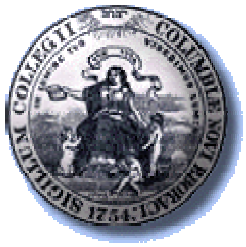

Columbia University

Department of Economics
Discussion Paper Series

Group Incentives for Teachers: The Impact of the NYC School-Wide Bonus Program on Educational Outcomes

Sarena Goodman
Lesley Turner

Sarena Goodman and Lesley Turner are PhD students in the Department of Economics, Columbia University. Faculty members Jonah Rockoff and Brendan O'Flaherty have recommended the inclusion of this paper in the Discussion Paper Series.

Discussion Paper No.: 0910-05

Department of Economics
Columbia University
New York, NY 10027

August 2009

Group Incentives for Teachers:
The Impact of the NYC School-Wide Bonus Program on Educational Outcomes

Sarena Goodman
Columbia University

Lesley Turner
Columbia University

August 2009*

Abstract

In current debates regarding the future of education, teacher compensation schemes are often criticized for their lack of performance-based pay. Proponents of merit pay for teachers argue that tying teacher salaries to student achievement will induce teachers to focus on the success of their students and stimulate innovation in the school system as a whole. In this paper, we use a randomized policy experiment conducted in the New York City public school system to explore the effects of one group-based pay scheme. We investigate potential impacts of incentive pay over two academic years (2007-2008 and 2008-2009) on student performance on annual math and reading exams, teacher absences, and responses to environmental surveys of teachers and students. We also consider whether the program had differential outcomes on groups within schools that were especially likely to be targeted, given the particular incentive structure of the program. Last, we explore relative impacts on the market for teachers by examining end-of-year teacher turnover and the quality composition of newly hired teachers. In general, we find no significant effects of this program. However, there is some evidence that the program reduced teacher absenteeism in schools with a small number of teachers, and that these effects were weakened in larger schools by the presence of free-riding.

* Correspondence should be sent to ljt2110@columbia.edu. We are especially grateful to Jonah Rockoff for his thoughtful comments and advice. We would like to also thank Doug Almond, Todd Kumler, Bentley MacLeod, Neil Mehrotra, Petra Persson, Jesse Rothstein, Miguel Urquiola, Till Von Wachter, and Reed Walker for helpful feedback as well as participants in the Columbia applied microeconomics colloquium. **Note:** In the interest of fairness, the ordering of the authors' names was determined by a coin flip.

1. Introduction

Teacher compensation schemes are often criticized for their lack of performance pay and relatively low pay in general. Critics claim that these features of teachers' salaries can lead to sorting and adverse selection in the market for teachers. Proponents of merit pay argue that tying teacher salaries to student achievement will induce teachers to focus on the success of their students and stimulate innovation in the school system as a whole, leading to enhanced efficiency in the public education system. In other sectors, incentive systems are used to extract efficient output from workers.[1] Pay contingent on individual performance can be valuable in a setting where employers are able to measure and reward on-the-job performance. However, education is a complex good and it is difficult to observe and appropriately monitor the behavior of educators and their respective contributions to the production of education. Thus, designing an incentive system compatible with our educational goals may be difficult.

In this paper, we investigate the impact of group-based incentive pay for teachers using a policy experiment conducted in New York City. In the fall of 2007, 185 schools were randomly selected from a pool of high-poverty schools to be eligible for school-level bonuses based primarily on student exam results.[2] We examine the impacts this program had on student performance on math and reading exams in the first and second years following program implementation, teacher effort (as measured by absence rates), and outcomes from surveys of teachers and students, including changes in classroom activities and school-level policies. We find no significant impact of the bonus program on student achievement in the first or second year of the program. Nor do we find any overall impact on teacher absences or changes in policies such as the availability of tutoring or teachers' use of student achievement data in lesson planning in the program's first year.

An individual teacher's ability to affect the probability of receiving a bonus is decreasing in the number of teachers with tested students; thus, free-riding may have dampened the incentive effects of the bonus program. We examine whether this is the case by testing for an

[1] See Macleod and Parent (1999) for an overview of other sectors that employ incentive-based pay. These compensation schemes are generally most effective in sales jobs and those that involve operating machines. Incentive pay is less commonly used and generally less effective in sectors where output is more difficult to measure.

[2] The program also included 39 high-poverty secondary schools. Since the measurable outcomes are different for high schools, we concentrate our analysis on elementary and middle schools and schools serving children in kindergarten through 8th grade (K-8 schools).

interaction of the bonus program with the number of teachers with tested students in a school. We find no evidence of an important free-rider effect on student achievement or school policies. However, we find some evidence that teachers of tested subjects responded to the program by reducing absences, but only in schools with relatively of these teachers.

The structure of the bonus program also contained incentives to target specific subgroups of students, including those at the bottom of the achievement distribution. Despite this incentive, we find some evidence that students with low prior achievement were negatively impacted by the program. Finally, we examine the impact the program had on teacher turnover and the characteristics of newly hired teachers after the first year of the program. We find no evidence that the bonus program reduced teacher turnover or improved the qualifications (e.g., prior experience, certification) of new hires.

We provide an overview of the bonus program in Section 2. Section 3 discusses the difficulties and theoretical implications associated with implementing merit-based pay in schools. Section 4 provides an overview of the data and outlines our empirical framework, Section 5 presents results, and Section 6 concludes.

2. The New York City School-Wide Bonus Program

In the school year 2007-2008, the New York City Department of Education (DOE) established the "School-Wide Bonus Program" (hereafter, the bonus program), under an agreement with the local teachers' union, the United Federation of Teachers (UFT). The DOE randomly selected 185 schools serving kindergarten through eighth grade from a group of schools designated as "high need." These schools were then eligible to participate contingent on a majority vote in favor of the program.[3] Teachers in participating schools could receive a lump-sum bonus if the school met goals based primarily on student achievement.

Schools that achieved a target score or were awarded an "A" accountability grade (explained below) received a pool of bonus money equal to $3,000 per union teacher, while schools that fell short but managed to meet 75 percent of the target score received a pool of bonus money equal to $1,500 per union teacher. Bonuses were distributed across teachers

[3] In order to participate, 55 percent of full-time United Federation of Teachers (UFT) staff in the school needed to vote in favor of the program. A school's principal was also given participation veto power. A total of 25 schools (14% of all eligible schools) did not approve the program.

according to a formula designed by each school's compensation committee, comprised of the principal and two representatives from the teacher's union. The distribution formula was chosen after the student exam period. In a school where the bonus payment was equally distributed across teachers, the full $3,000 award represents a 7 percent increase in the salary of teachers at the bottom of the pay scale and a 3 percent increase for those at the top.[4]

The timing of program announcement and the selection of schools into the treatment group did not allow much leeway for behavioral responses to the program in its first year. Selection into the program and the vote to participate took place in November 2007, less than two months before reading exams were taken in January and less than four months before math exams were taken in March. However, the program continued in the school year 2008-2009, and all schools in the program voted to participate in the second year.[5] Of the 160 treatment schools that voted to participate in the first year of the program, 89 (56 percent) received bonus payments. The bonus pool averaged $160,095 per school, and amounted to a total of $14.2 million district-wide. Bonus payments for the second year of the program have not yet been distributed.

The school year 2007-2008 also marked the implementation of the DOE's new accountability system, under which schools received progress reports and accountability grades designed to summarize a school's overall performance on a multidimensional metric of student learning.[6] Each school's performance was scored relative to the entire district and to a group of "peer schools," with similar student demographic characteristics or prior test scores.[7] Each school's progress report documented its score on this metric, the corresponding accountability grade, and a target score for the following school year. Schools that received lower accountability grades needed to achieve greater gains to reach their target scores.

---

[4] Teacher salary schedules are available at http://www.uft.org/member/contracts/moa/salary_schedules

[5] Four schools participating in the program were closed at the end of the school year 2007-2008, thus 181 schools participated in 2008-2009.

[6] The metric was calculated from measures of school environment (results of a learning environment survey and student attendance), student performance (average student achievement on reading and math exams, median proficiency, and percentage students achieving proficiency), student progress (average change and percent making progress on math and reading exams), and extra credit for exemplary student progress among high-need students.

[7] For elementary schools and those serving kindergarten through eighth grade (K-8), the index was based on a function of the percentage of students that were English language learner (ELL), special education, Title I free lunch, and minority. For middle schools, the peer index was based on the 4th grade reading and math test scores of current students. These different constructions actually encapsulate consistent metrics for relative disadvantage, as the components for the elementary/K-8 peer index are very strong predictors of 4th grade test scores. Therefore, the two methods should yield reasonably close measures.

The details of the accountability system are important for our analysis: schools were selected into the experimental sample based on their peer indices, and teacher bonuses were awarded based on whether a school was able to achieve its target score. Furthermore, the accountability system provided additional incentives to schools participating in the bonus program. Schools that earned an A or B accountability grade received rewards (e.g., principal bonuses, additional funds based on students transferring from schools receiving a poor grade), while schools that received D and F grades faced consequences (e.g., risk of school closure and removal of principal). Estimates of the effects of the citywide accountability system suggest that receiving an F or D had significant and positive impacts on test scores (Rockoff and Turner, 2008). It is important to note that our results estimate the interaction of bonus program and the NYC accountability system; they may be interpreted as the impact of teacher group performance pay in a district where there is already an accountability system providing incentives to schools.

3. Incentive Pay and Teacher Effort

Allowing compensation to vary with output can align workers' incentives with those of the employer, highlighting specific aspects of an employee's job that are the most valued. When a job involves several tasks or when the nature of such tasks are broadly defined, incentive pay can help resolve confusion as to how best to fulfill responsibilities. Additionally, properly-structured performance pay can offset shirking behavior and encourage employees to provide costly effort.[8] If at least some public school teachers exert an inefficiently-low amount of effort or focus their effort on tasks where the marginal returns for society are too low, then merit pay may be desirable. Teachers could respond to incentive payments by increasing effort along several margins; for instance, spending more time on lesson preparation, showing up to school more often, or spending extra time helping students outside of normal class hours.[9]

However, performance pay in the educational sector may not be as effective as it is in other occupations. Education is a complex good: teachers must complete multidimensional tasks and allocate their effort across several activities. Holmstrom and Milgrom (1991) demonstrate

[8] Effort extraction is just one motivation for incentive-based pay. Incentive systems arise for other reasons as well, such as better sorting of workers across jobs or in order to select quantity versus quality of output (Lazear, 1986).
[9] It need not be the case that these teaching activities immediately translate into higher test scores. Rather, it is only necessary that teachers themselves believe that these behaviors are correlated with student achievement. However, over time, we might expect to see persistent or increased use of teaching practices that were successful in the short run and a decrease in the use of those that were not.

that the performance metric to which compensation is tied affects how effort is allocated. Therefore, in designing a system that rewards teachers for performance, two aspects of the design require careful thought: the performance measure used to evaluate performance and teachers' potential responses. While test scores provide a measure of educational output, tying performance pay to testing outcomes may cause teachers to focus on narrowly-defined skills that appear on exams (e.g., "teaching to the test") or overtly manipulate test scores (e.g., Levitt and Jacob, 2003; Jacob, 2005; Figlio, 2006; Figlio and Getzler, 2006; Cullen and Reback, 2006).

3.1 Previous Results on Teacher Incentives

Current systems of performance pay for teachers range anywhere from competitive bonuses drawn from a fixed pot, where the teacher whose classroom or school experiences the greatest gains receives an award (e.g., Florida's 2006-2007 Special Teachers Are Rewarded (STAR) program), to bonuses tied to fixed achievement thresholds (e.g., Mexico's Carrera Magisterial program), and from direct incentives awarding bonuses to individual teachers (e.g., Denver's ProComp program) to group-based incentives (e.g., North Carolina's ABCs of Public Education program), where bonus payments are contingent on school- or district-wide performance. The specifics of how awards are allocated, the size of potential bonuses, and the metrics on which bonuses are based are all important.[10]

The empirical literature on teacher performance pay is relatively new, and has grown as innovative compensation schemes have emerged. Figlio and Kenny (2007) present evidence of a positive cross-sectional relationship between individual-based teacher performance pay and student achievement in U.S. schools. Systems where awards were difficult to earn and only a small number of teachers received incentive payments were most strongly related to student achievement. Experimental evidence on individual teacher incentives in Israel is consistent with these findings (Lavy, 2004). In Lavy's study, teachers were awarded cash prizes for the performance of their class relative to other classes in the same subject. The incentive payments, ranging from 6 to 30 percent of teachers' average annual salary, led to an increase in both the proportion of students taking a high school exit exam and the performance among test-takers.

[10] For example, Neal (2008) discusses the necessary conditions and considerations for constructing an optimal incentive pay system for educators. Even under optimal circumstances (e.g. an assessment exists for every academic skill with perfect reliability), a functional bonus-pay system requires both "a method for ranking schools or teachers according to performance" and "the assignment of specific rewards and penalties to the various performance ranks that schools or teachers may receive."

These student achievement gains likely stemmed from an increase in after-school sessions, evidence of increased teacher effort in response to potential rewards. Muralidharan and Sundararaman (2008) also find a large positive impact of individual teacher incentives using a randomized experiment in India. These effects were present across different grades levels and student competency levels, and the gains made by students persisted through a second year of testing.

There is less evidence on the effectiveness of group-based teacher incentives. In theory, group incentive payments will be the most effective when the production technology is truly joint, and there is some evidence that teachers' productivity is affected by the productivity of their peers (Jackson and Bruegmann, 2009). However, group incentives may dilute a given teacher's marginal benefit to increasing effort, and free-riding among group members may result. Glewwe et al. (2003) examine the effects of a school-based teacher incentive experiment in rural Kenya, where teachers in grades 4 to 8 received a fixed bonus if their school had the highest score within the district or the largest improvement based on performance in a baseline year. The authors find evidence of short-term improvements in test scores but no long-term gains. Lavy (2002) finds that incentive payments based on school-wide performance increased student test scores and participation on matriculation exams in Israel, but the percentage of students who received matriculation certificates, arguably the longer-run outcome of interest, was unchanged. Notably, Muralidharan and Sundararaman (2008) also examine group based incentives, and find similar increases in student achievement as with the individual incentive program. However, it is important to note that the Indian schools in their sample typically contained only a few teachers, mitigating the free-rider problem.

4. Data and Descriptive Results

Data on schools' test scores and information related to the accountability system were collected from publicly available files on the DOE website.[11] Our measures of academic achievement are average math and reading test scores for each school for the school years 2006-2007, 2007-2008, and 2008-2009 (hereafter 2007, 2008, and 2009). We use information on each school's performance under the new NYC accountability system, including each school's

[11] See http://schools.nyc.gov/Accountability/DOEData/default.htm for more details.

accountability grade, target score, and peer index.[12] Lists of schools participating in the bonus program and eligible schools that voted to not participate in the program are also available online.[13]

Our treatment group is includes the 185 schools classified as elementary, middle, and K-8 (schools serving kindergarten through 8th grade) that were eligible for the program. Of these schools, 25 schools voted not to participate. While we do not have information on which schools were potentially eligible but not randomly selected into the treatment group, we do know that all treatment schools were drawn from a group of "high needs" schools with the lowest peer index scores in the city. Thus, we can use treatment schools' peer indices to construct a group of likely control schools. Specifically, we calculate the maximum peer index for each type of school (elementary, middle, K-8) in the treatment group, and assume that all schools below these cut-offs that were not assigned to the treatment group are control schools. Using this methodology, our control group contains 162 schools.

We also use school-level information from annual surveys of teachers and students conducted by the DOE as part of the accountability system near the end of the school year. From these surveys, we construct a group of variables designed to measure teachers' perceptions of school-wide changes and students' reports of classroom activities. Specifically, we use the questions from the student survey on the degree to which: 1) students completed essays and research projects, and 2) classroom activities included group work, class discussions, and "hands-on activities such as science experiments." We also measure the availability of tutoring, using students' responses to questions on whether tutoring was offered before or after school or during free-periods. From the teacher survey, we construct measures of changes in school-wide policies. For instance, administrators could respond to the program by offering teachers additional professional development opportunities. To examine changes of this sort, we use questions on teachers' use of achievement data, such as students' test results from prior years or "periodic examinations" during the school year, to inform their lesson planning, and their views

[12] Middle schools and elementary/K-8 schools have different metrics underlying their respective peer indices that also have different scales. Thus, for descriptive purposes, we standardize each type of school's peer index to have a mean of zero and standard deviation of 1.

[13] The list of participating schools is available online at http://schools.nyc.gov/Accountability/RewardsandConsequences/PrincipalsBonusAnalysis/default.htm). A list of eligible schools that did not vote to participate is available online at: http://www.uft.org/news/issues/press/bonus_vote. Although it does not affect our analysis, it is worth noting that one school appears on both lists.

on the quality of professional development offered. We also create a measure of whether teachers believed students faced high standards and expectations.

We aggregate data from individual students and teachers to examine whether students with particular characteristics experience greater-than-expected gains, given the incentives provided by the accountability system, and whether the program had any effect on teacher absences, a measure of teacher effort.[14] We restrict our attention to absences classified as "self-treated" illnesses and personal days, since absences for other reasons (e.g., severe illness, jury duty, military service, funeral, etc.) are unlikely to be affected by teachers' effort decisions. Additionally, we use aggregated teacher data to test whether the bonus program had an effect on teacher turnover or the characteristics of newly-hired teachers. Finally, we use data on the number of teachers within each school providing instruction in the tested subjects and grades to test for the effects of free-riding.

When evaluating any intervention within an experimental setting, it is important to determine random selection successfully balanced observable characteristics of treatment and control group schools. Table 1 compares the characteristics of treatment and control schools prior to selection into the treatment group.[15] Treatment and control schools are similar in terms of enrollment, accountability outcomes, and student demographics, although teachers in treatment schools were significantly more likely to hold a master's degree and also, on average, had significantly higher absence rates.

However, it is more troubling that treatment schools had significantly higher test scores in both math and reading in 2007, while control schools experienced significantly greater gains in reading scores between 2006 and 2007. We address the first concern by including a control for the outcome of interest in the year prior to the intervention in all specifications (discussed in more detail in the next section). To address the second concern, we examine trends in test scores between the school years 2004-2005 and 2006-2007 to determine whether the growth in test scores differed between these two groups in prior years. If this was the case, we might be concerned that any estimated treatment effect was actually picking up differences in test score

[14] These data are not public, but researchers can apply to use them with the approval of the DOE.
[15] Appendix Table I compares the characteristics of treatment schools by whether or not they voted to participate in the program. Schools voting "no" are largely similar to schools that actually received the treatment, although these 25 schools were relatively less disadvantaged and had higher test scores on average.

trends. However, we find that treatment and control schools display quite similar trends (results available upon request).[16]

5. Regression Framework

The advantage of a randomized experiment in estimating the effect of teacher incentives is that we can apply a very simple regression specification:

$$Y_{jt} = \delta D_{jt} + \varepsilon_{jt},$$

where Y_{jt} is the outcome of interest for school j in year t (for example, average math scores in 2008), ε_{jt} is a stochastic error component, and D_{jt} is an indicator variable for whether the teachers within the school are in the treatment group. The identifying assumption of this approach requires that there be no contemporaneous shock that differentially affects the outcomes of the treatment schools in the same period as the treatment.

However, as shown in Table 1, we find some evidence that, even under random assignment, the treatment and controls schools differed on a few key characteristics, such as previous average test scores. Therefore, our primary regression framework takes the following form:

$$Y_{jt} = \delta D_{jt} + \gamma Y_{jt-1} + \varepsilon_{jt}$$

where δ is the coefficient of interest. We control for the outcome in the year prior to the intervention to address any baseline differences between treatment and control schools. We estimate the equation with ordinary least squares, weighting by group size (e.g., number of students tested when the dependent variable is average math scores, number of teacher survey respondents for teacher survey outcomes). Because our treatment group includes schools that were eligible for the bonus program but voted not to participate, our results should be interpreted as the impact of offering the program to schools. Since almost 90 percent of eligible schools

[16] Table 1 also compares the characteristics of the experimental sample to other schools in New York City serving students in kindergarten through eighth grade that received accountability grades, were not charter schools, and did not only serve special education students. Given that eligible schools were selected based on having a peer index, it is not surprising that the experimental sample differed from the remainder of NYC schools across a number of dimensions. On average, schools in the experimental sample had a higher proportion of English Language Learners (ELL), special education students, minority students, and students eligible for the Title I free lunch program, as well as lower average math and reading scores. Teachers in the experimental sample had slightly less experience and more absences, and experimental schools were smaller, with lower enrollment and fewer teachers.

voted to participate, the "treatment-on-treated" estimate of actually participating in the program would be roughly 15 percent larger in absolute value.

We also estimate a second specification that includes a vector of control variables to increase precision. These controls include school type (i.e., elementary, middle, or K-8), demographic composition (i.e., percentage of students that are ELL, special education, free lunch, and minority), peer index, and accountability score (since this determines a school's target score). Finally, we estimate a third specification that drops schools in the bottom 5 percent of the peer index. This serves as a robustness check for our other estimates since there is a long tail at the bottom of the peer index distribution, which includes two control schools that are extreme outliers.

6. Results

To preview our estimates of the impact of the bonus program on student achievement, we display the distribution of average math and reading scores within treatment and control schools in 2007, 2008, and 2009 (Figures 1 and 2). On average, all schools in the experimental sample experienced an increase in average student performance in the two years following the implementation of the program (as was true of other schools in NYC). If the bonus program had an impact on test scores, we should observe a shift in the distribution among treatment schools, relative to control schools. However, there are no obvious differences in the distribution of test scores in either subject after the baseline year.

Table 2, which displays the regression results estimating the impact of the program on average exam scores, confirms these findings. Columns 1 and 4 contain estimates of the effect of program eligibility on average math and reading test scores, controlling for achievement in the prior year and weighted by the number of students tested. We do not observe any significant impacts on aggregate school performance in either the first or second year of the program, and the point estimates are all negative and quite small.[17] Results do not noticeably change if we include school-level controls or drop outliers. While one might not expect to observe any effects

[17] In 2008, the student level standard deviations of math and reading scores were 39 and 34 points, respectively

in the initial year, the fact that the estimates are very similar in the second year of the program suggests the program had no impact on student achievement.[18]

Treatment schools face different incentives according to their accountability grades. Since schools that received an A on their progress report needed only to maintain this grade, the bonus program may not have provided a large incentive to teachers in these treatment schools to change their behavior. Conversely, both treatment and control schools receiving low grades had additional motivation to improve student test scores, as they faced closure or principal removal if student achievement did not improve in the following year. We investigate whether there is heterogeneity in schools' responsiveness to treatment along this dimension, grouping schools into three separate bins by their accountability grades: A, B/C, and D/F. Our estimates become quite noisy, likely due to the small sample size within each grade-grouping, but are consistent with those discussed above: point estimates are small, we find no significant treatment effects in any group, and we cannot reject equality of effects across the three groups (results available upon request).

6.1 Student and Teacher Survey Results

It is possible that teachers and school administrators responded to the bonus program, but that these behavioral changes did not translate into increased student achievement. Thus, we explore whether the bonus program led to changes in teacher behavior and school policy using results from the DOE's annual surveys of teachers and students. [19] We test whether the program induced any changes in classroom activities by examining the extent to which students reported working on "essays or projects" and "group work or hands-on activities." We also test whether the program increased opportunities for before- or after-school tutoring sessions. Since only students in grades six or higher completed the survey, we lose a number of our schools, mostly at the elementary level. We do not find significant effects of treatment on student reports of participating in group or hands-on learning activities or on whether they completed projects or essays in class, although both of these outcomes are negatively correlated with treatment and, in the third specification, the latter measure comes close to conventional significance levels (Table

[18] Four schools in the treatment group were closed at the end of the 2008 school year, thus, our sample decreases by eight observations for the second set of regressions. Our 2008 results remain unchanged when we restrict the sample to exclude these observations.

[19] For ease of interpreting results, all survey outcomes are standardized within school type to have a mean of zero and standard deviation of one across all NYC schools.

3, Panel A). Additionally, the bonus program appears to have little impact on the availability of tutoring.

Although the bonus program targets teachers, one might also expect it to induce changes in school-wide decisions. However, we do not find evidence of institutional responses to the intervention (Table 3, Panel B). There are no significant treatment effects on teachers' use of student data or the quality of professional development received, and the point estimates of the impact of treatment on these outcomes are quite small. The final measure from the teacher survey we examine – whether teachers believed students in their school were held to high expectations – is negative and approaches conventional significance levels in the second and third specifications. If anything, these results indicate that the bonus program may have weakened standards for students.

6.2 Teacher Effort and the Free-Rider Problem

Although we cannot directly examine many of the dimensions on which the bonus program may have led to higher effort provision on the part of teachers, we can measure whether teachers in treatment schools reduced absences. Absences are more common among teachers than in other occupations, and absenteeism has been shown to have a negative effect on student achievement (Clotfelter et al., 2009; Miller et al., 2008). Using data on absences among New York City teachers, Herrmann and Rockoff (2009) estimate that an additional 10 teacher absences results in a 0.01 standard deviation reduction in test scores.

To explore whether the bonus program had an impact on teacher attendance, we run a series of regressions where the dependent variable is average absences between the months of November 2007, when schools first learned of their eligibility for the bonus program, and March 2008, when the last exams were taken (Table 4).[20] We also separately examine absences among teachers with tested students (e.g., teachers for grades 3 through 5 in elementary schools and math and reading teachers in middle schools). We only consider absences taken for illness and personal business. We exclude days missed due to death in the family, injury, jury duty, absences required by the school system (e.g., for professional development activities), conference

[20] Since treatment schools did not vote until November, 2007, it does not make sense to include earlier absences. If teachers did not expect the bonus program to continue beyond its first year, then the incentives should be weaker following the end of student testing.

attendance, and religious holidays because these are largely outside the teachers' control.[21] The first three columns of Table 4 show little effect of the bonus program on average absences, both among all teachers and among teachers with tested students.

Theory predicts that incentives are weaker when the number of teachers who can directly affect whether a school receives the bonus increases. In the case of the bonus program, the teachers whose effort can directly affect the probability that a school qualifies for the bonus payments are those with tested students. Thus, we examine whether treatment effects are related to the number of teachers with tested students.[22] We first de-mean our measure of the number of such teachers, and then include an interaction with the indicator for treatment; thus, the point estimate for the treatment indicator can be interpreted as the effect for the school with an average number of teachers with tested students (Table 5, columns (4) through (6)). A negative coefficient on the interaction between number of teachers and treatment would provide evidence that the program impacts are diluted by free-riding.

Using estimates from the most conservative specification (column (5)), these results suggest that for schools with fewer than 8 teachers in tested classrooms (schools in the lowest quartile of this variable), the bonus program reduced absences by 0.2 days per teacher, which translates into 1.6 fewer absences over the five month period we examine. Considering the estimates of Herrmann and Rockoff (2009), it is not surprising that we do not see a corresponding increase in achievement even if we allow the effects of the bonus program to vary by the number of teachers in tested classrooms (results available upon request); their estimates would suggest this reduction in absences would translate into a 0.002 standard deviation improvement in test scores.

6.3 Targeting

Although the bonus program had no observable impact on average student achievement, tying bonuses to the structure of the NYC accountability system provided incentives for schools to focus on students at the bottom of the achievement distribution. Therefore, a measure of central tendency, such as mean student achievement, may not fully capture the potential impacts

[21] In support of this notion, absences for illness and personal days are more likely to occur on Mondays and Fridays than other types of absences.

[22] A small number of middle and K-8 schools do not have information on the number of teachers teaching tested subjects, thus, these schools are not included in regressions where the dependent variable is absences among teachers with tested students.

of the program. In line with recent research examining the effect of accountability systems on performance among different student subgroups (Cullen and Reback, 2002; Figlio and Getzler, 2002; Figlio, 2006), we test whether the bonus program led to greater-than-expected achievement gains among particular students whose performance was given greater weight in determining the school's score. Although the NYC accountability system's methodology for scoring schools is fairly complex, it still contains clear incentives to focus on some students more than others. The accountability system awards schools more "points" for the performance of students whose prior-year achievement placed them in the lowest third of their grade, either within their school or within the entire city, those students on the cusp of proficiency and close to the school median (because a school receive points for students making proficiency and for its median score), and for the performance of ELL and special education students. In short, the design of the accountability system creates additional incentives to target students other than those with the highest propensity to do well. [23]

To examine whether the treatment led to greater achievement gains for some groups of students, we examine average test scores of four mutually exclusive groups: students in each third of the distribution of prior year test scores (making three groups), and students who were classified as receiving either ELL services or as special education students. Since students entering third grade will not have prior year test scores, we exclude these students from our sample. Each school-group cell forms a unit of observation and all regressions are weighted by the number of students in the cell with standard errors clustered at the school-level. When we include student characteristics as covariates, these are also calculated within each school-group cell. Table 5 presents these results using a model where the indicator for group is fully interacted with the treatment indicator, and the treatment indicator itself is dropped from the regression. We find no evidence that the math scores of students in the groups given more weight in the accountability system were differentially affected by program eligibility. Indeed, the interaction with being in lowest third of achievement is negative and marginally significant.

[23] Schools might also respond by reclassifying higher performing students as either ELL or special education to take advantage of the increased weight placed on these students' scores. We do not find that the proportion of tested students classified as ELL or special education within treatment schools increased relative to control schools (Appendix Table A2).

6.4 Teacher Characteristics and Turnover

Finally, we investigate whether the program bonus program led to changes in the qualifications of new teachers and the rate of teacher turnover. Schools serving poor children traditionally have more difficulty hiring and retaining highly-qualified teachers than those in more affluent areas (Hanushek and Rivkin, 2007, Jackson 2009).[24] If the bonus program increased the supply of qualified teachers willing to work at treatment schools or reduced turnover, it could have effects on student achievement in the long run. However, consistent with our other findings, , the bonus program did not reduce either type of turnover (Table 6 Panel A) or the characteristics of new hires, as measured by the percent with a masters degree and the percent with prior teaching experience (Table 6, Panel B).

7 Conclusion

The conditional random assignment of eligibility for the school-wide bonus program in New York City offers a great opportunity to learn about the impact of group-based incentives on schools, teachers, and students. Interestingly, we find little evidence that the bonus program had any effect on student test scores in either the first or second years of the program, nor did it lead to significant behavioral changes, as measured by student and teacher surveys, teacher absences, teacher turnover, or the selection of new teachers. While we find suggestive evidence of a small reduction in teacher absences in schools where teachers had relatively strong incentives to increase their effort, we also find some evidence that students in the lowest portion of the test score distribution were negatively impacted by the program.

In general, prior studies have found that teacher incentive pay enhances student achievement and other desirable outcomes. However, our results underscore the fact that the structure of performance pay and the setting in which it is implemented may be very important in determining its effects. On one hand, we find some evidence that a group-based compensation scheme led to free-riding among teachers with no discernable short-run benefits for students. On the other, in schools where the incentives for individual teachers are strongest, there is evidence that teachers do increase effort by reducing absenteeism. It is important to note that the size of the bonuses offered in the New York City program were not small – constituting around five percent of teacher salaries – compared with programs studied elsewhere. For example, in the

[24] In a given year, approximately 10 percent of NYC teachers leave the city while an additional 8 percent switch schools within the city.

program that Lavy (2002) examines, which offered group incentives to teachers within schools, bonuses were also equivalent to five percent of a teacher's starting salary. Thus, it is unlikely that the lack of any significant impacts is due to the size of potential bonuses. Indeed, the city spent $14 million providing bonuses after the first year of the program, and, as student test scores in New York City rose in 2009, it will likely spend an equal or greater amount in the program's second year. The results of this paper suggest that students may benefit from a restructuring of the program, or another alternative use of these funds.

References:

Ballou, D. (2001) "Pay for Performance in Public and Private Schools," Economics of Education Review 20(1): 51-61.

Clotfelter, C., Ladd, H., and Vigdor, J. (2009) "Are Teacher Absences Worth Worrying About in the U.S.?" Education Finance and Policy, 4(2): 115-149.

Clotfelter, C., Ladd, H. and Vigdor, J. (2006) "Teacher-Student Matching and the Assessment of Teacher Effectiveness," National Bureau of Economic Research working paper #11936.

Cullen, J. B. and Reback, R. (2006) "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System," In T. Gronberg and D. Jansen (Eds), Advances in Applied Microeconomics, 14.

Decker, P., Mayer, D., and Glazerman, S. (2004). "The Effects of Teach for America on Students: Findings from a National Evaluation," Mathematica Policy Research Report, New York.

Figlio, D. and Getzler, L. (2006) "Accountability, Ability, and Disability: Gaming the System?" In T. Gronberg and D. Jansen (Eds), Advances in Applied Microeconomics, 14.

Figlio, D. and Kenny, L., (2007) "Individual teacher incentives and student performance," Journal of Public Economics, 91: 901-914.

Figlio, D. (2006) "Testing, Crime, and Punishment," Journal of Public Economics 90: 837-851.

Glewwe, P., Ilias, N., and Kremer, M. (2003) "Teacher Incentives," NBER working paper #9671.

Hanushek, E. (2006) "School Resources" in E. Hanushek and F. Welch (Eds), Handbook of the Economics of Education.

Hanushek, E. and Rivkin, S. (2007) "Pay, Working Conditions, and Teacher Quality," The Future of Children, 17(1): 69-86.

Herrmann, M., and Rockoff, J. (2009) "Work Disruption, Worker Health and Productivity: Evidence from Teaching," Columbia Business School.

Holmstrom, B. and Milgrom, P. (1991) "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," Journal of Law, Economics, and Organization, 7, Special Issue: Papers from the Conference on the New Science of Organization: 24-52.

Jackson, C. K. (2009) "Student Demographics, Teacher Sorting, and Teacher Quality: Evidence from the End of School Desegregation," Journal of Labor Economics 27(2): 213-56.

Jackson, C. K. and Bruegmann, E. (2009) "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," American Economic Journal: Applied Economics 1(4).

Jacob, B. (2005). "Accountability, Incentives and Behavior: Evidence from School Reform in Chicago," Journal of Public Economics, 89: 761-796.

Jacob, B. and Levitt, S. (2003) "Rotten Apples: An Investigation Of The Prevalence And Predictors Of Teacher Cheating," The Quarterly Journal of Economics, 118(3): 843-877.

Kane, T., Rockoff, J., and Staiger, D. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City," Economics of Education Review, 27, no. 6 (2008): 615-631.

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

Lavy, V. (2004) "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," NBER working paper #10622.

Lazear, E. (1986) "Salaries and Piece Rates," Journal of Business, 59(3): 405-31.

Miller, R., Murnane R., and Willett J. (2008) "Do Worker Absences Affect Productivity? The Case of Teachers," International Labour Review, 147(1): 71–89.

Muralidharan, K., and Sundararaman, V. (2008) "Teacher Incentives in Developing Countries: Experimental Evidence from India," 2008 Conference on Performance Incentives, Nashville, TN, National Center on Performance Incentives.

MacLeod, W. B. and Parent, D. (1999) "Job Characteristics and the Form of Compensation," Research in Labor Economics, Vol. 18, JAI Press: 177-242.

Neal, D. (2008) "Designing Incentive Systems for Schools," forthcoming in *Performance Incentives: Their Growing Impact on American K-12 Education*, ed. by Matthew Springer, Brookings Institution Press, 2008.

Reback, R. (2008) "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," Journal of Public Economics, 92: 1394–1415.

Rockoff, J., and Turner, L. J. (2008) "Short Run Impacts of Accountability on School Quality," NBER working paper #14564.

Table 1: Baseline School Characteristics by Treatment Status

	Treatment Schools		Control Schools	Non-Experimental Schools
Number of Schools	185		162	640
Average enrollment	560		548	690
Average enrollment, tested grades	359		364	450
Fraction elementary school	62%	*	51%	60%
Fraction middle school	26%	*	38%	28%
Fraction K-8 school	12%		11%	12%
School Accountability Outcomes				
Peer index (mean = 0, sd = 1)	-0.90		-0.93	0.50
Overall accountability score	52.8		51.6	54.5
Student Characteristics				
Average math scale score (2007)	656	**	651	677
Change in math scale score (2006 to 2007)	10.6		10.3	8.8
Average reading scale score (2007)	640	+	638	661
Change in reading scale score (2006 to 2007)	1.5	*	3.0	3.1
Fraction English Language Learner	19%		19%	10%
Fraction special education	12%		13%	9%
Fraction free lunch	88%		89%	62%
Fraction Hispanic	56%		53%	32%
Fraction Black	41%		43%	30%
Fraction White	1%		1%	20%
Teacher Characteristics				
Number of teachers	55		54	59
Number of teachers, tested classrooms	12		13	14
Average years of experience	7.8		7.8	8.4
Fraction with masters degree	50%	**	47%	47%
Average absences (2007)	4.1	*	3.8	3.7
Average absences, tested classrooms (2007)	4.2		4.1	3.8
Fraction teachers not retained by DOE (2007)	12%		12%	9%
Fraction teachers changing schools (2007)	7%		8%	5%
Fraction of new teachers TFA volunteer	12%		13%	3%
Fraction of new teachers Teaching Fellow	26%		24%	19%
Fraction of new teachers with MA	35%		36%	41%
Fraction of new teachers with prior experience	27%		30%	36%

Notes: Characteristics measured at beginning of 2007-2008 school year unless otherwise noted; + difference between treatment and control significant at 10%, * 5%, ** 1%; average absences measured between November 2006 and March 2007.

Table 2: Impact of Teacher Incentives on Student Academic Achievement

	2007-2008			2008-2009		
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable						
Math Scale Score	-0.644	-0.710	-0.432	-0.650	-0.665	-0.751
	(0.530)	(0.519)	(0.520)	(0.478)	(0.468)	(0.486)
Reading Scale Score	-0.158	-0.284	-0.270	-0.129	-0.179	-0.232
	(0.474)	(0.446)	(0.458)	(0.332)	(0.327)	(0.335)
Observations	347	347	330	343	343	326
Additional controls		X	X		X	X
Outliers dropped			X			X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each cell denotes a separate regression; all regressions control for prior (2007) scale score; additional controls include: school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic), columns (3) and (6) drop schools in bottom 5% of peer index, all regressions weighted by number of students tested.

Table 3: Impact of Teacher Incentives on Student and Teacher Survey Outcomes

	(1)	(2)	(3)
A. Student Survey Outcomes			
Essays and Projects	-0.125	-0.122	-0.185
	(0.123)	(0.123)	(0.119)
Group & Hands-on	-0.076	-0.069	-0.047
Learning Activities	(0.142)	(0.149)	(0.147)
Tutoring Offered	0.061	0.078	0.092
Before/After School	(0.133)	(0.139)	(0.139)
Observations	159	159	143
B. Teacher Survey Outcomes			
Use of Student Data	-0.034	-0.053	-0.043
	(0.098)	(0.099)	(0.098)
Quality of Professional	0.076	0.069	0.081
Development	(0.098)	(0.099)	(0.099)
High Expectations	-0.106	-0.132	-0.126
	(0.092)	(0.086)	(0.085)
Observations	347	347	330
Additional controls		X	X
Outliers dropped			X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each cell denotes a separate regression; all regressions control for prior (2007) survey outcome; additional controls include: school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic), column (3) drops schools in bottom 5% of peer index, all regressions weighted by number of survey respondents.

Table 4: Impact of Teacher Incentives on Absences taken for Personal and Sick Leave, November 2007 through March 2008

	(1)	(2)	(3)	(4)	(5)	(6)
A. All Teachers						
Treatment	0.067	0.069	0.058			
	(0.085)	(0.088)	(0.089)			
Observations	347	347	330			
B. Teachers with Tested Students						
Treatment	0.003	0.009	0.023	-0.127	-0.126	-0.138
	(0.140)	(0.143)	(0.146)	(0.135)	(0.137)	(0.138)
Number of teachers with				-0.007	-0.011	-0.008
tested students *(mean = 0)*				(0.012)	(0.013)	(0.015)
* Treatment				0.054	0.053	0.053
				(0.023)*	(0.023)*	(0.025)*
Observations	320	320	305	320	320	292
Additional controls		X	X		X	X
Outliers dropped			X			X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column within Panels A and B denotes a separate regression; in columns (4) through (6) the number of teachers with tested students is demeaned; all regressions control for prior absences; additional controls include: school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic), columns (3) and (6) drop schools in bottom 5% of peer index; regressions are unweighted; schools with no teachers linked to tested students are dropped in Panel B regressions.

Table 5: Heterogeneity in the Impact of Teacher Incentives on Student Academic Achievement by Prior Year Achievement and ELL/Special Education Status

	Math Scale Score			Reading Scale Score		
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment * ELL or Special Education	-0.218	-0.229	0.281	-0.595	-0.511	-0.434
	(0.861)	(0.828)	(0.829)	(0.796)	(0.766)	(0.789)
Treatment * Lowest 3rd	-1.409	-1.330	-1.096	-0.432	-0.370	-0.246
	(0.768)+	(0.751)+	(0.760)	(0.645)	(0.614)	(0.630)
Treatment * Middle 3rd	-0.810	-0.796	-0.501	-0.128	-0.182	-0.082
	(0.642)	(0.630)	(0.636)	(0.515)	(0.481)	(0.495)
Treatment * Highest 3rd	-0.943	-0.937	-0.584	-0.451	-0.467	-0.469
	(0.720)	(0.703)	(0.715)	(0.650)	(0.572)	(0.588)
Observations	1372	1372	1308	1368	1368	1308
Additional Controls		X	X		X	X
Outliers Dropped			X			X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each column denotes a separate regression; all regressions control for prior year (2007) scale score; additional controls include: school level, peer index, overall accountability score, percentage of students receiving free lunch, and student race (African American and Hispanic); columns (3) and (6) drop schools in bottom 5% of peer index , all regressions weighted by group size.

Table 6: The Impact of Teacher Incentives on Teacher Turnover and the Qualifications of New Teachers

	(1)	(2)	(3)
A. Teacher Turnover, 2008-2009			
Fraction of teachers not retained by school district	-0.005	-0.002	-0.003
	(0.006)	(0.006)	(0.006)
Fraction of teachers leaving for another NYC school	-0.002	0.003	0.002
	(0.006)	(0.006)	(0.006)
Observations	343	343	326
B. Characteristics of New Teachers, 2009			
Fraction of new teachers who are Teach for America volunteers	0.000	0.005	0.010
	(0.028)	(0.027)	(0.027)
Fraction of new teachers with MA	0.039	0.033	0.033
	(0.027)	(0.027)	(0.027)
Fraction of new teachers with prior teaching experience	-0.004	-0.006	-0.007
	(0.025)	(0.024)	(0.024)
Observations	292	292	279
Additional controls		X	X
Outliers dropped			X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each column with Panels A and B denotes a separate regression; Panel A regressions control for prior (2007-2008) fraction of teachers not retained or fraction of teachers leaving for another school; Panel B regressions control for prior (2008) outcome; additional controls include: school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic), column (3) drops schools in bottom 5% of peer index, all regressions weighted by number of teacher (panel A) or number of new teachers (panel B); schools without new teacher hires dropped from Panel B regressions.

Figure 1: Distribution of Math Scale Scores by Year and Treatment Status

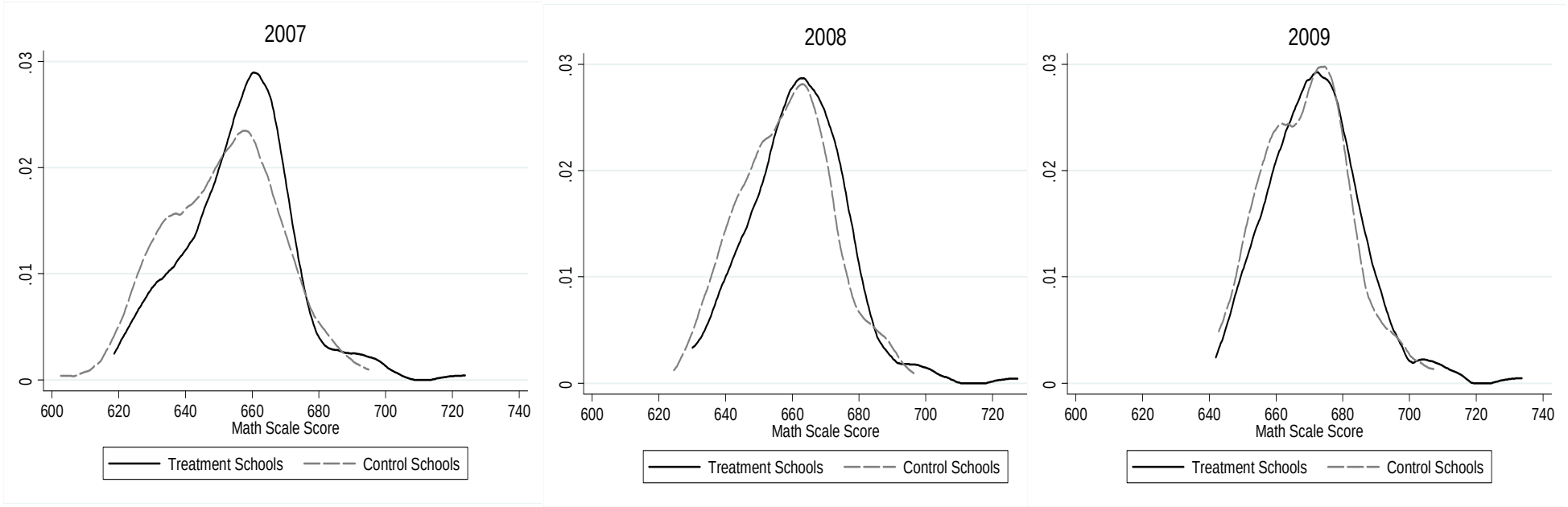

Figure 2: Distribution of Reading Scale Scores by Year and Treatment Status

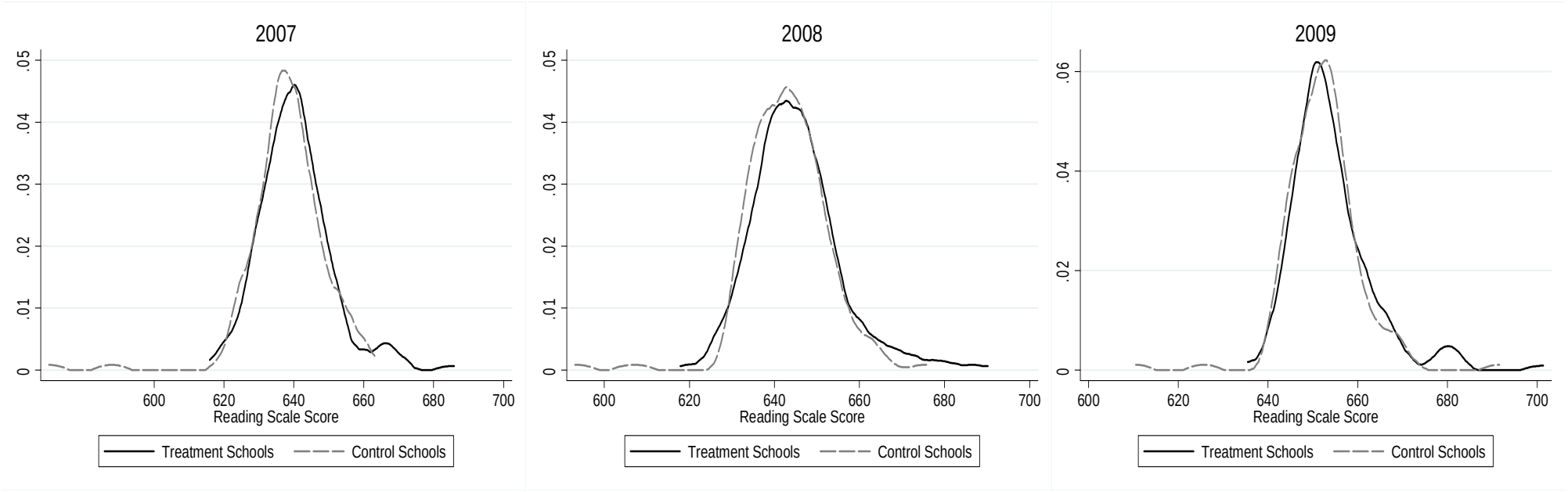

Table A1: Baseline Characteristics of Treatment Schools by Vote

	Voted "No"		Voted "Yes"
Number of Schools	25		160
Average enrollment	573		558
Average enrollment, tested grades	333		377
Fraction elementary school	0.72		0.60
Fraction middle school	0.08		0.13
Fraction K-8 school	0.20		0.28
School Accountability Outcomes			
Peer index (mean = 0, sd = 1)	-0.86		-0.91
Overall accountability score	56.2		52.3
Student Characteristics			
Average math scale score (2007)	662	*	655
Change in math scale score (2006 to 2007)	10.2		10.6
Average reading scale score (2007)	645	*	640
Change in reading scale score (2006 to 2007)	53%		170%
Fraction English Language Learner	18%		20%
Fraction special education	12%		12%
Fraction free lunch	88%		88%
Fraction Hispanic	57%		56%
Fraction Black	38%		41%
Fraction White	1%		1%
Teacher Characteristics			
Number of Teachers	55		54
Number of teachers, tested classrooms	13		12
Average years of experience	8.2		7.7
Fraction with masters degree	50%		49%
Average absences (2007)	3.9		4.1
Average absences, tested classrooms (2007)	4.3		4.2
Fraction teachers not retained by DOE (2007)	11%		12%
Fraction teachers changing schools (2007)	7%		7%
Fraction of new teachers TFA volunteer	8%		12%
Fraction of new teachers Teaching Fellow	20%		27%
Fraction of new teachers with MA	41%		34%
Fraction of new teachers with prior experience	38%	*	25%

Notes: Characteristics measured at beginning of 2007-2008 school year unless otherwise noted; + difference between treatment and control significant at 10%, * 5%, ** 1%; average absences measured between November 2006 and March 2007.

Table A2: Program Impacts on Percentage of Tested Students Classified as English Language Learner or Special Education, by Subject

	(1)	(2)	(3)
Math			
Percentage of tested students ELL	-0.003	0.001	0.001
	(0.003)	(0.001)	(0.001)
Observations	290	290	276
Percentage of tested students special education	-0.003	0.000	0.000
	(0.003)	(0.003)	(0.003)
Observations	322	322	308
Reading			
Percentage of tested students ELL	-0.001	0.002	0.003
	(0.003)	(0.002)	(0.002)+
Observations	284	284	270
Percentage of tested students special education	-0.002	0.001	0.002
	(0.003)	(0.003)	(0.003)
Observations	322	322	308
Additional controls		X	X
Outliers excluded			X

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each cell denotes a separate regression; all regressions control for prior (2007) percentage ELL or special education; additional controls include: school level, peer index, overall accountability score, percentage of students ELL , special education , free lunch recipients, and student race (African American and Hispanic), columns (3) and (6) drop schools in bottom 5% of peer index, all regressions weighted by number of students tested.