009
director - financial resources	0,309	0,046	6,770	0,000	0,219	0,398
director - turnover teachers	-0,009	0,040	-0,230	0,816	-0,088	0,069
director - computer for students	-1,097	0,067	-16,420	0,000	-1,227	-0,966
director - no library	0,248	0,082	3,040	0,002	0,088	0,409
director - no laboratory	0,411	0,042	9,830	0,000	0,329	0,493
_cons	0,281	0,492	0,570	0,568	-0,684	1,245

Figure 3 - percentage standardized bias before and after the matching - 5th grade with municipally run schools as controls

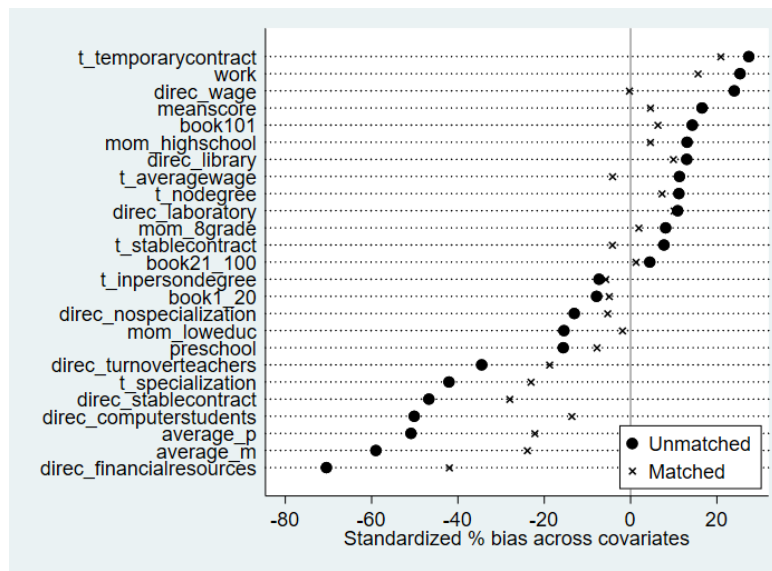


Figure 4 - percentage standardized bias before and after the matching - 9th grade with municipally run schools as controls

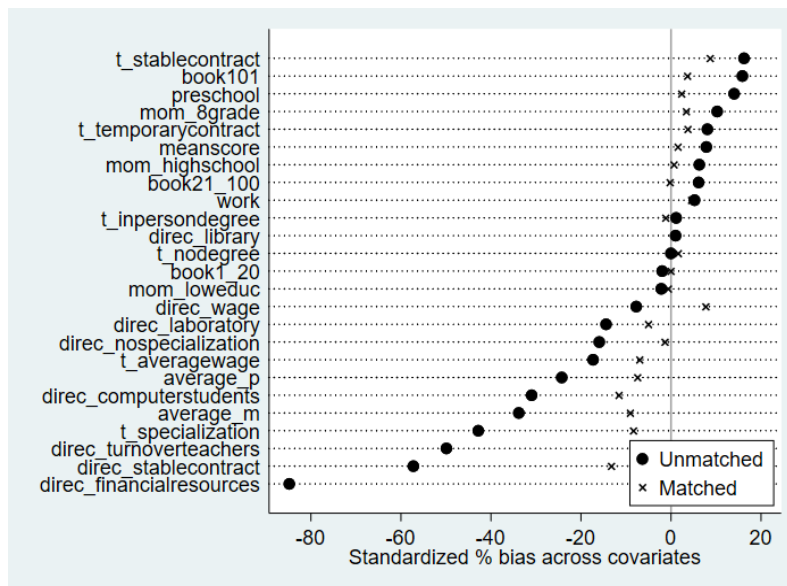


Figure 5 - percentage standardized bias before and after the matching - 5th grade with state run schools as controls

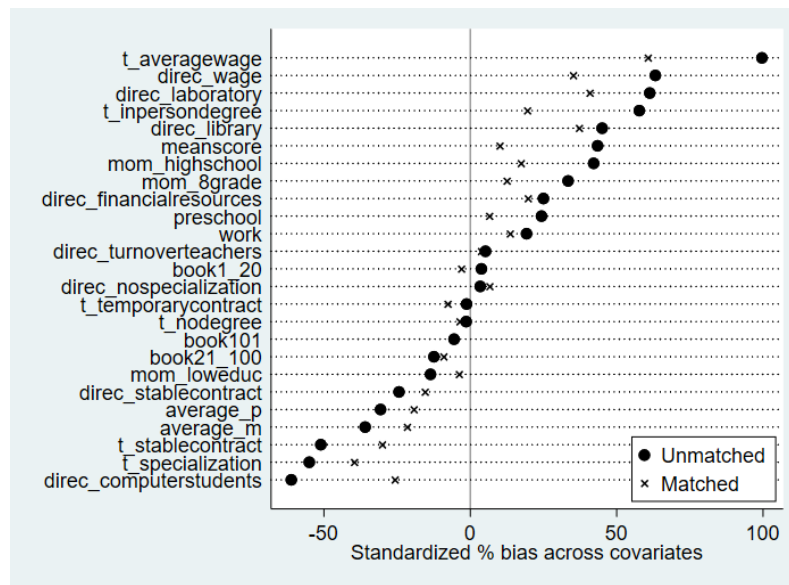


Figure 6 - percentage standardized bias before and after the matching - 9th grade with state run schools as controls

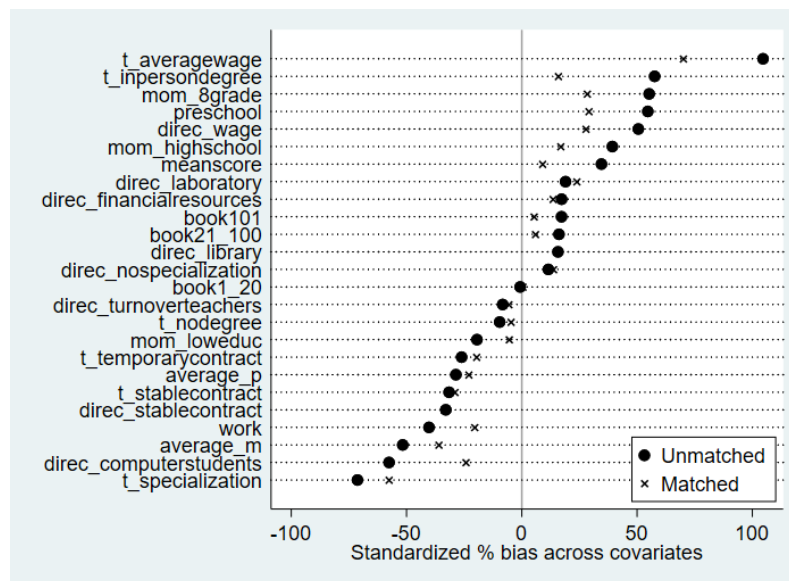


Figure 7 - average test scores for the 5th grade - municipally run schools as controls

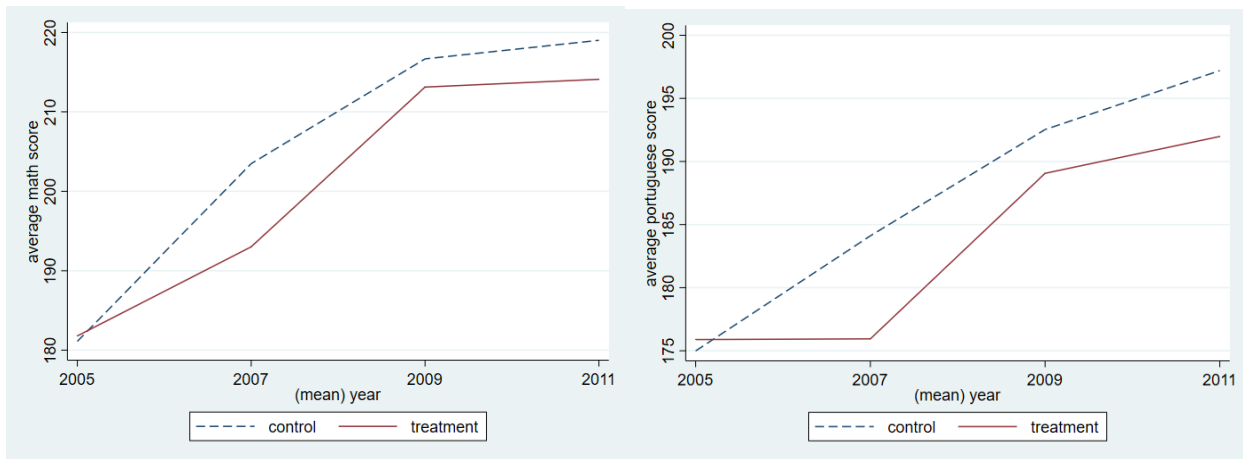


Figure 8 - average coefficient of variation in test scores for the 5th grade - municipally run schools as controls

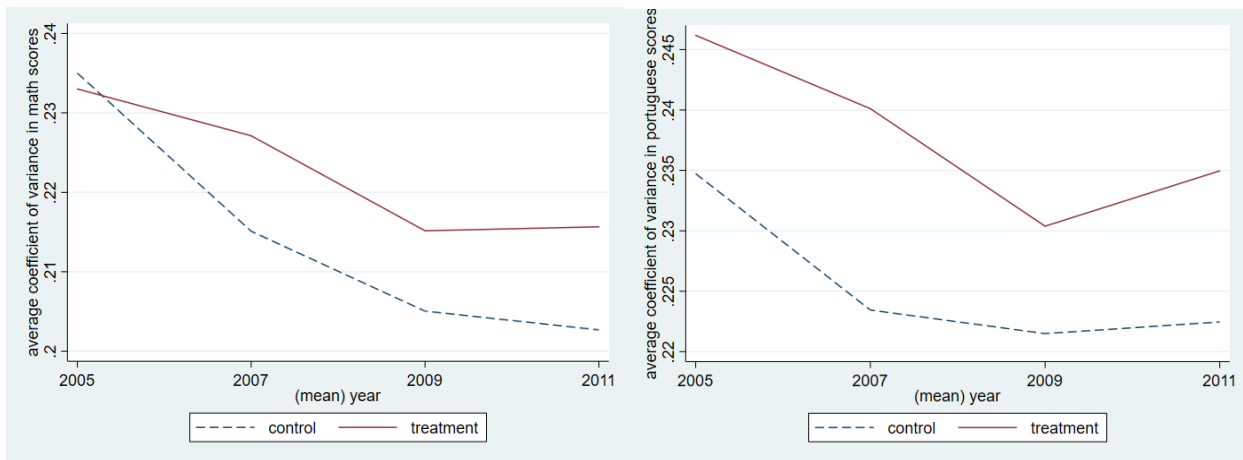


Figure 9 - average share of students below the basic in the 5th grade - municipally run schools as controls

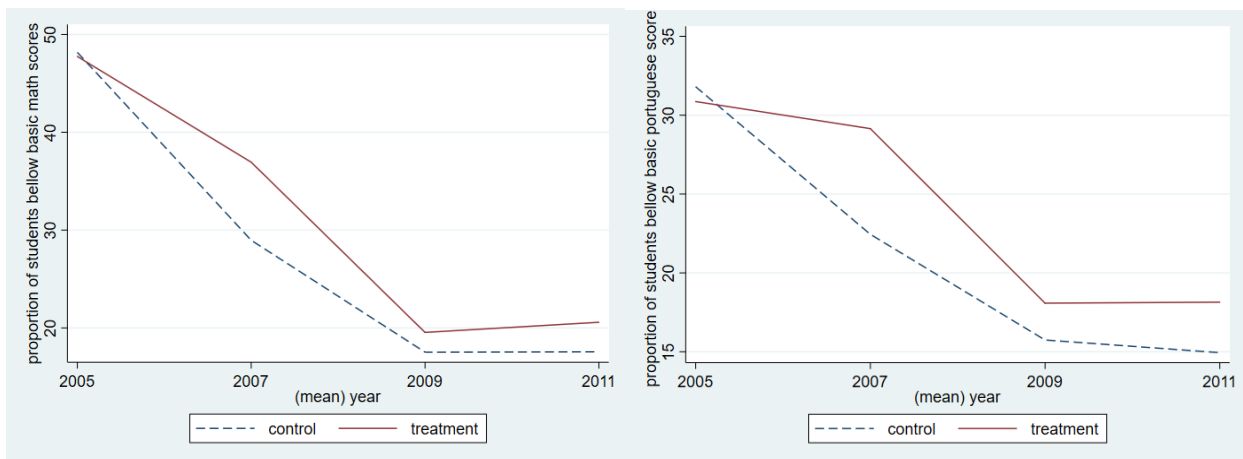


Figure 10 - average share of students at or below the basic in the 5th grade - municipally run schools as controls

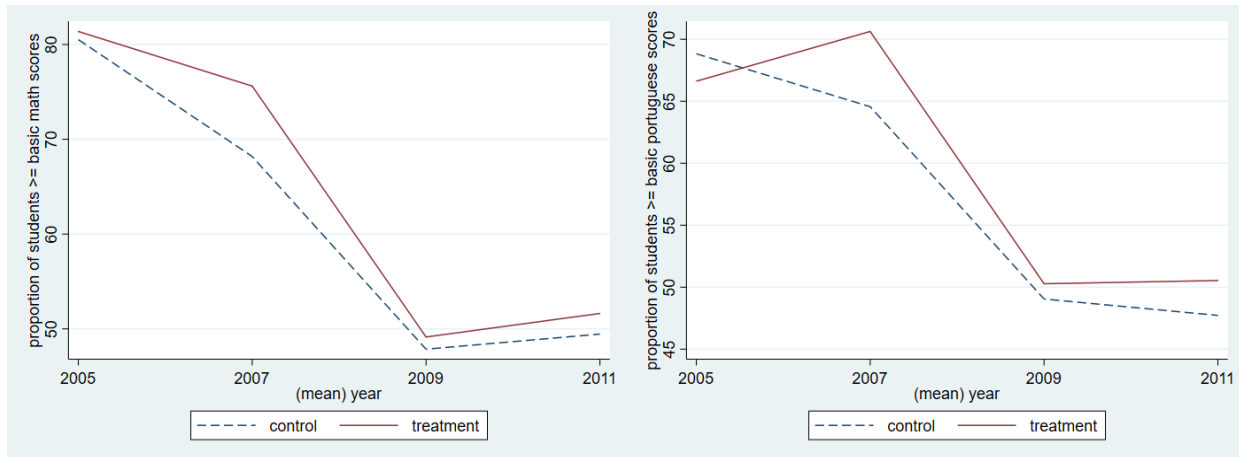


Figure 11 - average test scores for the 9th grade - municipally run schools as controls

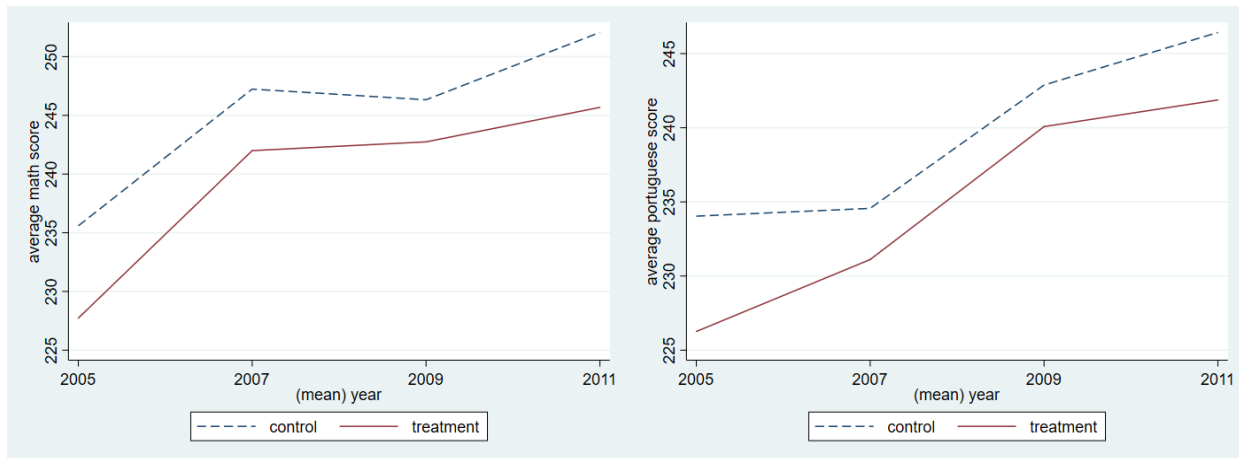


Figure 12 - average coefficient of variation in test scores for the 9th grade - municipally run schools as controls

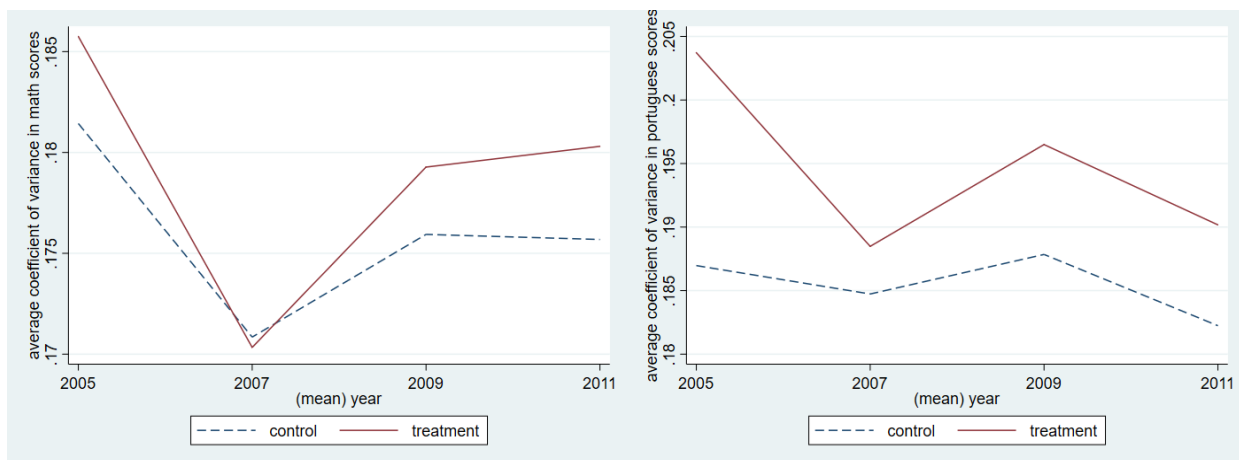


Figure 13 - average share of students below basic in the 9th grade - municipally run schools as controls

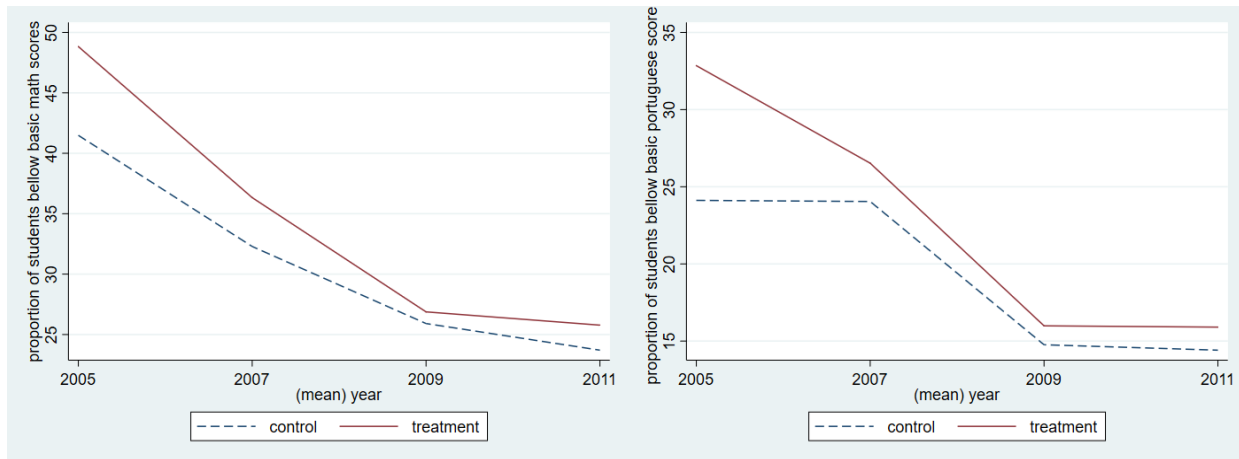


Figure 14 - average share of students at or below basic in the 9th grade - municipally run schools as controls

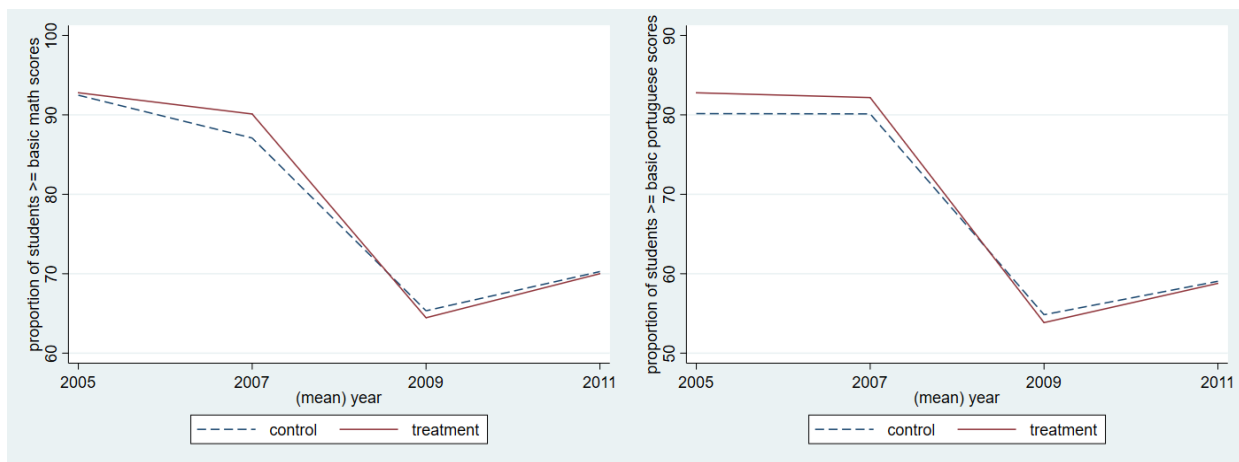


Figure 15 - average test scores for the 5th grade - state run schools of neighboring states as controls

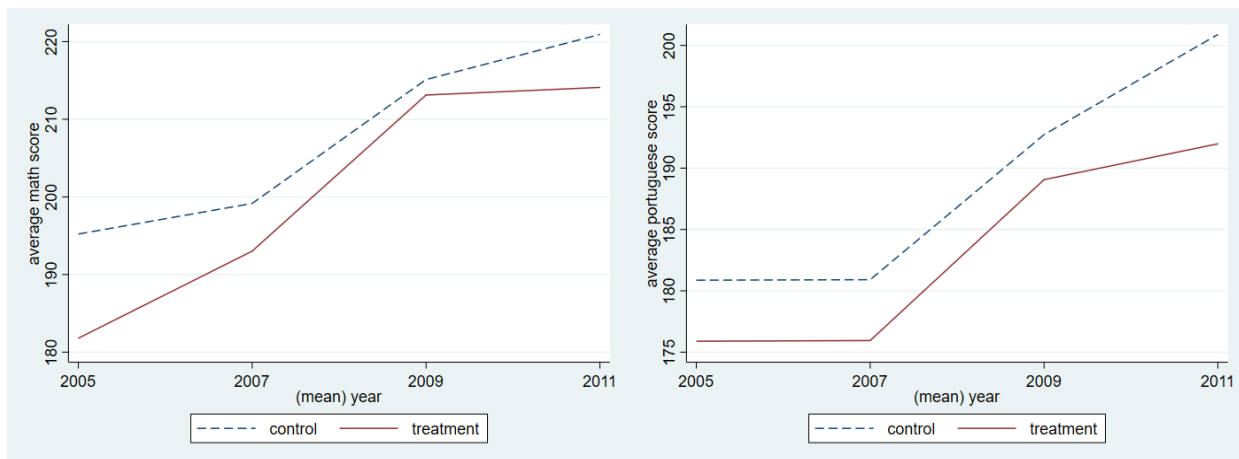


Figure 16 - average coefficient of variation in test scores for the 5th grade - state run schools of neighboring states as controls

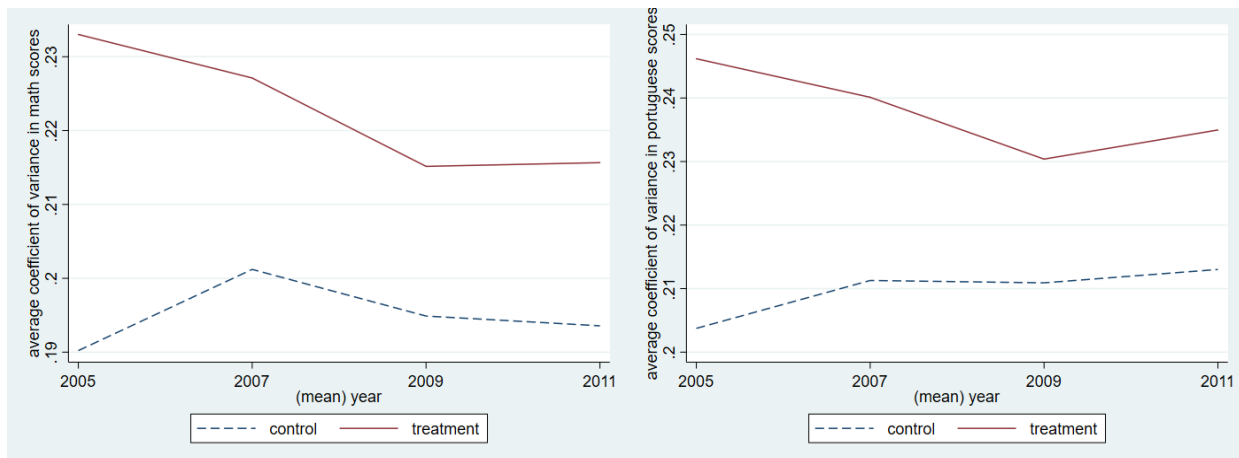


Figure 17 - average share of students below basic in the 5th grade - state run schools of neighboring states as controls

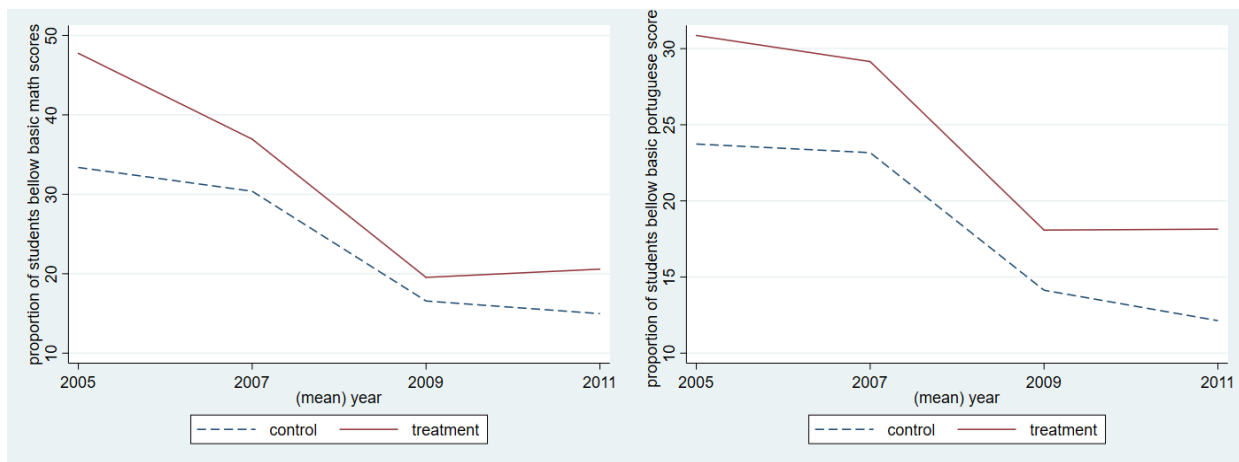


Figure 18 - average share of students at or below basic in the 5th grade - state run schools of neighboring states as controls

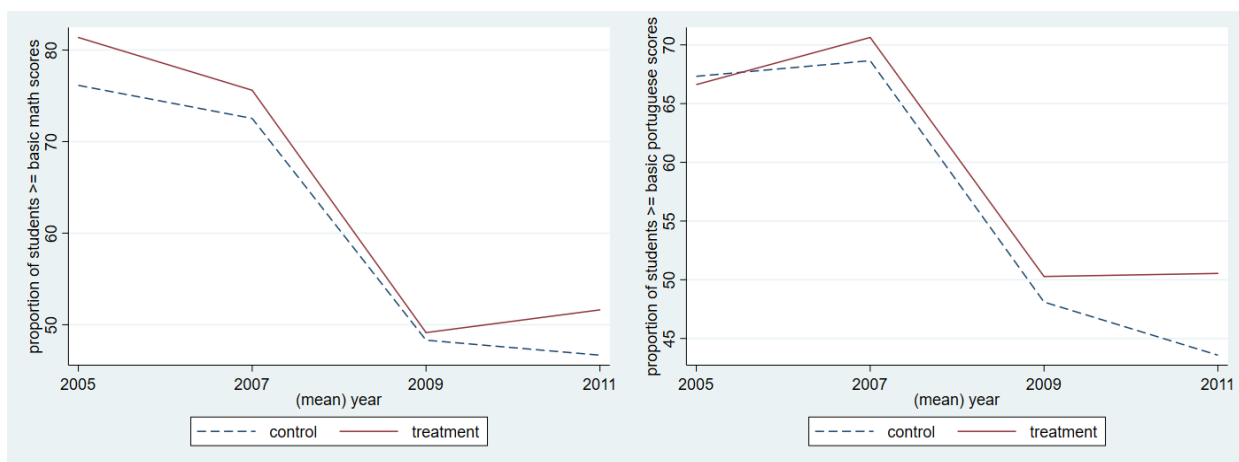


Figure 19 - average test scores for the 9th grade - state run schools of neighboring states as controls

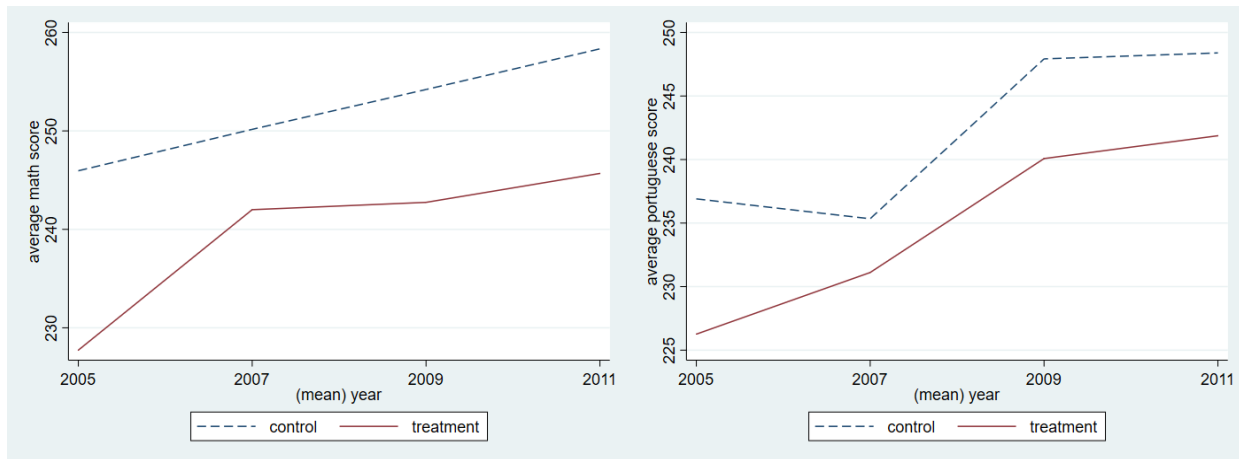


Figure 20 - average coefficient of variation in test scores for the 9th grade - state run schools of neighboring states as controls

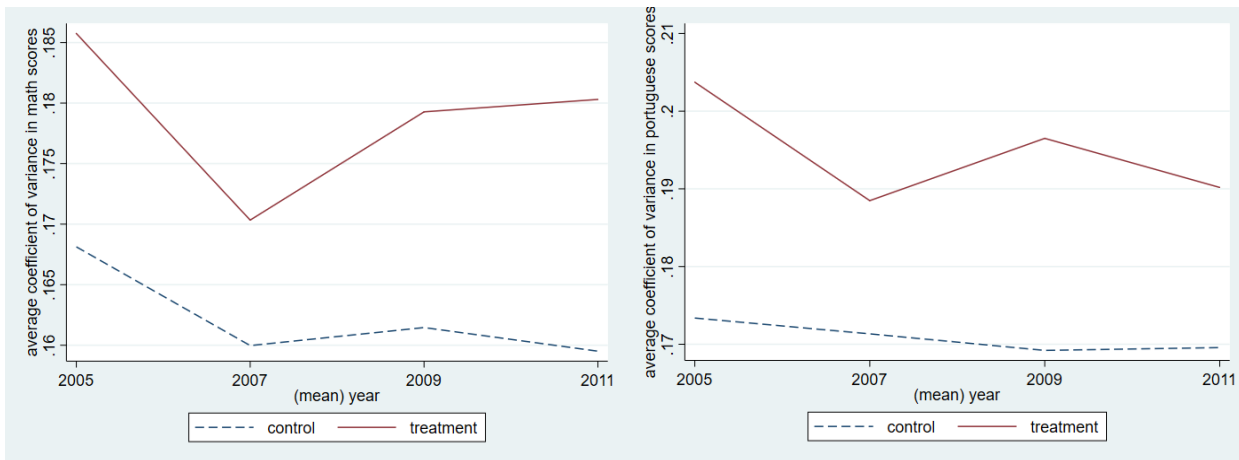


Figure 21 - average share of students below basic in the 9th grade - state run schools of neighboring states as controls

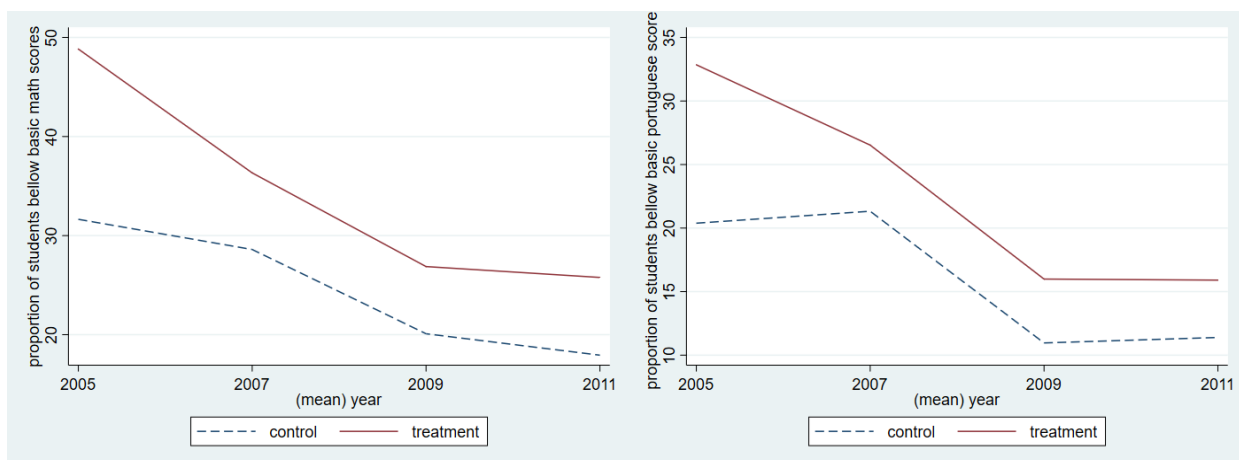


Figure 22 - average share of students at or below basic in the 9th grade - state run schools of neighboring states as controls

