Department of Economics- FEA/USP

Teacher Incentives and Student Performance: Evidence from Brazil

ANDREA LÉPINE

WORKING PAPER SERIES Nº 2016-18

DEPARTMENT OF ECONOMICS, FEA-USP
WORKING PAPER Nº 2016-18

Teacher Incentives and Student Performance: Evidence from Brazil

Andrea Lépine (alepine@usp.br)

Abstract:

This paper provides evidence on a large-scale teacher incentive program in the state of São Paulo, Brazil, which awarded group bonuses to teachers and school staff conditional on improvements in student performance. By using a difference-in-differences and triple-differences framework, I show that the program had overall positive effects on student achievement, although improvements vary across grades and subjects. The robustness of the results is assessed through the use of a series of alternative counterfactuals. I also investigate whether initial school characteristics affect the impact of the program. Although it could be expected that free-riding effects increase with the number of teachers in schools, therefore limiting the impact of the program, this does not seem to be the case. More sizeable differences are found according to school's previous performance. Initially low-performing schools improved much more than the average, suggesting there may be considerable differences in the ability of schools to respond to this type of policy.

Keywords: Pay for performance; Student achievement; Incentives

JEL Codes: I21; I28; J45

Teacher Incentives and Student Performance: Evidence from Brazil

Andrea Lépine[*]

September 2016

Abstract

This paper provides evidence on a large-scale teacher incentive program in the state of São Paulo, Brazil, which awarded group bonuses to teachers and school staff conditional on improvements in student performance. By using a difference-in-differences and triple-differences framework, I show that the program had overall positive effects on student achievement, although improvements vary across grades and subjects. The robustness of the results is assessed through the use of a series of alternative counterfactuals. I also investigate whether initial school characteristics affect the impact of the program. Although it could be expected that free-riding effects increase with the number of teachers in schools, therefore limiting the impact of the program, this does not seem to be the case. More sizeable differences are found according to school's previous performance. Initially low-performing schools improved much more than the average, suggesting there may be considerable differences in the ability of schools to respond to this type of policy.

Keywords: Pay for performance, Student achievement, Incentives

[*] Universidade de São Paulo. Email: alepine@usp.br

1 Introduction

Policies linking teacher pay to student performance have drawn considerable attention in recent years, and are seen as a promising way of improving student learning. This idea is motivated by the view that teachers generally face weak incentives and low accountability, especially in developing countries. And while teacher quality has been shown to be an important factor explaining student achievement (Rockoff 2004), teacher observable characteristics such as experience and qualifications, which are the main determinants of salaries in most school systems, do not seem to be good predictors of student performance (Rivkin, Hanushek and Kain 2005; Aaronson, Barrow and Sander 2007).

It is not obvious, however, that incentives will have the desired effects in the specific context of schools. While the standard principal-agent framework suggests that appropriate incentives can increase efficiency, schools have specific characteristics that could change the classic model's predictions. First, given that teachers' work involves multiple tasks, the introduction of performance pay could result in a reallocation of effort toward skills tested on exams linked to the incentive program (or "teaching to the test"), at the expense of human capital accumulation in a broader sense. Recent papers have discussed this issue (Muralidharan and Sundararaman 2011; Neal 2011) based on the theoretical model developed by Holmstrom and Milgrom (1991), and show that the final outcome depends on model parameters of a given school system, which are generally not known. Second, the fact that teaching involves teamwork means there can be complementarities in the education production function, which explains why teacher performance bonuses are often distributed at the group level. One issue related to this type of incentive design is the possible occurrence of free-riding (Holmstrom 1982), which is likely to increase as the group gets large and may limit the effectiveness of such policies. Third, it is possible that in organizations where agents are driven by intrinsic motivation such as schools, incentives can become less efficient (Besley and Ghatak 2005). Finally, incentives could be inefficient if they are not properly understood by teachers. Overall, the theoretical literature indicates that the impact of this type of program is likely to be context-dependent, pointing to the importance of empirical evidence.

Rigorous empirical evidence on the effectiveness of such policies is limited, and points to mixed findings. Fryer (2013) uses a randomized trial to analyze the effects of a teacher incentive program in New York, where schools could decide how they would distribute the incentive bonus among teachers, and finds no evidence that the program increased student performance or changed teacher behavior. Other studies in the United States (Glazerman and Seifullah 2012; Springer et al. 2010) also find no clear evidence that teacher incentive programs improved student achievement. In England, the impact of a pay scheme for teachers partly based on student performance was estimated by Atkinson et al. (2009). Although they find overall positive effects, the impacts are negative for some subjects. A similar program is analyzed by Martins (2009), who looks at the effect of a teacher pay reform in Portugal which conditioned teachers' progression on a pay scale on student test scores, and finds that the policy resulted in a decline in student performance and an increase in grade inflation.

Studies from developing countries provide more encouraging findings. In a well-known paper, Muralidharan and Sundararaman (2011) present results from a randomized evaluation in India assessing the impact of a program that provided group and individual bonuses for teachers, conditional on student performance. They find that both types of programs increased student achievement, and find no evidence of adverse behavior from teachers. In another randomized trial by Glewwe, Ilias and Kremer (2010) in Kenya, a school-based incentive program directed at teachers from primary schools was found to have positive but short-term effects on students' test scores. However, gains were only observed for exams directly linked to the incentives, showing evidence of "teaching to the test" and suggesting that the program did not lead to a broad increase in learning. Lavy (2002, 2009) looks at the impact of two different teacher incentive programs in Israeli high schools using natural experiments, and also finds positive effects on student test scores.

This paper contributes to the literature on teacher incentives by evaluating the effect of a group-based teacher pay for performance program in the state of São Paulo, Brazil, introduced in 2008. The program awards bonuses to teachers and staff from public schools run by the state government, conditional on improvements in student test scores. Using a difference-in-differences and triple differences framework and data from a standardized test not directly related to the incentive scheme (*Prova Brasil*), I estimate the effect of the program on 5th and 9th grade students up to 5 years after the program implementation. The robustness of the results is assessed through the use of a series of different counterfactuals. The estimates obtained show that the program had overall positive effects, although achievement gains among 9th grade students were more modest and less robust across counterfactuals. When averaging the estimates found across all counterfactuals, 5th grade students' math and language test scores improved by 0.15 and 0.09 standard deviations respectively, while the corresponding gains for 9th grade students were 0.06 and 0.03 standard deviations.

I also investigate whether initial school characteristics affect the impact of the program. First, I look at whether the number of teachers influences how schools react to the program, in an attempt to provide evidence on the presence of possible free-riding effects. This does not seem to be generally the case, although the number of teachers has a modest negative effect on 9th grade students' performance. I also estimate how schools' initial performance affect gains from the program, and find that initially low-performing schools improved much more than other schools, suggesting a reduction in inequality among schools.

A related study by Oshiro, Scorzafave and Dorigan (2015) assesses the impact of the teacher bonus program in São Paulo through propensity score matching and difference-in-differences and finds mixed results. This paper aims at bringing further evidence on this issue by using a different methodology and a longer span of data, and by exploring additional impacts of the policy.

This paper is organized as follows. Section 2 provides background information on the teacher performance program in São Paulo. Section 3 describes the data used and shows some descriptive statistics. Section 4 discusses the empirical strategy, and Section 5 presents the main results. Heterogeneous effects are analyzed in Section 6, and Section 7 concludes.

2 Background

São Paulo is the richest and most populous state in Brazil, with a population of over 40 million. As in the rest of Brazil, its school system is composed of private schools, and of public schools managed either by the federal, state or municipal government (from now on referred to as "federal schools", "state schools" and "municipal schools" respectively). According to school census data, the state government managed over 5500 schools in 2013, which accounted for around 20% of the total number of schools in São Paulo[1].

Despite improvements in recent years, student performance in São Paulo and in Brazil more generally remain low compared to high-income countries, and even compared to countries with similar levels of per capita income, as evidenced by international student assessments such as PISA[2]. As in many other developing countries, teachers working in the Brazilian public school system face low accountability and cannot be easily dismissed for poor performance. Teacher absenteeism is also an important problem.

The pay for performance program is part of an initiative launched in 2008 by the Secretary of Education of the State of São Paulo, with the objective of improving the quality of education in public schools run by the state government. According to this initiative, schools are assessed every year through an indicator called *Idesp* (*Índice de Desenvolvimento da Educação do Estado de São Paulo*), which serves as the basis for the calculation of teacher bonuses. Although discussions on the pay for performance program in São Paulo started in 2007, its practical details were not clear until the end of 2008 and the law determining its adoption was only passed in December 2008. First bonus payments were distributed in March 2009 based on the progress achieved by schools in 2008 relative to 2007 (the first year for which the *Idesp* was calculated).

The *Idesp* indicator combines information on student retention and on schools' performance measured through an annual standardized test, the *Saresp* (*Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo*). Only students' performance in math and language (Portuguese) are taken into account, even though the test covers other subjects such as science. The *Idesp* is calculated separately for three different school cycles: 1st to 5th grade, 6th to 9th grade, and 10th to 12th grade. It is obtained through the multiplication of an index of student performance (which involves a calculation based on the proportion of students that fall in pre-determined performance categories), and an index of student flow, determined by the share of students allowed to pass to the next grade. From 2013 onwards, the *Idesp* also started taking into account students' socio-economic composition.

[1] Private schools and municipal schools accounted for approximately 35% and 44% of schools respectively.

[2] Program for International Student Assessment of the OECD (Organisation for Economic Cooperation and Development).

The attribution of bonuses is linked to schools' *Idesp* improvement relative to targets that are individually set. These targets are determined both by schools' initial situation and by long terms goals for 2030 that are identical for all schools within the same cycle. The incentive scheme involves a group bonus, where school staff receive a payment that is proportional to how much each school has improved compared to its target. In schools that have exactly attained their targets for example, staff receive a payment equivalent to 2.4 minimum wages while the upper limit for schools that have improved beyond their targets is 2.9 minimum wages[3]. The 10% best schools of the state of São Paulo also receive a reward regardless of their improvement[4]. Although teacher salaries vary depending on the grade and subject taught, and on teachers' qualifications, the bonus amounts on average to a little less than one monthly salary for a full-time middle school teacher, a relatively large payment compared to most other incentive programs analyzed in the literature.

São Paulo is not the only state in Brazil to implement initiatives aimed at rewarding good educational practices; however, in the majority of cases there is no direct link between teacher compensation and student performance. To my knowledge, only four other states so far (from a total of 27 states in Brazil), have implemented programs explicitly tying teacher bonus pay to student performance. In particular, the state of Pernambuco in northeast Brazil has adopted a very similar pay for performance program as the one in São Paulo in 2008. Other states that have recently implemented similar programs are Espírito Santo, Rio de Janeiro, and Amazonas[5].

[3] Only staff who worked at least 244 days during the year are eligible for the bonus.

[4] These rewards are lower than the regular bonus (1.5 minimum wages).

[5] The bonus program in Espírito Santo, implemented in 2011, determines that teachers and school staff can receive up to one additional monthly salary per year based on student performance and other indicators such as absences. In Rio de Janeiro, where a similar program was implemented the same year, teachers can receive up to three monthly salaries. In Amazonas, a program implemented in 2008 determines that teachers from the best schools can receive up to two additional salaries, although its rules are different and a smaller proportion of schools is awarded bonuses.

3 Data

For the analysis in this paper I use data from *Prova Brasil*, a standardized test applied every two years to 5th and 9th grade students from all public schools[6], created in 2005. *Prova Brasil* is a low-stakes test designed as a tool to help teachers and policymakers assess the quality of education in Brazilian schools. In addition to taking an exam in math and language (Portuguese), students also provide socio-economic information through a survey. School principals and teachers also provide information about the school, its staff and working conditions through separate surveys.

The dataset used in this paper includes detailed information at the student level for 2007, 2009, 2011 and 2013, that is, one period of data before the implementation of the program, and three periods of data after the program was implemented. I also use 2007 School Census data to obtain information on total student enrollment (which allows the calculation of the percentage of test-takers in each school), and on the number of teachers by school. Columns 1-2 of Table 1 present descriptive statistics from the sample of 5th and 9th grade students before the implementation of the policy in 2007. Only students with available data on test scores are included, and who are enrolled in schools managed by the state or municipal government (students from other types of school represent less than 0.01% of the total). The share of test-takers relative to the total number of enrolled students is high, averaging around 90% for 5th grade students and 80% for 9th grade students. The data show that schools from the state of São Paulo are larger on average than schools from the rest of Brazil, and that their student population comes from a slightly more privileged socio-economic background, with a higher proportion of students whose parents finished high school.

In São Paulo as in the rest of Brazil, the majority of schools offering 5th grade education is managed by the municipal government, while the majority of schools offering 9th grade education is managed by the state government. However, during the period of analysis some schools switch from being run by the state government to being run by the municipal government, following a trend toward the decentralization of school management started in the 1990's. Overall, the share of municipal schools increases by around 10 percentage points between 2007 and 2013. As a result, some schools initially

[6] Only schools with a least 20 students enrolled in each grade are tested, and a very small number of private schools were tested during the period of analysis.

managed by the state government in São Paulo stop participating from the teacher pay for performance program in the course of this period. As most of these schools adopt new school identifiers, it is not possible to distinguish them from schools that drop out of the sample due to attrition. Given that the specific rules of the decentralization process might differ across states, I only include in the analysis schools that have not changed management in the period 2007-2013 to avoid selection issues[7]. Columns 3-4 of Table 1 show statistics for the final sample which only includes schools that have all 4 periods of data available. The new sample is about 70% the size of the original sample of schools offering 5th grade education and about 80% of the original sample of schools offering 9th grade education, and presents very similar observable characteristics as the original sample.

4 Empirical Strategy

In the absence of experimental data, the difference-in-differences (DD) method provides a way of estimating the impact of a program under the assumption that the outcomes of interest of the treated and control groups would have followed parallel trends over time in the absence of the treatment. If this identifying hypothesis is valid, the method allows the estimation of the average treatment effect on the treated (ATT). However, the estimator will be biased if there are time-varying unobserved factors that affect the outcome of both groups differently.

A typical concern in DD designs is the motivation behind the decision to participate in a program, as individuals who decide to enter a program may have specific unobserved characteristics that can affect their outcomes' trajectory. In the case of the São Paulo teacher incentive program, the participation decision has been taken at a centralized level, and concerns all state schools from the state of São Paulo regardless of their characteristics. It is still possible, however, that São Paulo state schools share distinct unobservable characteristics that would have led them to experience a different trend in performance compared to schools from other states in the absence of the program. Although the parallel trend hypothesis is impossible to verify, since only treated or control schools can be observed at a given time, I use a series of different counterfactuals

[7] The results of the analysis are very similar, however, when using the whole sample of schools.

in the estimations to test the robustness of the results.

First, I use students from São Paulo municipal schools as the comparison group, as they are not affected by the policy but are likely to be subject to the same state-specific time trends as students from São Paulo state schools. I then use students from state schools in other Brazilian states and regions to form several other comparison groups. The fact that there are both state and municipal schools within each state allows for a more robust analysis by using a triple differences framework (DDD), in order to remove state-specific time trends that could bias simple difference-in-differences estimations. Intuitively, the triple differences estimator is obtained by subtracting from the DD estimator differences in trends between São Paulo municipal schools and municipal schools from other states. If, however, there are trends that are specific to states and to the type of school, this method would not be effective in removing all bias.

The data available allow for a panel data analysis at the school level, as well as a student-level analysis using repeated cross-sections. I present the main results using cross sectional data at the student level with school fixed effects, as this allows for an analysis at a more disaggregated level and a larger number of observations[8]. School-level estimations provide very similar results and are used subsequently for further analysis.

Formally, the difference-in-differences equation can be written as follows:

$$Y_{ijt} = \alpha + \beta X_{ijt} + \sum_{\tau=2009}^{2013} \gamma_\tau \mathbb{1}(Year = \tau) + \sum_{\tau=2009}^{2013} \delta_\tau Treat_j * \mathbb{1}(Year = \tau) + \mu_j + \epsilon_{ijt} \quad (1)$$

Where Y_{ijt} is the outcome of student i in school j and year t, X_{ijt} is a vector of student-level covariates, $\mathbb{1}(Year = \tau)$ are a set of year dummies, $Treat_j$ is a dummy indicating treatment status (which equals 1 for the treated group, composed of schools managed by the state government of São Paulo), and μ_j are school fixed effects. The coefficients of interest are δ_τ, which capture the effect of the interaction between the treatment dummy and dummies for years in which the program was implemented. The fact that three periods of data are available after the implementation of the program allows me

[8] Hanushek, Rivkin and Taylor (1996) discuss aggregation bias in analyses of student performance and suggests that aggregation can increase bias in some cases.

to use a flexible specification where treatment effects can vary over time. I also run an alternative specification where the treatment effect is constant over time. School fixed effects are intended to capture school characteristics that are fixed in time and could potentially influence the outcome, which includes any factors that are common to all schools in the state of São Paulo.

Similarly, the triple-differences estimator can be obtained through the following regression:

$$\begin{aligned} Y_{ijt} = \alpha + \beta X_{ijt} + \sum_{\tau=2009}^{2013} \gamma_\tau \mathbb{1}(Year = \tau) + \sum_{\tau=2009}^{2013} \delta_\tau SP_j * \mathbb{1}(Year = \tau) \\ + \sum_{\tau=2009}^{2013} \theta_\tau SS_j * \mathbb{1}(Year = \tau) + \sum_{\tau=2009}^{2013} \phi_\tau SP_j * SS_j * \mathbb{1}(Year = \tau) + \mu_j + \epsilon_{ijt} \quad (2) \end{aligned}$$

In addition to year dummies, the DDD equation includes interactions between year dummies and SP_j, a dummy indicating whether the school is located in the state of São Paulo, and interactions between year dummies and SS_j, a dummy indicating whether the school is managed by a state government. The three coefficients of interest are ϕ_{2009}, ϕ_{2011}, ϕ_{2013}, associated with the triple interaction, and as in the previous case I also run an alternative specification where the treatment effect is constant over time.

To assess whether student composition effects are driving the results (which might be the case for example if low-performing students are encouraged to dropout or change schools), I present estimates both with and without student-level covariates.

A potential concern with DD estimates is the fact that standard errors may be correlated within groups and over time, which could bias their estimation and understate the standard deviation of the estimator. This issue has been discussed by Bertrand, Duflo and Mullainathan (2004), who suggest that when the number of groups is large, one solution is to allow for the auto-correlation of standard errors. Following this idea, I cluster standard errors at the school level in all specifications.

5 Main Results

5.1 Student Performance

Tables 2 and 3 present the main results for 5th and 9th grade students respectively. In columns 1-2, I present simple DD estimates where the comparison group is composed by municipal schools in the state of São Paulo. Next, I present DDD estimates using a series of different counterfactuals from other Brazilian states. In columns 3-4 the comparison group is composed of all Brazilian state schools excluding São Paulo, and in columns 5-6 the comparison group is limited to state schools located in São Paulo adjacent states. Next, I only include in the analysis treated and control schools located in regions that are geographically close to each other, in an attempt to restrict the comparison to schools that are more similar in terms of unobservable characteristics. In columns 7-8 only state schools located in "micro-regions"[9] that are at the boundary between São Paulo and its neighbor states are included in the analysis. Similarly, in columns 9-10 only state schools located in municipalities that are at the boundary between São Paulo and its neighbor states are included in the analysis. In all DDD regressions, I exclude states that have had pay for performance programs over the period of analysis as mentioned previously (Rio de Janeiro, Espírito Santo, Amazonas, and Pernambuco).

For a large proportion of students, basic socio-economic information is not available, and as a result when including covariates in the regression the sample size drops by more than half. In order to disentangle selection effects related to the availability of data from the effect of introducing covariates, I also run estimates without covariates using the restricted sample of students for which covariates are available. While I do not include these estimates in the tables for simplicity, they show that in cases where coefficients change with the introduction of covariates, the variation comes mainly from selection effects related to the availability of data, and that once we restrict the sample to students with basic socio-economic information, the coefficients remain practically unchanged when including covariates.

The estimated coefficients indicate that the program had overall positive and significant

[9] Micro-regions are administrative divisions which include groups of municipalities based on socio-economic similarities. The state of São Paulo has 645 municipalities, grouped in 63 micro-regions.

effects for 5th grade students. The size of the treatment effect varies according to the counterfactual chosen: overall gains in math range from 0.04 to 0.24 standard deviations, averaging 0.15 standard deviations, while gains in language are more modest, varying from close to zero to 0.14 standard deviations, and averaging 0.09 standard deviations. Interestingly, the estimated coefficients tend to increase as I progressively restrict the sample to geographically closer regions, suggesting unobserved factors might be leading to an underestimation of the effect size when using larger samples. The estimated effects of the program are more modest for 9th grade students, and the coefficients obtained are not statistically significant in several specifications. Improvements in math average 0.06 standard deviations, and overall effects in language average 0.03 standard deviations, although in the latter case coefficients are close to zero in most specifications. Overall, the estimated program impact is not constant across years: the coefficients obtained are generally smaller in 2011 than in 2009, but tend to increase again in 2013.

Tables A1 and A3 in the Appendix present simple DD estimates using the same counterfactuals used in previous triple differences estimations. These estimates show the coefficients obtained without taking into account the differential trends in student performance across states. Estimates for 5th grade students show positive and significant coefficients in most cases, although the coefficients are negative when using neighbor states as the comparison group. Meanwhile, simple DD estimates for 9th grade students show modest negative effects for both math and language test scores. Trends in student performance from these counterfactual groups are shown in Figures B1-B5 in the Appendix.

Tables A2 and A4 in the Appendix show placebo DD estimates using the same counterfactuals, but assuming municipal schools in São Paulo are the treated group. Results show that over the period of analysis the performance of municipal schools from the state of São Paulo has evolved less favorably than the performance of municipal schools from other states. The magnitude of the estimated coefficients is similar for 5th and 9th grade students. This highlights the importance of using triple differences to account for state-specific trends that are unrelated to the program and may bias simple DD estimates. Final DDD estimates are obtained by subtracting placebo DD estimates from simple DD estimates, and provide more stable coefficients across counterfactuals.

5.2 Placebo Test

In this subsection, I use 2005 *Prova Brasil* data to check whether São Paulo state schools were already experiencing larger improvements in student performance relative to the counterfactuals groups considered prior to the implementation of the teacher incentive program. A limitation of this exercise, however, is that I only have access to data for 5th grade students in 2005. By using the same empirical strategy and sample of schools as in previous estimations, I estimate a placebo treatment effect assuming the program took place in 2007. The results, shown in Table A5 in the Appendix, indicate that the performance of 5th grade students in São Paulo state schools deteriorated by around 0.1 to 0.3 standard deviations over this period depending on the counterfactual group used, with similar coefficients for math and language. This might indicate that previous results are a lower bound for the actual impact of the program, although it is not possible to rule out the possibility that this deterioration was related to a specific occurrence.

6 Heterogeneous Effects

6.1 Number of Teachers and Free-Riding Effects

The theoretical literature on group incentives suggests that free-riding behavior is more likely to occur as the group size increases. In large groups, each individual has less incentive to make efforts as his own impact on the overall outcome is lower, and given that he can expect to benefit from the effort of other workers. Goodman and Turner (2013) present evidence on this mechanism in the context of a group-based teacher incentive program in New York, and show that although the program was ineffective in general, schools with a smaller number of teachers experienced a modest increase in student achievement. Imberman and Lovenheim (2015) also provide evidence on the presence of free-riding among teachers participating in a group incentive program in Houston. They argue that the higher the share of students that a given teacher instructs, the stronger incentives become, as this increases teachers' impact on the overall outcome and reduces incentives to free ride. Accordingly, they show that student performance improved more among students whose teachers taught a larger share of students.

I take a similar approach as Goodman and Turner (2013), and look at whether the number of teachers teaching a specific grade affects the impact of the São Paulo teacher incentive program. Information on the number of teachers is obtained from 2007 School Census data. In 2007, the average number of 5th and 9th grade teachers in São Paulo state schools was 14 and 34 respectively. In order to assess how the number of teacher affects the impact of the program, I divide treated schools into four quartiles according to their number of teachers in 2007, and create dummies for three of the quartiles (d2, d3 and d4). I then interact each of these dummies with the treatment effect. This method has the advantage of giving more complete information than simply interacting the treatment variable with the number of teachers, while allowing other coefficients to be the same for all schools. One limitation of this approach is that there can be other factors related to the number of teachers which also affect student performance, although the inclusion of school fixed effects in the estimations already controls for the influence of school size.

Results are shown in Table 4. For simplicity, I only present DD estimates using municipal schools from the State of São Paulo as the comparison group, as estimates are very similar using other counterfactuals. The first line shows the overall program effects for reference, while the other coefficients are obtained from estimating equation (1) with the inclusion of the interacted variables mentioned above (the number of teachers is already captured by school fixed effects and is not included in the regressions). The estimates show that the number of teachers does not significantly impact the performance of 5th grade students. However, the performance of 9th grade students seems to be negatively affected by the number of teachers in a given school, although the estimated effects are modest. While 9th grade students from schools in the lowest quartile show no improvement in language test scores and a modest improvement of 0.03 standard deviations in math, the performance of those in the highest quartile deteriorates by 0.05-0.06 standard deviations (obtained by subtracting the highest quartile coefficient to the reference group coefficient).

6.2 Heterogeneity by Previous School Performance

In this section I assess whether the program's impact varies depending on schools' previous performance. The fact that the design of the program determines that rewards are proportional to schools' improvement (and not only given to schools that attain a discrete performance threshold), could be expected to lead to improvements in the entire distribution of schools, according to Lazear (2003). Nevertheless, it is possible that the program does not affect all schools equally if for example initially low-performing schools have more scope for improvement, or on the contrary lack the skills and capacity to respond to the program and improve performance.

To assess how treatment effects differ according to schools' initial performance, I take a similar approach as in the previous section and divide treated schools into four quartiles according to the distribution of test scores in 2007. Quartile dummies are then interacted with the treatment dummy. Table 5 reports the estimated coefficients. As in previous estimations, I only present results using municipal schools from the State of São Paulo as the counterfactual group.

The coefficients point to considerably different effects according to schools' initial performance. In the case of 5th grade students, while schools in the lowest quartile show an improvement of around 0.2 standard deviations in language scores, this effect decreases for schools in the following quartiles and drops to close to zero for schools in the highest quartile. A similar trend is observed for 5th grade math outcomes, although all types of schools experience positive gains in this case. Schools serving 9th grade students in the lowest quartile show gains of 0.07-0.08 standard deviations in both math and language, while schools in the highest quartile experience a deterioration in performance of around 0.1 standard deviations.

7 Conclusion

Despite being highly controversial, monetary incentives for teachers are an increasingly popular policy. However, the effects of this type of policy are not clear from a theoretical perspective and empirical evidence is limited so far. This paper contributes to the literature on teacher incentives and school accountability by studying the impact of a large teacher incentive program in the State of São Paulo, Brazil, which awarded bonuses at the group level for teachers and school staff conditional on improvements in student performance.

The results suggest the program had overall positive effects in performance, but that gains were more modest and less robust across specifications for 9th grade students than for 5th grade students. A potential explanation for this finding lies in the fact that 5th grade students generally have one main teacher, while 9th grade students have different teachers for different subjects taught, which might make coordination more difficult in the context of a group incentive program. Additionally, it may be easier to improve the performance of younger students, while older students may have learning gaps that are more difficult to close. The results also point to stronger achievement gains in math than in language for both grades, in line with findings from Muralidharan and Sundararaman (2011) and Lavy (2009), suggesting it might be more difficult to improve language skills in the short-term.

The fact that the magnitude of the estimated coefficients varies across specifications points to the difficulty of finding an appropriate counterfactual in difference-in difference analysis and the importance of dealing with potential confounding effects. In this study, I try to deal with state-specific trends that could potentially affect results through triple differences estimations, and assess the robustness of the results by using a series of different comparison groups.

The results also shows that the impact of the incentive program varies according to schools' initial characteristics. For 9th grade students, the number of teachers in a given school is associated with lower gains from the program, although the estimated effects are modest. More sizeable differences are found according to school's previous performance. Initially low-performing schools improved much more than the average, suggesting there may be considerable differences in the ability of schools to respond to this type of policy. Further research is needed to understand the factors behind the heterogeneity in gains among grades, subjects, and schools' initial characteristics; so that teacher incentive programs can be better targeted and more efficient in the future.

References

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, 25(1):95–135.

Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H., and Wilson, D. (2009). Evaluating the impact of performance-related pay for teachers in England. *Labour Economics*, 16(3):251–261.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-in-Differences Estimates? *The Quarterly Journal of Economics*, 119(1)(February):249–75.

Besley, T. and Ghatak, M. (2005). Competition and Incentives with Motivated Agents. *American Economic Review*, 95(3):616–636.

Fryer, R. (2013). Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, 31(2).

Glazerman, S. and Seifullah, A. (2012). An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years. Final Report. *Mathematica Policy Research, Inc.*, pages 1–106.

Glewwe, P., Ilias, N., and Kremer, M. (2010). Teacher Incentives. *American Economic Journal: Applied Economics*, 2(3):205–227.

Goodman, S. F. and Turner, L. J. (2013). The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. *Journal of Labor Economics*, 31(2):409–420.

Hanushek, E. a., Rivkin, S. G., and Taylor, L. L. (1996). Aggregation and the Estimated Effects of School Resources. *The Review of Economics and Statistics*, 78(4):611–627.

Holmstrom, B. (1982). Moral Hazard in Teams. *The Bell Journal of Economics*, 13(2):324–340.

Holmstrom, B. and Milgrom, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, 7(Special Issue):24–52.

Imberman, S. A. and Lovenheim, M. F. (2015). Incentive Strenght and Teacher Productivity: Evidence from a Group-based Teacher Incentive Pay System. *The Review of Economic Studies*, 97(2):364–386.

Lavy, V. (2002). Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement. *Journal of Political Economy*, 110(6):1286–1317.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review*, 99(5):1979–2011.

Lazear, E. P. (2003). Teacher incentives. *Swedish Economic Policy Review*, 10:179–214.

Martins, P. S. (2009). Individual Teacher Incentives, Student Achievement, and Grade Inflation. *Applied Economics*, (4051):43.

Muralidharan, K. and Sundararaman, V. (2011). Teacher Performance Pay : Experimental Evidence from India. *Journal of Political Economy*, 119(1):39–77.

Neal, D. (2011). The Desgin of Performance Pay in Education. *Handbook of Economics of Education*, Vol. 4.

Oshiro, C. H., Scorzafave, L. G., and Dorigan, T. A. (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(2):213–249.

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

Rockoff, J. E. . (2004). The Impact of Individual Teachers on Student Achievement : Evidence from Panel Data. *American Economic Review*, 94(2):247–252.

Springer, M. G., Dale, B., Hamilton, L., Le, V.-N., Lockwood, J., McCaffrey, D., Pepper, M., and Stecher, B. M. (2010). Teacher Pay for Performance Exprimental Experimental Evidence from the Project on Incentives Teaching. *National Center on Performance Incentives, Project on Incentives in Teaching.*

Table 1: Descriptive statistics (year 2007)

	Brazil (all schools)	São Paulo (all schools)	Brazil (schools w/ 4 years of data)	São Paulo (schools w/ 4 years of data)
	(1)	(2)	(3)	(4)
5th grade				
No. of test-takers	2,293,687	580,675	1,809,660	441,555
No. of schools	37,104	5,589	25,972	4,120
Avg. test-takers per school	62	104	70	107
% test-takers relative to total	90	92	91	92
% state schools	33	40	30	41
% municipal schools	67	60	70	59
Avg. age	10.8	10.4	10.8	10.4
% black	12	9	12	9
% father w/ high school	34	40	34	40
% mother w/ high school	31	36	32	36
Avg. Test score - language (0-100)	50	52	50	52
Avg. Test score - math (0-100)	51	53	51	54
9th grade				
No. of test-takers	1,785,895	494,487	1,546,368	467,442
No. of schools	27,163	4,666	21,273	4,308
Avg. test-takers per school	66	106	73	109
% test-takers relative to total	82	84	81	84
% state schools	69	80	69	80
% municipal schools	31	20	31	20
Avg. age	15.4	15	15.3	15
% black	11	10	11	10
% father w/ high school	30	34	31	35
% mother w/ high school	31	32	32	32
Avg. Test score - language (0-100)	57	58	57	58
Avg. Test score - math (0-100)	57	57	57	57

Note: 5th grade students are graded in a scale of 0-350 for language (Portuguese) and 0-375 for math, while 9th grade students are graded in a scale of 0-400 for Portuguese and 0-425 for math. For ease of comparability, grades are rescaled in a range of 0-100.

Table 2: Difference-in-differences and triple differences estimates - 5th grade

Comparison group	São Paulo municipal schools (DD)		Rest of Brazil (DDD)		Neighbor states (DDD)		Neighbor micro-regions (DDD)		Neighbor municipalities (DDD)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Language (Portuguese)										
Overall effect	0.086***	0.111***	0.093***	0.065***	0.108***	-0.007	0.125***	0.082**	0.140***	0.120**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)	(0.04)	(0.06)
2009	0.098***	0.119***	0.110***	0.078***	0.104***	-0.018	0.124***	0.079*	0.131***	0.106
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.05)	(0.07)
2011	0.051***	0.073***	0.068***	0.033***	0.081***	-0.031*	0.077**	0.032	0.122**	0.096
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)	(0.05)	(0.05)	(0.07)
2013	0.117***	0.153***	0.102***	0.087***	0.146***	0.038**	0.184***	0.139***	0.172***	0.166**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.06)	(0.07)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	1,567,525	659,445	5,843,338	2,226,555	3,017,941	1,202,437	478,137	198,413	188,325	76,653
Math										
Overall effect	0.120***	0.152***	0.116***	0.095***	0.135***	0.042***	0.210***	0.181***	0.242***	0.244***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.05)	(0.06)
2009	0.129***	0.156***	0.117***	0.091***	0.104***	0.005	0.187***	0.185***	0.238***	0.263***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)	(0.05)	(0.06)	(0.07)
2011	0.086***	0.115***	0.101***	0.072***	0.125***	0.037**	0.175***	0.139***	0.246***	0.251***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.06)	(0.07)
2013	0.153***	0.203***	0.134***	0.134***	0.188***	0.106***	0.278***	0.222***	0.242***	0.208***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.05)	(0.06)	(0.06)	(0.08)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	1,567,287	659,425	5,842,494	2,226,402	3,017,504	1,202,372	478,069	198,399	188,307	76,650

Note: Outcomes are standardized test scores. All regressions use school and year fixed effects. Student level controls include: gender, race (a dummy=1 for black students), mother and father education (a dummy=1 if the mother/father have completed high school). Standard errors clustered at the school level in parentheses.

$^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$

Table 3: Difference-in-differences and triple differences estimates - 9th grade

Comparison group	São Paulo municipal schools (DD)		Rest of Brazil (DDD)		Neighbor states (DDD)		Neighbor micro-regions (DDD)		Neighbor municipalities (DDD)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Language (Portuguese)										
Overall effect	-0.027***	-0.023***	-0.002	-0.006	0.000	-0.006	0.119***	0.117***	0.078	0.066
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.06)	(0.06)
2009	-0.015	-0.011	-0.010	-0.012	-0.043***	-0.041**	0.107**	0.112**	0.085	0.074
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)	(0.05)	(0.07)	(0.06)
2011	-0.053***	-0.054***	-0.031***	-0.044***	0.032*	0.013	0.073*	0.043	0.044	-0.001
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.07)	(0.07)
2013	-0.012	-0.006	0.037***	0.035***	0.013	0.017	0.181***	0.193***	0.108	0.123
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)	(0.06)	(0.08)	(0.08)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	1,796,341	1,190,513	5,398,627	3,387,310	3,103,450	2,032,720	472,196	320,711	182,931	123,970
Math										
Overall effect	-0.012	-0.008	0.023**	0.018*	0.039***	0.031**	0.150***	0.136***	0.105*	0.080
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)	(0.04)	(0.06)	(0.06)
2009	0.006	0.006	0.022*	0.017	0.010	0.007	0.117**	0.110**	0.104*	0.081
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)	(0.05)	(0.06)	(0.06)
2011	-0.051***	-0.053***	-0.017	-0.029**	0.048***	0.032*	0.087*	0.053	0.044	-0.005
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)	(0.05)	(0.07)	(0.07)
2013	0.011	0.018	0.064***	0.064***	0.060***	0.059***	0.246***	0.244***	0.171**	0.164**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)	(0.06)	(0.08)	(0.08)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	1,796,295	1,190,496	5,398,269	3,387,163	3,103,324	2,032,677	472,192	320,711	182,928	123,976

Note: Outcomes are standardized test scores. All regressions use school and year fixed effects. Student level controls include: gender, race (a dummy=1 for black students), mother and father education (a dummy=1 if the mother/father have completed high school). Standard errors clustered at the school level in parentheses.

$^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

Table 4: Heterogeneous effects by number of teachers

Comparison group	São Paulo municipal schools (DD)			
	5th grade		9th grade	
	(1)	(2)	(3)	(4)
	Language (Portuguese)			
Overall effect	0.093***	0.098***	-0.021**	-0.019**
	(0.01)	(0.01)	(0.01)	(0.01)
Treatment (ref. group=1)	0.091***	0.098***	0.011	0.009
	(0.01)	(0.01)	(0.01)	(0.01)
Treatment*d2	-0.008	-0.015	-0.021*	-0.015
	(0.02)	(0.02)	(0.01)	(0.01)
Treatment*d3	0.008	0.015	-0.045***	-0.042***
	(0.02)	(0.02)	(0.01)	(0.01)
Treatment*d4	0.009	0.003	-0.056***	-0.053***
	(0.01)	(0.01)	(0.01)	(0.01)
Controls	No	Yes	No	Yes
Obs.	16,477	16,473	17,227	17,226
	Math			
Overall effect	0.132***	0.137***	-0.007	-0.004
	(0.01)	(0.01)	(0.01)	(0.01)
Treatment (ref. group=1)	0.125***	0.132***	0.025**	0.026**
	(0.01)	(0.01)	(0.01)	(0.01)
Treatment*d2	0.033	0.026	-0.021*	-0.017
	(0.02)	(0.02)	(0.01)	(0.01)
Treatment*d3	-0.021	-0.013	-0.050***	-0.045***
	(0.02)	(0.02)	(0.01)	(0.01)
Treatment*d4	-0.009	-0.014	-0.057***	-0.055***
	(0.02)	(0.02)	(0.01)	(0.01)
Controls	No	Yes	No	Yes
Obs.	16,477	16,473	17,227	17,226

Note: Outcomes are standardized test scores. All regressions use school and year fixed effects. School level controls include: % of girls, % of black students, % of students whose mother completed high school, % of students whose father completed high school. Only schools with at least 10 test-takers are included in the regressions. Standard errors clustered at the school level in parentheses. The number of observations range from 16,319 to 17,227 for 9th grade estimations, and from 15,845 to 16,477 for 5th grade estimations.

$^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$

Table 5: Heterogeneous effects by initial school performance

Comparison group	São Paulo municipal schools (DD)			
	5th grade		9th grade	
	(1)	(2)	(3)	(4)
	Language (Portuguese)			
Overall effect	0.093***	0.098***	-0.021**	-0.019**
	(0.01)	(0.01)	(0.01)	(0.01)
Treatment (ref. group=1)	0.213***	0.215***	0.088***	0.072***
	(0.01)	(0.01)	(0.01)	(0.01)
Treatment*d2	-0.116***	-0.113***	-0.095***	-0.080***
	(0.02)	(0.02)	(0.01)	(0.01)
Treatment*d3	-0.136***	-0.136***	-0.131***	-0.108***
	(0.02)	(0.01)	(0.01)	(0.01)
Treatment*d4	-0.226***	-0.220***	-0.209***	-0.180***
	(0.02)	(0.02)	(0.01)	(0.01)
Controls	No	Yes	No	Yes
Obs.	16,477	16,473	17,227	17,226
	Math			
Overall effect	0.132***	0.137***	-0.007	-0.004
	(0.01)	(0.01)	(0.01)	(0.01)
Treatment (ref. group=1)	0.212***	0.218***	0.083***	0.076***
	(0.01)	(0.01)	(0.01)	(0.01)
Treatment*d2	-0.057***	-0.058***	-0.065***	-0.057***
	(0.02)	(0.02)	(0.01)	(0.01)
Treatment*d3	-0.094***	-0.099***	-0.117***	-0.102***
	(0.02)	(0.02)	(0.01)	(0.01)
Treatment*d4	-0.169***	-0.168***	-0.180***	-0.162***
	(0.02)	(0.02)	(0.01)	(0.01)
Controls	No	Yes	No	Yes
Obs.	16,477	16,473	17,227	17,226

Note: Outcomes are standardized test scores. All regressions use school and year fixed effects. School level controls include: % of girls, % of black students, % of students whose mother completed high school, % of students whose father completed high school. Only schools with at least 10 test-takers are included in the regressions. Standard errors clustered at the school level in parentheses.

${}^{*}p < 0.1, {}^{**} p < 0.05, {}^{***} p < 0.01$

Appendix

A Tables

Table A.1: Simple difference-in-differences estimates - 5th grade

Comparison group	Rest of Brazil		Neighbor states		Neighbor micro-regions		Neighbor municipalities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Language (Portuguese)				
Overall effect	0.029***	0.042***	-0.074***	-0.055***	0.009	0.029	0.030	0.063
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.03)	(0.05)
2009	0.064***	0.078***	-0.055***	-0.050***	0.027	0.044	0.025	0.053
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)	(0.04)	(0.05)
2011	-0.035***	-0.018*	-0.116***	-0.085***	-0.053	-0.032	-0.018	0.014
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)	(0.04)	(0.05)
2013	0.064***	0.063***	-0.042***	-0.026*	0.059*	0.075*	0.094**	0.130**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)	(0.05)	(0.06)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	1,788,321	713,597	1,027,571	422,536	84,532	36,513	51,596	22,057
					Math			
Overall effect	0.104***	0.128***	-0.023**	0.006	0.105***	0.150***	0.143***	0.202***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)	(0.04)	(0.05)
2009	0.137***	0.157***	-0.030***	-0.016	0.105***	0.160***	0.132***	0.207***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.04)	(0.06)
2011	0.048***	0.076***	-0.054***	-0.02	0.057	0.104**	0.115***	0.180***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.04)	(0.06)
2013	0.131***	0.152***	0.028**	0.066***	0.164***	0.185***	0.192***	0.217***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.05)	(0.05)	(0.06)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	1,788,321	713,597	1,027,439	422,521	84,509	36,509	51,582	22,058

Note: Outcomes are standardized test scores. Only state schools are included in the regressions. All regressions use school and year fixed effects. Student level controls include: gender, race (a dummy=1 for black students) mother education, father education (a dummy=1 if the mother/father have completed high school). Standard errors clustered at the school level in parentheses.

$^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$

Table A.2: Placebo difference-in-differences estimates - 5th grade

Comparison group	Rest of Brazil		Neighbor states		Neighbor micro-regions		Neighbor municipalities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Language (Portuguese)				
Overall effect	-0.063***	-0.022***	-0.180***	-0.048***	-0.116***	-0.051**	-0.109***	-0.055
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
2009	-0.046***	0.001	-0.159***	-0.032***	-0.097***	-0.035	-0.106***	-0.052
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.04)
2011	-0.104***	-0.051***	-0.197***	-0.054***	-0.130***	-0.062**	-0.140***	-0.081**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.03)	(0.04)
2013	-0.038***	-0.023**	-0.188***	-0.063***	-0.124***	-0.063**	-0.078**	-0.034
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.03)	(0.04)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	4,055,017	1,512,958	1,990,370	779,901	393,605	161,900	136,729	54,596
				Math				
Overall effect	-0.011	0.033***	-0.157***	-0.035***	-0.104***	-0.03	-0.085***	-0.027
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.04)
2009	0.020***	0.066***	-0.134***	-0.021*	-0.082***	-0.024	-0.084**	-0.034
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.04)	(0.04)
2011	-0.053***	0.006	-0.179***	-0.048***	-0.119***	-0.033	-0.119***	-0.057
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.03)	(0.04)
2013	-0.003	0.019*	-0.160***	-0.039***	-0.114***	-0.035	-0.05	0.012
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.03)	(0.03)	(0.05)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	4,054,419	1,512,850	1,990,065	779,851	393,560	161,890	136,725	54,592

Note: Outcomes are standardized test scores. Only municipal schools are included in the regressions. All regressions use school and year fixed effects. Student level controls include: gender, race (a dummy=1 for black students) mother education, father education (a dummy=1 if the mother/father have completed high school). Standard errors clustered at the school level in parentheses.

$^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$

Table A.3: Simple difference-in-differences estimates - 9th grade

Comparison group	Rest of Brazil		Neighbor states		Neighbor micro-regions		Neighbor municipalities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Language (Portuguese)				
Overall effect	-0.039***	-0.038***	-0.068***	-0.059***	-0.018	-0.007	-0.027	-0.028
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
2009	-0.057***	-0.047***	-0.087***	-0.067***	-0.013	0.007	0.001	0.008
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
2011	-0.031***	-0.031***	-0.067***	-0.064***	-0.032**	-0.034**	-0.052**	-0.068**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)
2013	-0.030***	-0.034***	-0.049***	-0.044***	-0.008	0.003	-0.03	-0.033
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	3,795,289	2,423,248	2,504,379	1,647,043	379,193	256,892	144,396	97,773
				Math				
Overall effect	-0.033***	-0.026***	-0.049***	-0.036***	0.004	0.016	0.013	0.01
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
2009	-0.039***	-0.034***	-0.054***	-0.040***	0.001	0.02	0.019	0.022
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)
2011	-0.071***	-0.065***	-0.092***	-0.082***	-0.047***	-0.048***	-0.054**	-0.071**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
2013	0.011*	0.018***	0	0.013	0.064***	0.074***	0.077***	0.075**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.03)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	3,795,069	2,423,157	2,504,283	1,647,006	379,195	256,894	144,395	97,778

Note: Outcomes are standardized test scores. Only state schools are included in the regressions. All regressions use school and year fixed effects. Student level controls include: gender, race (a dummy=1 for black students) mother education, father education (a dummy=1 if the mother/father have completed high school). Standard errors clustered at the school level in parentheses.

$^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

Table A.4: Placebo difference-in-differences estimates - 9th grade

Comparison group	Rest of Brazil		Neighbor states		Neighbor micro-regions		Neighbor municipalities	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Language (Portuguese)								
Overall effect	-0.037***	-0.031***	-0.069***	-0.053***	-0.137***	-0.123***	-0.106*	-0.093
	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.06)	(0.06)
2009	-0.046***	-0.035***	-0.044***	-0.026*	-0.120***	-0.104**	-0.084	-0.064
	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)	(0.06)	(0.06)
2011	0.000	0.013	-0.099***	-0.077***	-0.104**	-0.077*	-0.096	-0.065
	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.04)	(0.06)	(0.07)
2013	-0.067***	-0.068***	-0.063***	-0.060***	-0.188***	-0.189***	-0.138*	-0.153*
	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)	(0.05)	(0.08)	(0.08)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	1,603,338	964,062	599,071	385,677	93,003	63,819	38,535	26,197
Math								
Overall effect	-0.056***	-0.044***	-0.088***	-0.066***	-0.145***	-0.118***	-0.092*	-0.069
	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.05)	(0.05)
2009	-0.061***	-0.051***	-0.064***	-0.046***	-0.116***	-0.088**	-0.085	-0.059
	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)	(0.04)	(0.06)	(0.05)
2011	-0.054***	-0.035***	-0.140***	-0.113***	-0.135***	-0.100**	-0.098	-0.064
	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.06)	(0.07)
2013	-0.054***	-0.045***	-0.060***	-0.045**	-0.183***	-0.168***	-0.094	-0.086
	(0.01)	(0.01)	(0.02)	(0.02)	(0.05)	(0.05)	(0.07)	(0.08)
Controls	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	1,603,200	964,006	599,041	385,671	92,997	63,817	38,533	26,198

Note: Outcomes are standardized test scores. Only municipal schools are included in the regressions. All regressions use school and year fixed effects. Student level controls include: gender, race (a dummy=1 for black students) mother education, father education (a dummy=1 if the mother/father have completed high school). Standard errors clustered at the school level in parentheses.
$^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$

Table A.5: Placebo estimates 2005-2007 - 5th grade

Comparison	São Paulo		Rest of Brazil		Neighbor states		Neighbor micro-regions		Neighbor municipalities	
group	municipal schools (DD)		(DDD)		(DDD)		(DDD)		(DDD)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Language (Portuguese)										
Placebo effect	-0.201***	-0.204***	-0.230***	-0.206***	-0.336***	-0.272	-0.139***	-0.157**	-0.155***	-0.174***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.04)	(0.04)	(0.05)	(0.05)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Obs.	613,959	308,177	2,607,918	1,167,448	1,328,752	598,306	201,215	95,724	82,556	37,736
Math										
Placebo effect	-0.176***	-0.176***	-0.210***	-0.194***	-0.295***	-0.253***	-0.114**	-0.158***	-0.141**	-0.174**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.06)	(0.07)
Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	613,827	308,161	2,607,593	1,167,381	1,328,538	598,274	201,181	95,715	82,539	37,734

Note: outcomes are standardized test scores. All regressions use school and year fixed effects. Student level controls include: gender, race (a dummy=1 for black students), mother and father education (a dummy=1 if the mother/father have completed high school). Standard errors clustered at the school level in parentheses.

${}^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$

B Figures

Figure B.1: Trends in student performance - state and municipal schools in São Paulo

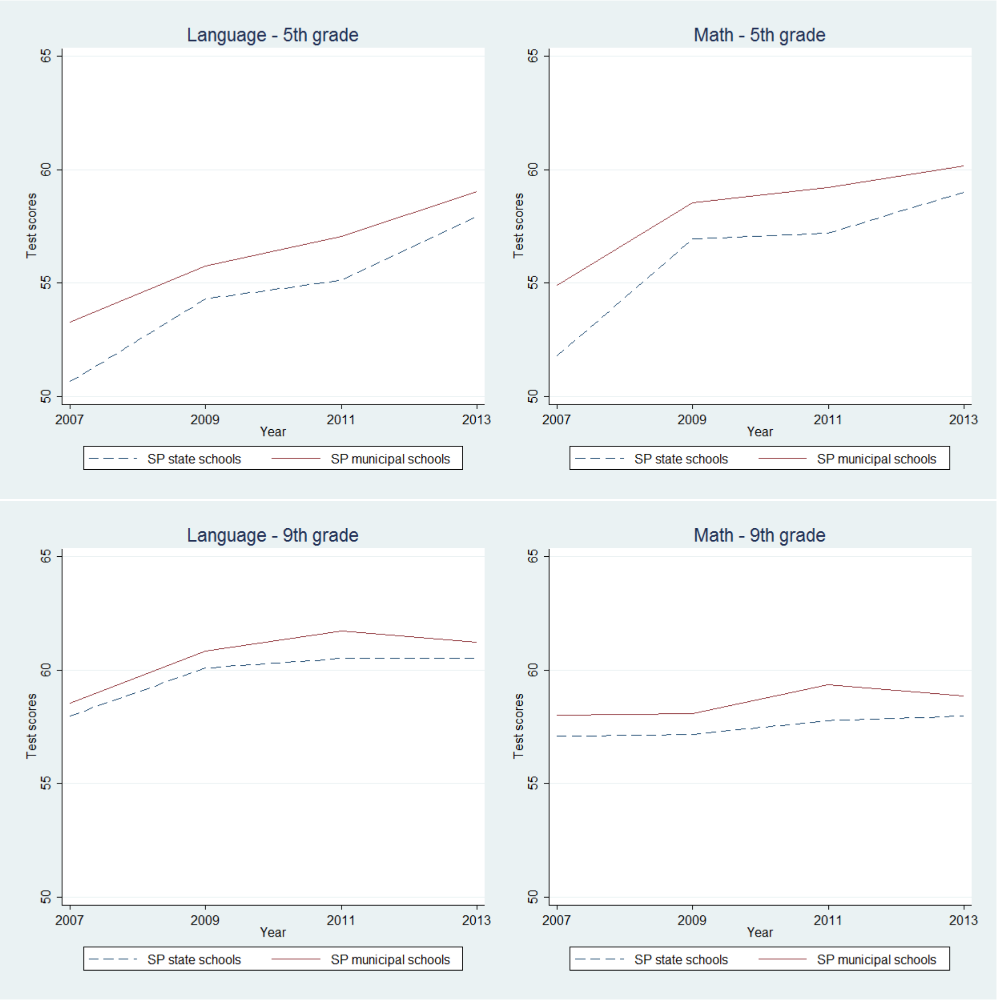


Figure B.2: Trends in student performance - state schools in São Paulo and the rest of Brazil

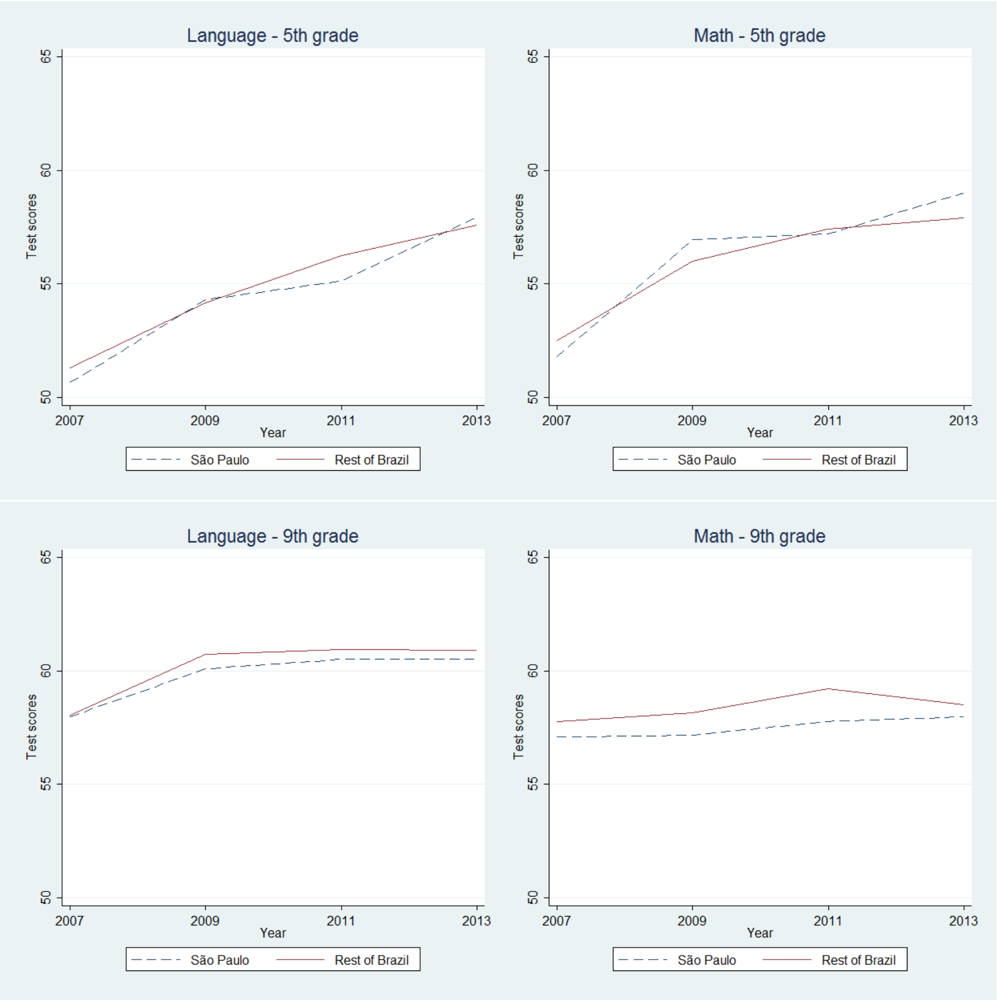


Figure B.3: Trends in student performance - state schools in São Paulo and neighbor states

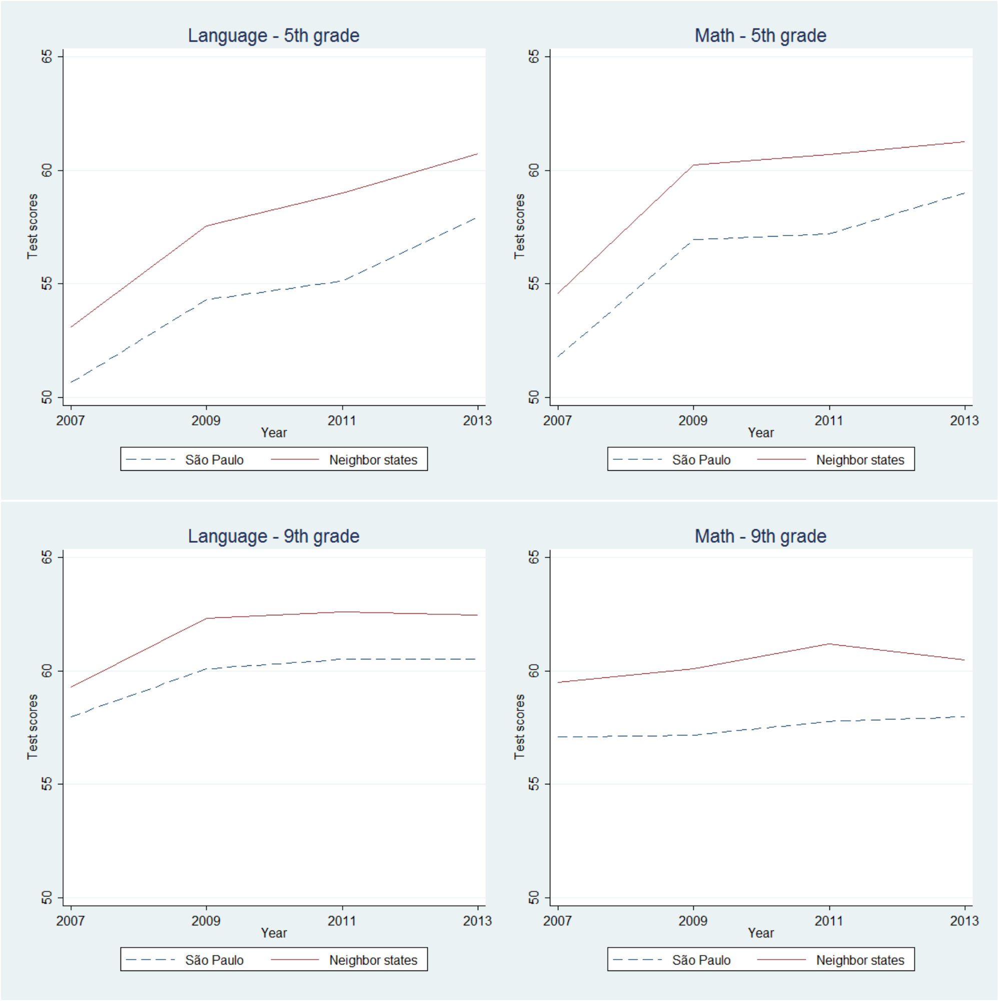


Figure B.4: Trends in student performance - state schools in São Paulo and neighbor states, adjacent micro-regions only

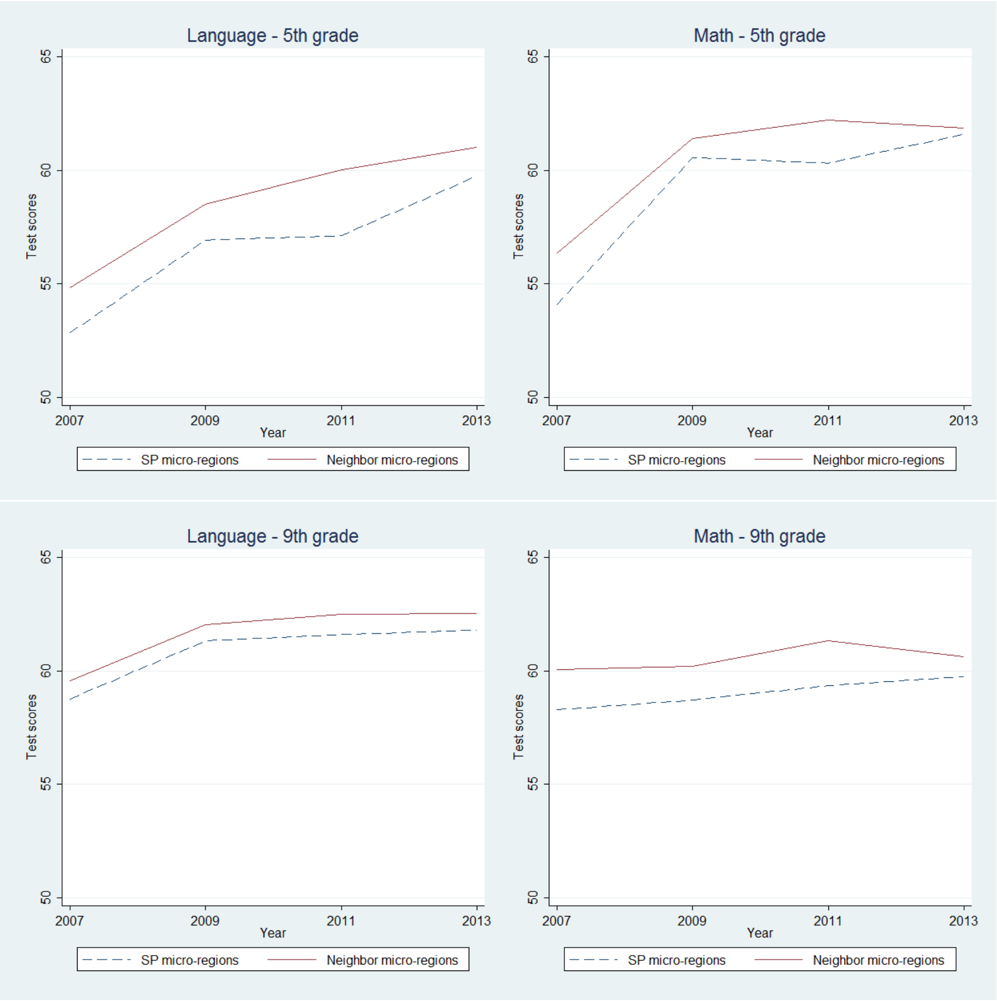
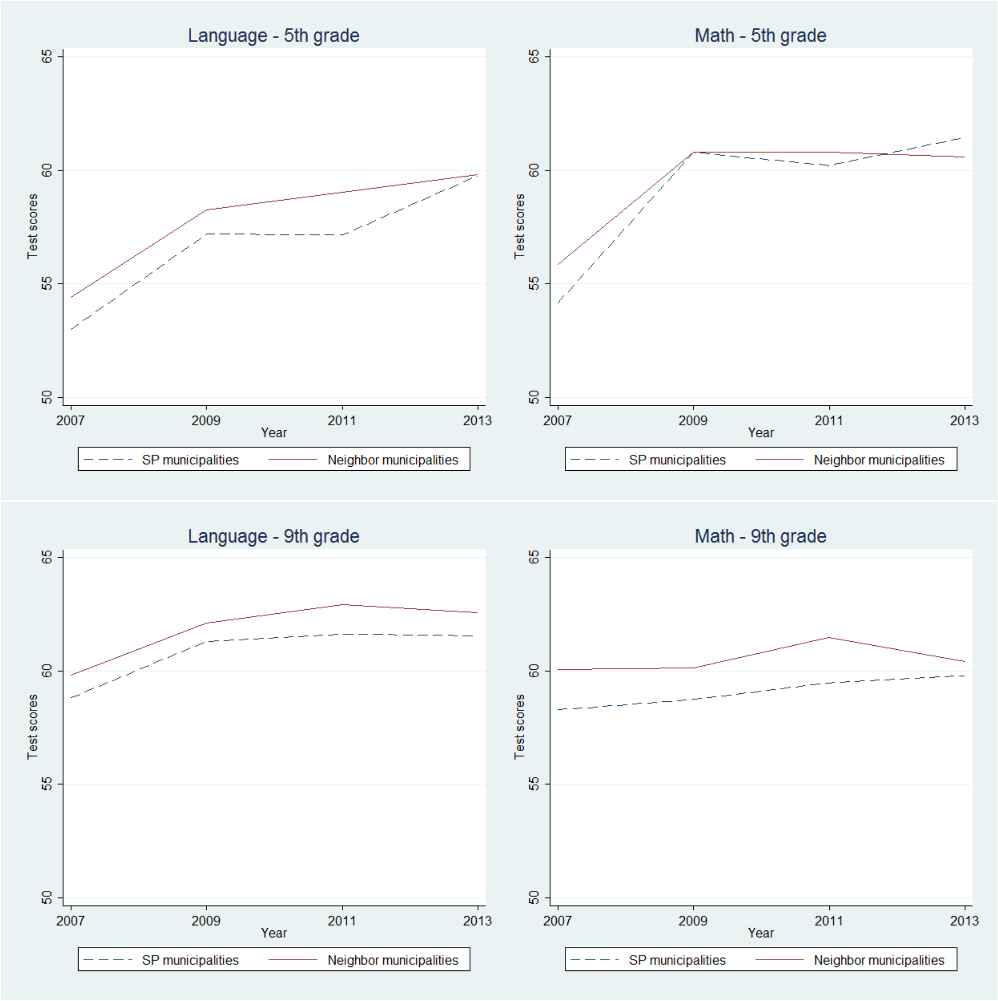


Figure B.5: Trends in student performance - state schools in São Paulo and neighbor states, adjacent municipalities only


Language - 5th grade
Math - 5th grade
Language - 9th grade
Math - 9th grade
Test scores
Year
SP municipalities
Neighbor municipalities