# Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program

Sarena F. Goodman
Columbia University

Lesley J. Turner
Columbia University

November 2010*

Abstract

Teacher compensation schemes are criticized for lacking a performance-based component. Proponents argue that teacher incentive pay can raise student achievement and stimulate system-wide innovation. We use a policy experiment conducted in the New York City public school system to explore the effects of a performance-based, school-wide bonus scheme on student achievement, teacher absenteeism, classroom activities, and teacher quality. Teacher incentive pay had little effect on these outcomes. We provide evidence that the group bonuses led to free-riding and show that in schools where incentives to free-ride were weakest, the program led small increases in math achievement.

* 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 Todd Kumler, Bentley MacLeod, Ben Marx, Petra Persson, Maya Rossin, Jesse Rothstein, Miguel Urquiola, Till Von Wachter, and Reed Walker and seminar participants at the Columbia applied microeconomics colloquium, AEFA annual meeting, Teacher's College Economics of Education Workshop, and the Harvard Kennedy School's Program on Education Policy and Governance's Merit Pay Conference for helpful feedback. We are grateful to the New York City Department of Education for the data used in this paper.

## 1. Introduction

Teacher compensation schemes are often criticized for their lack of performance pay. A large body of empirical research shows that in many sectors, incentive pay increases worker effort and output.[1] Properly structured pay schemes align the interests of workers and employers, provide information about the most valued aspects of an employee's job, and motivate workers to provide costly effort. If in at least some schools, teachers exert an inefficiently low amount of effort or focus their effort on tasks with low marginal returns, teacher incentive pay may increase student achievement. Additionally, in the long-run, a performance-based element of teacher pay may combat wage compression in the profession and increase the ability of individuals choosing to enter the teaching profession (Lazear, 2003; Hoxby and Leigh, 2005). Public school systems rarely use performance pay schemes for teachers, especially in comparison to their private school counterparts (Ballou, 2001; Ballou and Podgursky, 1997).

However, several features of the educational sector may dilute the effect of performance pay. First, incentive pay is most effective when employers have good measures of worker output or observable effort is closely tied to firm productivity. It is costly to monitor teachers and difficult to quantify individual contributions to a student's education since production depends not only on a student's current teacher but also upon the effort provided by past teachers. Education is a complex good; educators must complete multidimensional tasks and allocate their effort across several activities. Tying incentives to a single measure, such as student test scores, may lead teachers to focus their effort away from classroom activities that are also important for student learning (Holmstrom and Milgrom, 1991), focus on narrowly-defined basic skills that appear on exams (e.g., "teaching to the test"), or overtly manipulate test scores (e.g., Jacob and Levitt, 2003; Jacob, 2005; Figlio, 2006; Figlio and Getzler, 2006; Cullen and Reback, 2006). Finally, to the extent that current accountability systems, such as No Child Left Behind (NCLB), already provide significant negative incentives for teachers to improve test scores, it is unclear whether reasonably sized monetary incentives can induce additional effort provision.

In this paper, we investigate the impact of group-based teacher incentive pay, taking advantage of a policy experiment conducted in New York City. In the fall of 2007, 181 schools

[1] These compensation schemes are generally most effective in sales jobs and those that involve operating machines. Macleod and Parent (1999) provide an overview of other sectors that employ incentive-based pay schemes. Gibbons (1998) and Lazear and Oyer (2010) review the performance pay literature.

were randomly selected from a group of high-poverty schools.[2] These schools were eligible to earn school-wide bonuses if they achieved goals based primarily on student achievement on state math and reading exams. Schools that reached a set threshold received lump sum payments equal to $3000 per union teacher (between three and seven percent of annual teacher pay).

The best evidence on the effectiveness and optimal form of teacher merit pay comes from outside the United States. Experimental evidence from India (Muralidharan and Sundararaman, 2009) and quasi-experimental evidence from Israel (Lavy 2002, 2009) suggests both individual and group-based teacher incentive pay lead to increases in teacher effort and student achievement, although individual bonuses are the most effective.[3] Tournaments, where a certain percentage of top performers are rewarded, may be optimal if all schools or teachers are exposed to aggregate shocks (Lazear and Rosen, 1981). The tournaments Lavy (2002, 2009) examines both lead to positive outcomes. However, evidence from a tournament-based incentive pay program in Chile suggests that only a subset of schools experienced positive achievement gains (Rau and Contreras, 2009). Muralidharan and Sundararaman's (2009) treatments utilize a piece-rate payment scheme: teachers or schools receive bonus payments for incremental improvements in student achievement. Most other incentive schemes, including the NYC program we examine, instead provide bonus payments above an absolute threshold, which may dilute incentives for schools with a probability of bonus receipt that approaches either zero or one.

At least one study suggests that the rewarding test score gains may lead teachers to focus on test preparation activities, with little impact on long term achievement. Glewwe, Ilias, and Kremer (2010) study a school-based teacher incentive experiment in rural Kenya where non-monetary prizes were awarded based on both absolute and relative performance goals. The

[2] The program also included 39 secondary schools. Since bonus receipt for high schools was based on different outcomes for high schools, we focus on elementary and middle schools and schools serving children in kindergarten through 8th grade (K-8 schools). We exclude schools that served both K-8 students and high school students and schools in a special district that serve only special education students.

[3] Muralidharan and Sundararaman (2009) test the impacts of individual and group-based rewards using a randomized experiment in rural India and find positive returns to both types of incentives, but larger returns to individual incentives in the second year of the program. Lavy (2002) shows that school-wide incentives increased student test scores and participation on matriculation exams in Israel; the percentage of students who received matriculation certificates was not affected. Lavy (2009) examines a program in which teachers were awarded cash prizes for their students' relative performance. Incentive payments led to an increase in both the proportion of students taking a high school exit exam and the performance among test-takers. These student achievement gains likely stemmed from an increase in after-school sessions, evidence of increased teacher effort in response to potential rewards. Ahn (2009) shows evidence of free-riding among teachers in a system involving group bonuses, although his results suggest individual incentives may actually lead to lower effort if schools contain both high and low ability teachers.

program increased test-taking tutorials in treatment schools and led to short-term test score gains in the subjects used for bonus determination. However, the authors find no evidence of long-term gains in human capital or spillovers on other subjects.

There is less evidence of the impacts of teacher incentive pay in the United States. Figlio and Kenny (2007) document a positive cross-sectional relationship between individual-based teacher performance pay and student achievement in the United States. The most effective systems appear to be those where awards were difficult to earn and only a small number of teachers received incentive payments. However, these results are confounded by the possibility that better schools might be more willing to adopt bonus pay, leaving the direction of causation unclear. Preliminary results from experiments in Chicago and Nashville suggest teacher incentives in the U.S. have little effect on student achievement (Glazerman and Seifullah, 2010 and Springer et al. 2010). Springer and Winters (2009) also examine the NYC bonus program and find no discernable impact on student achievement. Our paper goes beyond documenting the null impacts on student test scores and investigates what features of the NYC bonus program may have diluted the program's incentives. Given the large amount of funding federal initiatives link to performance pay (e.g., Race to the Top and the Teacher Incentive Fund), our paper provides important evidence on which designs are most likely to be effective.

We examine the effect of this incentive pay program on average student achievement in math and reading, measured by performance on statewide exams. We also investigate a wide range of other outcomes that likely contribute to human capital development but may not immediately manifest as higher test scores: teacher effort, measured by absenteeism, and reported classroom activities and school policies, from surveys of teachers and students. To determine whether the program increased relatively disadvantaged schools' ability to recruit or retain qualified teachers, we test whether eligibility to earn bonuses affected teacher turnover and the quality of newly hired teachers, measured by experience and other qualifications. The bonus program had little impact on any of these outcomes. If anything, the program resulted in a slight reduction in math achievement and the percentage of students classified as proficient in math in its second year.

We investigate which features of the bonus program may have led to its ineffectiveness. In theory, group incentive pay is most effective with a joint production technology (Itoh, 1991). If an individual teacher's effort has a positive effect on the effort chosen by other teachers (e.g.,

Jackson and Bruegmann, 2009), then group incentives are optimal. Otherwise, group incentives decrease individual returns to effort and may lead to free-riding unless workers monitor each other's effort. We test for free-riding by allowing the program's impacts to vary by the number of teachers with students who are tested (and therefore contribute to the probability that a school qualifies for the bonus award). To test for the importance of joint production and monitoring, we examine whether program impacts vary by the degree to which teachers report collaborating in lesson planning and instruction on a survey administered in the year prior to the program's implementation. We find evidence that the bonus program raised math achievement in schools with a small number of teachers with tested students, although these program impacts are small (approximately 0.08 student-level standard deviations) and insignificant in the second year of the program. We also find suggestive evidence of positive program impacts in schools where instruction involves a high degree of collaboration across teachers.

In the fall of 2007, NYC also implemented an accountability system that contained significant incentives for schools to improve student achievement. Thus, our results represent the impact of group-based teacher performance pay for schools already under accountability pressure. However, given that many states have implemented accountability systems and all school districts in the United States are subject to NCLB, this may be the most appropriate parameter to estimate. Additionally, we show that schools under the least amount of accountability pressure were similarly affected by the bonus program, suggesting that our results not driven by dilution.

The second section of our paper describes the bonus program and Section 3 provides an overview of the data. In Section 4, we outline our estimation framework and present empirical results. Section 5 concludes.

## 2. The New York City School-Wide Bonus Program

We use a policy experiment implemented by the New York City Department of Education (DOE) in the fall of 2007, the "School-Wide Performance Bonus Program" (hereafter, the bonus program).[4] Both the DOE and the United Federation of Teachers (UFT) endorsed the program as an innovative model for teacher performance pay. In November 2007, 181 schools serving kindergarten through eighth grade were randomly selected from a group of 309 schools

[4] The original randomization of the schools in the experimental sample was led and conducted by Roland Fryer.

designated as "high need"; 128 schools were assigned to the treatment group. Two of the 181 schools originally assigned to the treatment group were moved to the control group prior to notification of their assignment; for the purposes of our analyses, we consider these schools as part of the original treatment group. Treatment schools were eligible to participate in the program, contingent on 55 percent of a school's full-time United Federal of Teachers (UFT) staff voting in favor of participation. Twenty-five schools voted not to participate or withdrew from the program following a vote of approval. Finally, four schools originally assigned to the control group were allowed to vote and ultimately chose to participate in the bonus program; for the purposes of our analyses, we consider these schools as part of the original control group. However, the group of schools that ultimately could earn bonus payments totaled 158.

The schools that voted in favor of the program could earn a lump-sum bonus by meeting a school-wide goal. These goals were tied to the NYC accountability system which awarded letter grades to schools (explained below) and were primarily based on student achievement on state math and reading exams. Schools that achieved their goals received lump sum bonuses equal to $3,000 per union teacher, while schools that fell short but managed to meet 75 percent of their goal received $1,500 per union teacher. Thus, although total bonus awards varied across schools with different numbers of union teachers, the expected bonus payment was equal across these schools.[5] Schools that did not reach their target suffered no consequences beyond the absence of bonus pay. The full $3,000 award represents a seven percent increase in the salary of teachers at the bottom of the pay scale and a three percent increase for the most experienced teachers.[6]

Each participating school selected a four-member compensation committee, consisting of the principal, a second administrator, and two union representatives elected by the school's UFT members.[7] In the program's first year, this committee was required to submit a bonus distribution scheme after students took the state math and reading exams but before exam results were released. Thus, at least in the first year of the program, teachers' effort decisions should not be affected by the distribution that was ultimately chosen. Program guidelines stipulated that within

[5] Schools that received the highest accountability grade for two concurrent years also earned $1,500 per union teacher. However, this condition was announced in June 2008, after the state tests were taken (Springer and Winters, 2009).

[6] Similar to the vast majority of public school districts in the United States, New York City teacher salaries are determined through a schedule that only takes into account years of experience and graduate coursework (Podgursky and Springer, 2007). Schedules are available at http://www.uft.org/member/contracts/moa/salary_schedules

[7] See http://www.uft.org/member/rights/bonus/moa/.

schools reaching their goal, all union teachers must receive a bonus payment and individual bonuses could not be explicitly based on seniority. Beyond these requirements, committees had complete freedom in determining individual teachers' bonus payments and could also provide bonus payments to other school employees. Around half of treatment schools chose an approximately equal distribution (i.e., the difference between the highest and lowest bonus payment was less than $100). In the remainder of schools, the difference between the highest and lowest bonus ranged from $200 to $5000 (Figure 1).

The 2007-2008 school year also marked the implementation of the DOE's new accountability system. Under this system, schools received accountability grades designed to summarize a school's overall performance on a multidimensional metric of student learning.[8] Each school's performance was scored relative to the entire district and to a group of peer schools. This group included the 40 schools that were most similar according to a "peer index" that was based on student demographic characteristics and prior achievement.[9] Each school received a progress report documenting its overall performance, the corresponding accountability grade, and a target score for the following year. Schools with lower accountability grades needed to make larger improvements to reach their targets. Importantly, these target scores determined which schools participating in the bonus program received awards.

Moreover, the accountability system provided additional incentives to improve student achievement, regardless of bonus program participation. Schools that earned an A or B accountability grade received rewards (e.g., principal bonuses, additional funds when students transferred from schools receiving a poor grade), while schools that received D and F grades faced consequences (e.g., school closure and principal removal). Although this accountability system was more complex than systems based on a single metric (e.g., the percentage of students achieving proficiency), teachers and administrators received training on how to interpret the complicated set of measures determining a school's grade, and it was clear that grades were

[8] The metric includes a measure of school environment (student attendance and results from survey of parents, teachers, and students), student performance (average student achievement on reading and math exams, median proficiency, and percentage students achieving proficiency), and student progress (average change and percent making progress on math and reading exams). Schools received extra credit for student progress among high-need students.

[9] For elementary schools and those serving kindergarten through eighth grade (K-8), the peer 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 provide consistent metrics for relative disadvantage, as the components for the elementary/K-8 peer index are very strong predictors of 4th grade test scores.

largely determined by student performance on math and reading exams. Rockoff and Turner (2010) find that receiving an F or D led to a significant improvement in student test scores, a result consistent with school employees understanding that performance under the accountability system was dependent on student achievement. Bonus program impacts do not vary across schools with different accountability grades (see Section 4). However, it is still important to note that our results represent the impact of group-based teacher performance pay for schools already under accountability pressure.

The timing of program announcement and the selection of schools into the treatment group did not allow much room for behavioral responses to the program in its first year. The school vote took place in November 2007, less than two months before the January reading exam and less than four months before the March math exam.[10] The program continued into the 2008-2009 school year and all but three of the participating schools voted to continue participation.[11] Of the 158 schools that voted to participate in the first year of the program, 87 (55 percent) received bonus payments. The bonus pool averaged approximately $160,500 per school, and totaled $14.0 million in the first year. In the second year of the program, of the 151 schools that were eligible to receive bonuses, the vast majority (91 percent) earned awards, totaling $27.1 million.

## 3. Data and Descriptive Results

Our analyses focus on schools classified as elementary, middle, and K-8 (schools serving kindergarten through grade 8) eligible for selection into the bonus program.[12] A total of 181 schools were selected into the treatment group, while the control group includes 128 schools. The majority of our data are publicly available on the DOE website.[13] We measure academic achievement using the average math and reading test scores for each school for the 2006-2007, 2007-2008, and 2008-2009 school years (hereafter 2007, 2008, and 2009 school years). We also construct a measure of the share of students within each school classified as proficient in each

[10] However, even given this short time period, the NYC accountability system led to significant improvements in math and, albeit smaller, improvements in reading (Rockoff and Turner, 2010).
[11] Schools that voted no in the first year of the program were not given a second chance to vote on the program. However, we still consider these schools as part of the group originally assigned to the treatment group.
[12] A small number of schools initially belonging to the experimental sample were excluded *prior* to random assignment. The exclusion of these schools does not affect the internal validity of our results. We examine the characteristics of these schools to determine if the external validity of our results is compromised, and find little differences between these schools and the final experimental sample (results available upon request).
[13] See http://schools.nyc.gov/Accountability/data/default.htm for details.

subject. We take advantage of school-level results from annual teacher and student surveys conducted by the DOE as part of the accountability system.[14] Specifically, we use the questions from the student survey on the extent to which: 1) students completed essays and research projects and 2) classroom activities including group work, class discussions, and "hands-on activities such as science experiments." We also measure the availability of tutoring, using questions on whether sessions were offered before or after school. From the teacher survey, we use a question addressing whether teachers use student achievement data, such as students' test results from prior years or "periodic examinations" during the school year, to inform their lesson planning. We also create a measure of whether teachers believed students faced high standards and expectations.

We measure teacher absences, teacher turnover, and the characteristics of newly hired teachers using aggregate statistics from data on individual teachers.[15] In some specifications, we include information on school demographic characteristics (the percentage of students in each school that are English Language Learners (ELL), special education students, Title I free lunch recipients, and minorities) and each schools performance under the new NYC accountability system, including each school's accountability score and peer index.

*3.1 Was Randomization Successful?*

Our ability to make causal inferences about the effects of teacher incentive pay depends on the success of random assignment. In Table 1, we present comparisons of the characteristics of treatment and control groups prior to random assignment, where the treatment group includes schools that were initially selected but did not participate.[16] Treatment and control schools are similar in terms of enrollment, accountability outcomes, student demographics, and teacher characteristics. We find no significant differences between the observable characteristics of treatment and control schools, suggesting a causal interpretation of our results is valid.

We also compare the characteristics of the 309 schools in the experimental sample to other schools in NYC.[17] Given that schools with low peer indices were eligible for selection into

[14] Available at http://schools.nyc.gov/Accountability/tools/survey/default.htm.
[15] We thank Jonah Rockoff for constructing these aggregate statistics for the purpose of this research.
[16] Appendix Table A1 compares the characteristics of schools by whether or not they voted to participate in the program. Schools voting "no" are largely similar to schools that voted in favor of the program, although, on average, these 25 schools were relatively less disadvantaged and their students had higher test scores.
[17] We restrict our universe to the 923 schools serving students in kindergarten through eighth grade that received accountability grades and were not charter schools or schools that only serve special education students.

the bonus program, it is not surprising that the experimental sample differs from the remainder of NYC schools across a number of dimensions. Schools in the experimental sample had a higher proportion of English Language Learners (ELL), special education, 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 almost twice as many absences than teachers in other NYC schools. Finally, experimental schools had lower enrollment and fewer teachers than other schools.

## 4. Regression Framework and Results

We take advantage of this randomized experiment to estimate the effect of teacher incentives using the following simple model:

$$(1) \qquad Y_{jt} = \delta D_{jt} + \varepsilon_{jt},$$

where $Y_{jt}$ is the outcome of interest for school $j$ in year $t$ (for example, average math scores in 2008), $D_{jt}$ is an indicator for whether a school is *eligible* for the bonus program (regardless of whether the school ultimately participated), and $\varepsilon_{jt}$ is a stochastic error component. These "intent-to-treat" estimates tell us the impact of offering a school the opportunity to participate in the bonus program. We estimate the equation with ordinary least squares, where school observations are weighted by the 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).[18] With successful randomization, $D_{jt}$ is independent of omitted variables and this approach should estimate the true effect of the bonus program. The identifying assumption requires that there be no contemporaneous shock that affects the relative outcomes of the treatment schools in the same period as the treatment. Such a shock would be highly unlikely in our setting given the experimental framework. We estimate a second specification that includes a vector of control variables, including the outcome measured in 2007, the year prior to the intervention, to reduce residual variance. Additional controls include indicators for 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), and 2007 peer index and accountability score. In a third specification, we instrument for actual participation in the bonus program with a

[18] If bonus program treatment effects are homogenous across students and teachers, estimates from weighted regressions will be efficient. Results obtained from unweighted regressions are similar (available upon request).

school's original assignment using two-stage least squares. These "treatment-on-the-treated" estimates can be interpreted as the impact of the program on schools that choose to participate.

*4.1 Student Math and Reading Achievement*

To preview our estimates of the impact of the bonus program on student achievement, Figures 2 and 3 display the distribution of average math and reading scores within treatment and control schools in 2007, 2008, and 2009. On average, all NYC schools experienced an increase in average student performance in the two years following the implementation of the program; this pattern holds in the experimental sample. If the bonus program had an impact on test scores, we should observe a rightward shift in the distribution among treatment schools, relative to control schools. The distribution of math and reading scores do not differ significantly between treatment and control schools in either 2008 or 2009.

Table 2, which displays results from regressions estimating the impact of the program on average math and reading exam scores, confirms these findings. We find little evidence that the program led to increases in math and reading achievement and, if anything, it appears that eligibility to earn bonuses had a negative impact on math achievement. Panels A and B examine the first and second years of the program separately. The point estimates for 2008 are negative and quite small, although precisely estimated.[19] In the second year of the program, eligibility to earn bonuses had no effect on student achievement in reading and a small negative impact on math scores, leading to an approximately 0.08 standard deviation reduction in math achievement.[20]

Next we examine whether the bonus program had any effect on the proportion of students achieving proficiency (Table 3).[21] A measure of central tendency, such as mean test scores, may not capture distributional effects. Teachers could have focused on improving achievement among particular student subgroups or altered their teaching practices in a way that had differential effects for students along the achievement distribution. We do find that the bonus program led to a significant decrease in student proficiency; however, the magnitude of this effect is small –

[19] For instance, our IV estimates reject effects as small as a 0.7 point increase in reading achievement and a 0.2 point increase in math. These effects are quite small in magnitude, given the 2008 student level standard deviation in test scores was 35 points for reading and 31 points for math.

[20] Four schools in the treatment group were closed at the end of the 2008 school year, thus, our sample decreases by four in the second set of regressions. Our 2008 results remain unchanged when we restrict the sample to only include schools open in both 2008 and 2009.

[21] Students are considered proficient if they achieve a set score on the state exams and are regarded as meeting learning standards.

approximately a 3 percent reduction in proficiency for both subjects. This decrease in proficiency rates appears to be caused by an overall reduction in student achievement rather than a differential change in achievement for high achieving students (discussed further in Section 4.3).

*4.2 Group Bonuses and the Free-Rider Problem*

Theory suggests that teachers will respond to the bonus program by increasing effort until the expected marginal benefit is equal to the marginal cost. However, the probability that a treatment school reaches its goal and receives a bonus award depends primarily on student math and reading performance. Thus, the impact of an individual's teacher's effort on her expected bonus is decreasing as the number of teachers with tested students grows large.[22] In other words, the diffusion of responsibility for test score gains across many teachers may dilute the incentives of the bonus scheme. Moreover, monitoring may be more difficult in schools with more teachers, amplifying free-riding incentives (Holmstrom, 1982).

We test for evidence of free-riding by allowing treatment effects on math and reading scores to vary by the number of math and reading teachers, respectively. A small number of middle and K-8 schools do not have information on the number of teachers teaching tested subjects; these schools are excluded. The first set of regressions in Table 4 show that our basic results hold for this sample of schools. We then add an interaction between the number of math/reading teachers (relative to the mean number of such teachers in the sample) and the treatment indicator (columns 2 and 5), and finally, interact treatment status with an indicator for schools in the bottom quartile of the number of teachers with tested students (approximately 10 or fewer teachers in elementary and K-8 schools and 5 or fewer in middle schools). We only present results from specifications that include covariates, however, results are similar when we exclude covariates or instrument for actual treatment with initial assignment.

We find evidence of free-riding. For schools at the bottom of the distribution of teachers, we estimate a positive effect of the bonus program on math achievement in the first year of the program and a positive, but insignificant effect in the second year, although we cannot reject a

[22] Consider two extremes, a school with only one teacher with tested students and a school with an infinite number of such teachers. In the first case, the teacher will either respond to the program choose to increase her effort to the expected level necessary to achieve the school's goal or not respond (if the size of the bonus is less than the cost of exerting this level of effort). In the second case, each individual teacher has no ability to determine whether the school receives a payment and will optimally not respond.

test of equality of treatment effects across years. In 2008, the bonus program resulted in a 3.2 point (0.08 student standard deviation) increase in math achievement.

Group-based incentive pay can be more effective than individual-based performance pay when production is joint. If the degree to which teachers work together varies across schools, group bonus payments may be effective in schools with a high level of cooperation between teachers. We construct a measure of school cohesiveness using teachers' answers to a set of five survey questions in spring 2007 – prior to the announcement of the bonus program – to proxy for the extent of joint production in a school.[23] This measure may also pick up on the degree to which teachers are able to monitor their colleagues. We sum responses across survey questions and standardize the index so it has a mean of zero and standard deviation equal to one. Schools with a teacher survey response rate lower than 10 percent are excluded. This index has a small, negative, and statistically insignificant correlation with the number of math and reading teachers in a school (i.e., schools with high levels of cohesion are distinct from those with a small number of teachers with tested students). Table 5 tests for heterogeneity in treatment effects according to the level of school cohesion. We first interact treatment with the linear index (columns 2 and 5) and then interact treatment with an indicator for above average cohesion (columns 3 and 6). Results provide suggestive evidence that the program may have had detrimental effects in schools with low levels of cohesion, and small positive effects on achievement in cohesive schools. The point estimates for schools with below average cohesion are (at least marginally) statistically significant and negative in both subjects and both years. The interaction of treatment and the indicator for above average cohesion is significant, positive, and of greater magnitude.[24]

*4.3 Bonuses and School Accountability*

Although the bonus program had little overall effect on student achievement, tying bonuses to the structure of the NYC accountability system provided incentives for schools to focus on students at different points in the achievement distribution. In line with recent research examining the effect of accountability systems on the performance of different groups of students (Neal and Schanzenbach, 2010; Cullen and Reback, 2006; Figlio and Getzler, 2006;

[23] These questions include: (1) the extent to which teachers report feeling supported by fellow teachers, (2) whether curriculum and instruction is aligned within and across school grades, (3) whether the principal involves teachers in decision making, (4) whether school leaders encourage collaboration, and (5) whether teachers collaborate to improve instruction.

[24] As a placebo test, we replicate Table 5 using pre-treatment (2007) math and reading achievement and find no significant treatment or interaction effects (results available upon request).

Figlio, 2006), we test whether the bonus program had heterogeneous impacts for students whose scores were weighted differentially in bonus determination. While the NYC accountability system takes into account average school performance and changes in performance for individual students, students with certain characteristics may be double or even triple weighted: those whose prior-year achievement placed them in the lowest third of their grade, those on the cusp of proficiency and those close to the school median, and those designated as ELL and special education.[25] We divide students into terciles based on their prior-year achievement within the school, treating ELL and special education students as a fourth, mutually exclusive group. We find no difference in the impact of the bonus program across these different groups of students (Appendix Table A2).

Schools could also respond to the bonus program by removing students from the test-taking pool or reclassifying higher performing students as either ELL or special education to take advantage of the increased weight placed on these students' achievement. We find no impact on the overall proportion of students taking math and reading exams or the proportion tested students classified as ELL or special education (Appendix Table A3).

An additional concern is that teachers had already adjusted their effort or teaching practices in response to the NYC accountability system's incentives. If teachers face decreasing marginal returns or increasing marginal costs to effort, the size of potential bonus payments may not be large enough to induce additional effort. To evaluate this possibility, we take advantage of the fact that treatment schools face different incentives according to their accountability grades. Both treatment and control schools receiving low grades had additional motivation to improve student test scores, as they faced school closure or principal removal if student achievement did not improve in the following year. Conversely, schools receiving an A on their progress report generally needed to make the smallest gains to receive a bonus, thus, the program may not have provided a large incentive to teachers in treatment schools to alter their behavior. Treatment and control schools in the middle of the grade distribution faced the largest difference in incentives. We test whether treatment effects vary along this dimension, grouping schools into three separate

[25] Unlike accountability systems that depend on the number of students within a given group reaching an absolute threshold of proficiency (e.g., NCLB), the NYC accountability system awards points based on a school's performance *relative* to both the entire district and a to group of peer schools. Thus, determining exactly how much a particular student's achievement contributes to the probability a treatment school receives performance is quite difficult. The categories we group students into are, at best, blunt measures of how much these students contribute to the probability of bonus receipt.

bins by their accountability grades: A, B or C, and D or F. We find no significant differences in treatment effects between these grade groupings or for schools at the center of the grade distribution where the difference in incentives between treatment and control schools are largest (Appendix Table A4).

*4.4 Bonus Receipt and Year 2 Impacts*

As previously mentioned, treatment group schools were notified of their eligibility in November of 2007, leaving teachers with little time to respond to incentives, especially when preparing for reading exams. Although we find the program had little overall impact in the first or second year, bonus receipt (or lack of receipt) may have incentivized teachers to alter their behavior in the second year of the program. Under the assumption that the program had no effect in its first year, we test whether the receipt of a bonus had an impact on student achievement in the program's second year. We simulate bonus receipt in the control group and interact treatment with predicted bonus receipt. We find no evidence that bonus receipt led to any changes in student achievement in 2009 (Appendix Table A5).

*4.5 Teacher Effort*

A primary motivation for the use of performance-based pay is to provide teachers with incentives to increase effort devoted to raising student achievement. Although we do not directly observe teacher effort, we can measure teacher attendance, which may be correlated with effort decisions. Absences are more common among teachers than in other sectors and absenteeism has been shown to have a negative effect on student achievement (Clotfelter, Ladd, and Vigdor, 2009; Miller, Murnane, and Willett, 2008). Using data on absences among NYC teachers, Herrmann and Rockoff (2010) estimate that an additional 10 absences reduce test scores by 0.01 standard deviations.

We run a series of regressions where the dependent variable is the average number of absences taken during the months when schools first learned of their eligibility for the bonus program and when the last exams were taken (November 2007 and March 2008 in the first year and September 2008 and March 2009 in the second year of the program).[26] If teachers believe that their attendance can affect the probability of bonus receipt through increasing student achievement, changes in behavior should be largest over this period. We only consider absences

[26] Results are robust to alternate definitions of the time period (e.g., November to March in the second year or September to March in the first year).

that teachers are likely to have some control over – those taken for illness and personal business – and exclude days missed due to death in the family, injury, jury duty, absences required by the school system (e.g., professional development activities), conference attendance, and religious holidays.

Table 6 presents these results. Each column within a panel contains the estimates from separate regressions of the effect of the bonus program on the number of absences per teacher. The first column examines program impacts on absences across all teachers within a school. The bonus program had no measurable impact on school-wide absences. Column 2 focuses on teachers with tested students, while the third and fourth columns follow the same approach as Table 4 and interact the treatment indicator with the number of teachers with tested students (column 3) or an indicator for whether a school falls in the bottom quartile of the number of such teachers (column 4). Program impacts are inconsistent across years. We find some evidence of free-riding in the first year of the program: teachers in schools with a small number of teachers with tested students increased attendance, although impacts are only significant for schools at the $10^{th}$ percentile in the distribution of number of teachers (available upon request). In the second year of the program, we find positive but insignificant impacts on absenteeism and no evidence of free-riding.[27]

*4.6 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. Additionally, incentives to focus on student achievement may lead teachers to substitute away from other classroom activities (Holmstrom and Milgrom, 1991). We explore whether the bonus program led to changes in teacher behavior and school policies using results from the DOE's annual surveys of teachers and students. [28] 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

[27] We also test whether the bonus program had heterogeneous impacts according to initial teacher effort. For instance, initially low effort (high absence) teachers may be the only group with room to respond through increasing attendance. However, we find no evidence that this is the case (results available upon request).

[28] For ease of interpreting results, all survey outcomes are standardized to have a mean of 0 and standard deviation of 1 across all NYC schools, according to school type.

opportunities for before- or after-school tutoring. Only students in grades six or higher completed the environmental survey, thus, we lose a number of schools, primarily elementary schools.

We do not find significant treatment effects 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 positively correlated with treatment and in the third specification, the latter measure comes close to conventional significance levels (Table 7, Panel A). Additionally, the bonus program had no significant impact on tutoring. Estimates are similar when we do not include covariates or instrument for actual treatment with original assignment.

Although the bonus program targets teachers, one might also expect it to induce changes in school-wide decisions. However, we find no evidence of institutional responses to the intervention (Table 7, Panel B). There are no significant treatment effects on teachers' use of student data. The second measure we examine from the teacher survey – whether teachers believed students in their school were held to high expectations – is negative but insignificant. These results provide little evidence that teachers substituted test prep for more complicated activities, which is not surprising given that we find no positive impacts on test scores.

*4.7 Teacher Characteristics and Turnover*

A second motivation for the use of performance-based pay is to increase the supply of high-ability individuals in an occupation. We investigate whether the bonus program led to changes in the quality of new teachers and reduced teacher turnover, in line with literature on sorting, which shows how schools serving disadvantaged students have difficulty hiring and retaining highly-qualified teachers (e.g., Clotfelter et al., 2006; Hanushek and Rivkin, 2007). If the bonus program increased the supply of qualified teachers willing to work at treatment schools, any resulting impacts on student achievement will lag these changes by at least a year.

We first examine whether the bonus program led to a reduction in teacher turnover. In a given year, approximately 10 percent of NYC teachers leave the city and 8 percent switch schools within the city. As shown in Panel A of Table 8, the bonus program did not reduce either type of turnover. Second, we examine whether treatment schools experienced an increase in the qualifications of newly hired teachers (Table 8, Panel B). The bonus program had positive but insignificant impacts on the proportion of new hires with a master's degree or prior teaching experience. These results are not surprising, given that, in the short run, the ability of the bonus program to increase retention and teacher quality is limited by the pool of existing teachers. It is

possible that widespread changes to the structure of teacher pay could induce individuals to enter the profession who might have otherwise chosen other occupations. Unfortunately, a small-scale program, such as the NYC bonus program, cannot speak to the long-run impacts of changing teacher pay more broadly.

## 5 Conclusion

In many sectors, performance-based pay enhances effort, output, and other desirable outcomes. However, despite significant expenditures on the NYC bonus program, we find little evidence that the program led to an overall increase in student achievement or had any impact on a variety of other outcomes, including classroom activities, tutoring, or administrative decisions. Nor did the program reduce teacher turnover or improve the quality of the teaching pool within eligible schools. We present suggestive evidence that students in treatment schools with fewer teachers or a more cohesive group of teachers experienced significantly higher math achievement. These results indicate that the group-based structure of the program may have been detrimental for the majority of schools and the diffusion of responsibility for test score gains among many teachers diluted the incentives of the opportunity to earn bonuses. Our results are consistent with the long-standing literature in economics on the importance of taking into consideration free-riding, joint production, and monitoring when designing incentive systems and suggest that a one-size-fits-all approach may not be the most effective when implementing incentive pay within a district.

Research provides evidence that threats of sanctions under NCLB and other accountability systems, including the NYC accountability system, increase student achievement. One interpretation of our results is that negative incentives are more effective than positive ones. Alternatively, incentive pay programs that come about as a compromise between school districts and teachers unions' might contain incentives that are so diluted they are destined to fail.[29] Finally, the extensive margin may be most important. In other professions, merit pay has equally large impacts on sorting into professions and effort provided by the existing workforce (Lazear, 2000). Small-scale teacher incentive pay experiments can not provide information concerning the

[29] A related concern is that teachers were aware that results from the experiment could impact future incentive pay policies and strategically did not respond due to union preferences against incentive pay. However, this would not lead to the free-riding effects we find.

general equilibrium effects of overall increase in teacher pay or movement towards performance-based compensation.

Currently, the U.S. government provides significant funding for school systems to pilot programs that introduce incentive pay for teachers. In 2010, 62 school districts and nonprofit groups received $442 million in funding from the federal Teacher Incentive Fund.[30] Eligibility for Race to the Top funding depends on districts' ability and willingness to link student achievement to individual teachers and use this data in teacher evaluations.[31] Our results underscore that the structure of performance pay is important and policy innovations in this area should be carefully designed.

[30] See http://www2.ed.gov/programs/teacherincentive/index.html and http://www.ed.gov/news/press-releases/department-education-announces-442-million-teacher-quality-grants-62-winners-27- for more information.
[31] See section D.2 in the Race to the Top evaluation criteria; available at http://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

**References**


Ahn, Tom. 2009. The missing link: estimating the impact of incentives on effort and effort on production using teacher accountability legislation. Unpublished manuscript, University of Kentucky Department of Economics.

Ballou, Dale. 2001. Pay for performance in public and private schools. *Economics of Education Review* 20, no. 1:51–61.

Ballou, Dale and Michael Podgursky. 1997. *Teacher pay and teacher quality.* Kalamazoo, MI: W.E. Upjohn Institution for Employment Research.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41, no. 4: 778–820.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2009. Are teacher absences worth worrying about in the U.S.? *Education Finance and Policy* 4, no. 2: 115–49.

Cullen, Julie Berry and Randall Reback. 2006. Tinkering toward accolades: school gaming under a performance accountability system. In *Advances in applied microeconomics volume 14: improving school accountability*, ed. Timothy J. Gronberg and Dennis W. Jansen. Oxford, UK: JAI Press.

Figlio, David N. 2006. Testing, crime, and punishment. *Journal of Public Economics* 90: 837-51.

Figlio, David N. and Lawrence S. Getzler. 2006. Accountability, ability, and disability: gaming the system? In *Advances in applied microeconomics volume 14: improving school accountability*, ed. Timothy J. Gronberg and Dennis W. Jansen. Oxford, UK: JAI Press.

Figlio, David N. and Lawrence W. Kenny. 2007. Individual teacher incentives and student performance. *Journal of Public Economics* 91: 901–14.

Gibbons, Robert. 1998. Incentives in organizations. *Journal of Economic Perspectives* 12, no. 4: 115–32.

Glazerman, Steven and Allison Seifullah. 2010. An evaluation of the teacher advancement program (TAP) in Chicago: year two impact report. Washington, DC: Mathematica Policy Research.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. Teacher incentives. *American Economic Journal: Applied Economics* 2, no. 3: 205–27.

Hanushek, Eric A. and Steven G. Rivkin. 2007. Pay, working conditions, and teacher quality. *The Future of Children* 17, no. 1: 69–86.

Herrmann, Mariesa A. and Jonah E. Rockoff. 2010. Work disruption, worker health and productivity: evidence from teaching. Unpublished manuscript, Columbia Business School.

Holmstrom, Bengt. 1982. Moral hazard in teams. *The Bell Journal of Economics* 13, no. 2: 324–40.

Holmstrom, Bengt and Paul Milgrom. 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7 (January): 24–52.

Hoxby, Caroline M. and Andrew Leigh. 2005. Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States. *American Economic Review*: 94, no. 2: 236–40.

Itoh, Hideshi. 1991. Incentives to help in multi-agent situations. *Econometrica* 59, no. 3: 611-36.

Jackson, C. Kirabo and Elias Bruegmann. 2009. Teaching students and teaching each other: the importance of peer learning for teachers. *American Economic Journal: Applied Economics* 1, no. 4: 1–27.

Jacob, Brian A. 2005. Accountability, incentives and behavior: evidence from school reform in Chicago. *Journal of Public Economics* 89: 761–96.

Jacob, Brian A. and Steven D. Levitt. 2003. Rotten apples: an investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics* 118, no. 3: 843–77.

Lavy, Victor. 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. *The Journal of Political Economy* 110, no. 6: 1286–1317.

———. 2009. Performance pay and teachers' effort, productivity and grading ethics. *American Economic Review* 99, no. 5: 1979–2011.

Lazear, Edward P. 2000. Performance pay and productivity. *American Economic Review* 90, no. 5: 1346–61.

———. 2003. Teacher Incentives. *Swedish Economic Policy Review* 10: 179–214.

Lazear, Edward P. and Paul Oyer. Forthcoming. Personnel economics. In *Handbook of organizational economics*, ed. Robert Gibbons and D. John Roberts. Princeton, NJ: Princeton University Press.

Lazear, Edward P. and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimal Labor Contracts." *Journal of Political Economy* 89, no. 5: 841–64.

Miller, Raegen T. Richard J. Murnane, and John B. Willett. 2008. Do worker absences affect productivity? The case of teachers. *International Labour Review* 147, no. 1: 71–89.

Muralidharan, Karthik and Venkatesh Sundararaman. 2009. Teacher performance pay: experimental evidence from India. Working paper no. 15323, National Bureau of Economic Research, Cambridge, MA.

Neal, Derek and Diane Whitmore Schanzenbach. 2010. Left behind by design: proficiency counts and test-based accountability. *Review of Economics and Statistics* 92, no. 2: 263–83.

Podgursky, Michael J. and Matthew G. Springer. 2007. Teacher performance pay: a review. *Journal of Policy Analysis and Management* 26, no. 4: 909–49.

Rau, Tomas and Dante Contreras. 2009. Tournaments, gift exchanges, and the effect of monetary incentive for teachers: the case of Chile. Working Paper no. 305, University of Chile, Department of Economics.

Rockoff, Jonah and Lesley J. Turner. 2010. Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy* 2, no. 4: 119–47.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, Brian M. Stecher. 2010. Teacher pay for performance: experimental evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Springer, Matthew G. and Marcus A. Winters. 2009. The NYC teacher pay-for-performance program: early evidence from a randomized trial. Manhattan Institute Civic Report No. 56.

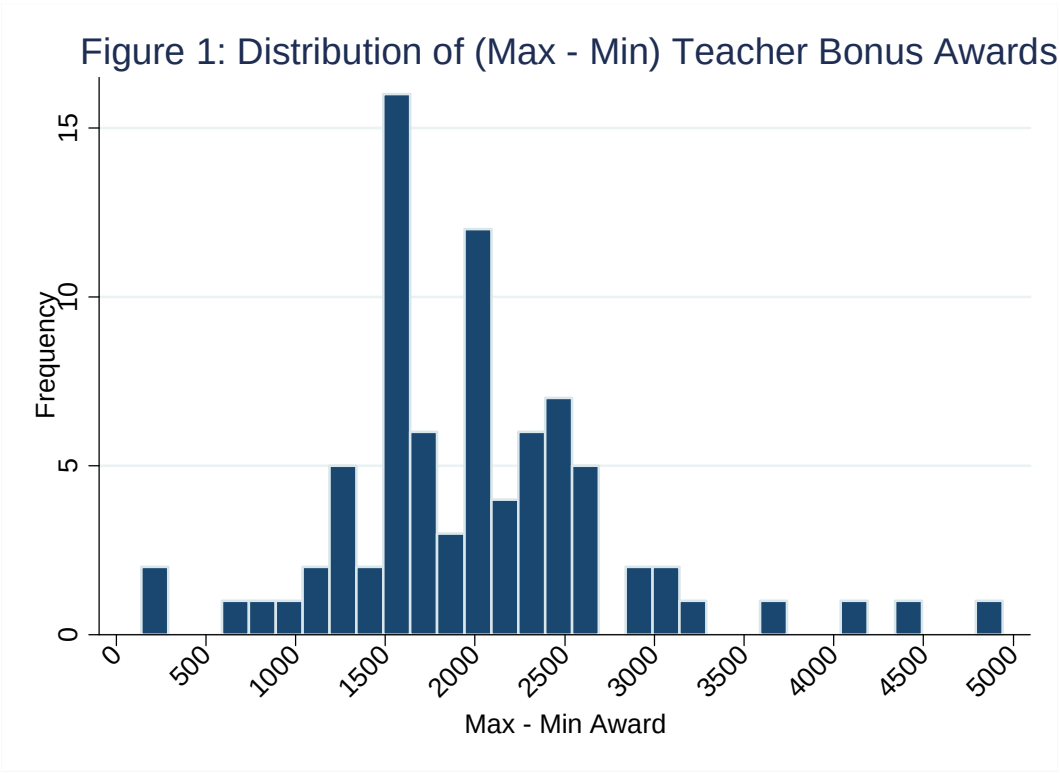
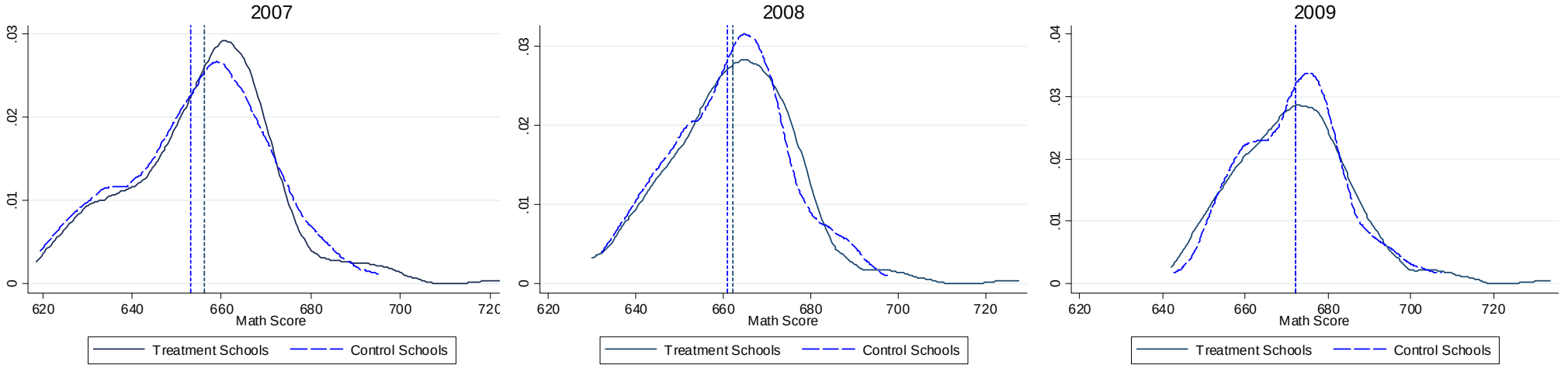


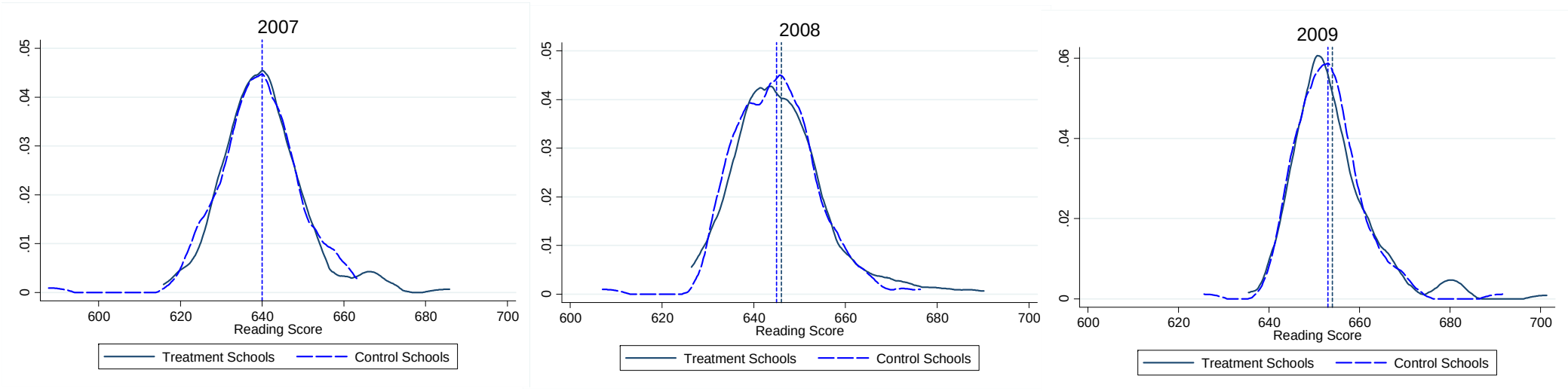
Figure 1: Distribution of (Max - Min) Teacher Bonus Awards
Frequency
0
5
10
15
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
Max - Min Award

Figure 2: Distribution of Average Math Scores by Year and Treatment Status




Note: Dashed vertical lines denote mean math scores for treatment and control schools.

Figure 3: Distribution of Average Reading Scores by Year and Treatment Status




Note: Dashed vertical lines denote mean reading scores for treatment and control schools.

Table 1: Baseline School Characteristics by Original Assignment to Treatment and Control Groups

| | Treatment Schools | Control Schools | Difference | p-value | Non-Experimental Schools |
|---|---|---|---|---|---|
| Number of Schools | 181 | 128 | | | 614 |
| Average enrollment | 558 | 558 | 0 | 0.852 | 687 |
| Average enrollment, tested grades | 363 | 367 | -4 | 0.912 | 459 |
| Fraction elementary school | 0.62 | 0.63 | -0.01 | 0.788 | 0.63 |
| Fraction middle school | 0.26 | 0.27 | -0.01 | 0.586 | 0.24 |
| Fraction K-8 school | 0.12 | 0.10 | 0.02 | 0.452 | 0.13 |
| *School Accountability Outcomes* | | | | | |
| Peer index (mean = 0, sd = 1) | -0.91 | -0.93 | 0.02 | 0.452 | 0.44 |
| Overall accountability score | 52.6 | 52.1 | 0.6 | 0.750 | 54.6 |
| Target score | 66.3 | 65.9 | 0.4 | 0.772 | 67.8 |
| *Student Characteristics* | | | | | |
| Average math scale score (2007) | 656 | 655 | 1 | 0.497 | 677 |
| Change in math scale score (2006 to 2007) | 10.5 | 10.3 | 0.2 | 0.741 | 8.7 |
| Average reading scale score (2007) | 640 | 640 | 1 | 0.603 | 660 |
| Change in reading scale score (2006 to 2007) | 1.4 | 1.9 | -0.5 | 0.511 | 0.9 |
| Fraction English Language Learner | 0.19 | 0.19 | 0.01 | 0.614 | 0.11 |
| Fraction special education | 0.12 | 0.13 | -0.01 | 0.246 | 0.09 |
| Fraction free lunch | 0.87 | 0.89 | -0.02 | 0.315 | 0.62 |
| Fraction Hispanic | 0.56 | 0.53 | 0.03 | 0.428 | 0.33 |
| Fraction Black | 0.41 | 0.43 | -0.03 | 0.425 | 0.29 |
| Fraction White | 0.01 | 0.01 | 0.00 | 0.640 | 0.20 |
| *Teacher Characteristics* | | | | | |
| Number of teachers | 55 | 55 | 0 | 0.952 | 60 |
| Number of teachers, tested classrooms | 16 | 16 | -1 | 0.431 | 17 |
| Average years of experience | 7.9 | 8.0 | -0.1 | 0.703 | 8.6 |
| Fraction with masters degree | 0.49 | 0.49 | 0.01 | 0.579 | 0.47 |
| Average absences/teacher (2007) | 7.2 | 7.0 | 0.2 | 0.447 | 6.7 |
| Average absences/teacher, tested classrooms (2007) | 7.4 | 7.2 | 0.3 | 0.377 | 7.0 |
| Fraction teachers not retained by DOE (2007) | 0.11 | 0.12 | 0.00 | 0.513 | 0.09 |
| Fraction teachers changing schools (2007) | 0.07 | 0.07 | 0.01 | 0.321 | 0.04 |
| Fraction of new teachers with MA | 0.34 | 0.37 | -0.03 | 0.407 | 0.45 |
| Fraction of new teachers with prior experience | 0.26 | 0.28 | -0.02 | 0.493 | 0.39 |

Notes: Characteristics measured at beginning of 2007-2008 school year unless otherwise noted; average absences per teacher include absences taken for personal or sick leave.

Table 2: Impact of Teacher Incentives on Student Math and Reading Achievement

| | Reading | | | | Math | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean (sd) | (1) OLS | (2) OLS | (3) IV | Mean (sd) | (4) OLS | (5) OLS | (6) IV |
| *A. Year 1: 2007 - 2008* | | | | | | | | |
| Treatment | 655 | -0.876 | -0.395 | -0.486 | 672 | -1.418 | -0.789 | -0.970 |
| | (35) | (1.084) | (0.488) | (0.589) | (40) | (1.737) | (0.524) | (0.632) |
| Observations | | 309 | 309 | 309 | | 309 | 309 | 309 |
| *B. Year 2: 2008 - 2009* | | | | | | | | |
| Treatment | 662 | -0.852 | -0.584 | -0.734 | 680 | -1.637 | -1.385 | -1.740 |
| | (31) | (0.930) | (0.533) | (0.660) | (37) | (1.652) | (0.655)* | (0.813)* |
| Observations | | 305 | 305 | 305 | | 305 | 305 | 305 |
| Additional covariates | | | X | X | | | X | X |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; first column displays mean and sd score across all NYC students; each cell denotes a separate regression, dependent variable: school average reading or math score; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: indicators for school level, pre-treatment (2007) math or reading scale score, pre-treatment (2007) peer index and overall accountability score, and pre-treatment (fall 2007) student demographic characteristics: percentage ELL, special education, free lunch recipients, and student race (African American and Hispanic); sample sizes differ across years due to the closure of four schools at the end of the 2007-2008 school year.

Table 3: Impact of Teacher Incentives on the Percentage of Students Achieving Proficiency

| | Reading | | | | Math | | | |
|---|---|---|---|---|---|---|---|---|
| | Sample Mean | (1) OLS | (2) OLS | (3) IV | Sample Mean | (4) OLS | (5) OLS | (6) IV |
| *B. Year 1: 2007 - 2008* | | | | | | | | |
| Treatment | 0.45 | -0.020 | -0.009 | -0.011 | 0.67 | -0.014 | -0.009 | -0.012 |
| | | (0.017) | (0.006) | (0.008) | | (0.020) | (0.007) | (0.008) |
| Observations | | 309 | 309 | 309 | | 309 | 309 | 309 |
| *C. Year 2: 2008 - 2009* | | | | | | | | |
| Treatment | 0.58 | -0.019 | -0.013 | -0.016 | 0.77 | -0.018 | -0.017 | -0.021 |
| | | (0.014) | (0.008) | (0.010) | | (0.016) | (0.007)* | (0.008)* |
| Observations | | 305 | 305 | 305 | | 305 | 305 | 305 |
| Additional covariates | | | X | X | | | X | X |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; first column displays mean and sd score across all NYC students; each cell denotes a separate regression, dependent variable: percentage of students proficient in math or reading; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: indicators for school level, pre-treatment (2007) percentage of students proficient, pre-treatment (2007) peer index and overall accountability score, and pre-treatment (fall 2007) student demographic characteristics: percentage ELL, special education, free lunch recipients, and student race (African American and Hispanic); sample sizes differ across years due to the closure of four schools at the end of the 2007-2008 school year.

Table 4: Free-riding and the Impact of Teacher Incentives on Student Math and Reading Achievement

| | Reading | | | Math | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *A. Year 1: 2007-2008* | | | | | | |
| Treatment | -0.372 | 0.046 | -0.667 | -0.871 | -0.536 | -1.445 |
| | (0.490) | (0.499) | (0.519) | (0.530) | (0.568) | (0.561)* |
| * Number of teachers *(mean = 0)* | | -0.233 | | | -0.176 | |
| | | (0.089)** | | | (0.097)+ | |
| * First quartile of number of teachers | | | 2.044 | | | 4.670 |
| | | | (1.575) | | | (1.483)** |
| Treatment effect: schools in first quartile | | | 1.377 | | | 3.225 |
| | | | (1.481) | | | (1.395)* |
| Observations | 300 | 300 | 300 | 301 | 301 | 301 |
| *B. Year 2: 2008-2009* | | | | | | |
| Treatment | -0.579 | -0.395 | -0.909 | -1.297 | -0.979 | -1.893 |
| | (0.539) | (0.572) | (0.556) | (0.668)+ | (0.726) | (0.689)** |
| * Number of teachers *(mean = 0)* | | -0.126 | | | -0.171 | |
| | | (0.099) | | | (0.144) | |
| * First quartile of number of teachers | | | 2.122 | | | 4.826 |
| | | | (2.067) | | | (2.579)+ |
| Treatment effect: schools in first quartile | | | 1.213 | | | 2.933 |
| | | | (1.968) | | | (2.461) |
| Observations | 294 | 294 | 294 | 294 | 294 | 294 |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column denotes a separate regression; robust standard errors in parentheses; measures of the number of reading/math teachers are demeaned; additional controls include: pre-treatment (2007) school test score, school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); regressions are weighted by number of tested students; schools with no teachers linked to tested students are dropped; the number of math teachers for schools in the first quartile is less than or equal to: 10 (elementary and K-8 schools), 5 (middle schools); the number of reading teachers for schools in the first quartile is less than or equal to: 10 (elementary and K-8 schools), 6 (middle schools).

Table 5: School Cohesion and the Impact of Teacher Incentives on Student Math and Reading Achievement

| | Reading | | | Math | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *A. Year 1: 2007-2008* | | | | | | |
| Treatment | -0.310 | -0.067 | -0.912 | -0.643 | -0.274 | -1.288 |
| | (0.492) | (0.511) | (0.592) | (0.529) | (0.549) | (0.674)+ |
| * Cohesion index | | 0.316 | | | 0.797 | |
| | | (0.545) | | | (0.620) | |
| * Above average cohesion index | | | 1.888 | | | 1.987 |
| | | | (0.968)+ | | | (1.128)+ |
| Treatment effect: schools with above average cohesion | | | 0.976 | | | 0.698 |
| | | | (0.778) | | | (0.887) |
| Observations | 300 | 300 | | 301 | 301 | 301 |
| *B. Year 2: 2008-2009* | | | | | | |
| Treatment | -0.498 | -0.267 | -1.153 | -1.044 | -0.537 | -2.326 |
| | (0.533) | (0.554) | (0.661)+ | (0.654) | (0.669) | (0.838)** |
| * Cohesion index | | 0.406 | | | 1.266 | |
| | | (0.598) | | | (0.774) | |
| * Above average cohesion index | | | 1.982 | | | 3.692 |
| | | | (1.074)+ | | | (1.390)** |
| Treatment effect: schools with above average cohesion | | | 0.829 | | | 1.367 |
| | | | (0.847) | | | (1.070) |
| Observations | 299 | 299 | 299 | 300 | 300 | 300 |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; robust standard errors in parentheses; each column denotes a separate regression; teacher cohesion index mean = 0, sd = 1 across all NYC schools; additional controls include: pre-treatment (2007) school test score, school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); regressions are weighted by number of tested students; schools with teacher survey response rate below 10% are dropped.

Table 6: The Impact of Teacher Incentives on Average Absences per Teacher Due to Personal and Sick Leave

| | All | Teachers of Tested Students | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *A. Year 1: 2007-2008* | | | | |
| Treatment | 0.001 | -0.158 | -0.217 | -0.156 |
| | (0.091) | (0.146) | (0.148) | (0.163) |
| * Number of teachers *(mean = 0)* | | | 0.013 | |
| | | | (0.022) | |
| * First quartile of number of teachers | | | | -0.236 |
| | | | | (0.390) |
| Treatment effect: schools in first quartile | | | | -0.391 |
| | | | | (0.352) |
| Observations | 301 | 301 | 301 | 301 |
| *B. Year 2: 2008-2009* | | | | |
| Treatment | 0.045 | 0.151 | 0.203 | 0.161 |
| | (0.119) | (0.175) | (0.192) | (0.200) |
| * Number of teachers *(mean = 0)* | | | 0.005 | |
| | | | (0.032) | |
| * First quartile of number of teachers | | | | 0.158 |
| | | | | (0.621) |
| Treatment effect: schools in first quartile | | | | 0.319 |
| | | | | (0.576) |
| Observations | 294 | 294 | 294 | 294 |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column within a panel denotes a separate regression; measures of the number of reading/math teachers are demeaned; dependent variable is average absences per teacher taken for personel or sick leave between November and March (Panel A) or September and March (Panel B); additional controls include: pre-treatment (2007) school test score, school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); regressions are weighted by number of tested students; schools with no teachers linked to tested students are dropped; the number of teachers for schools in the first quartile is less than or equal to: 10 (elementary schools),11 (middle and K-8 schools).

Table 7: Impact of Teacher Incentives on Student and Teacher Survey Outcomes

| | (1) 2008 | (2) 2009 |
|---|---|---|
| *A. Student Survey Outcomes* | | |
| Essays and Projects | 0.056 | -0.028 |
| | (0.148) | (0.158) |
| Group & Hands-on Learning Activities | 0.081 | 0.089 |
| | (0.180) | (0.149) |
| Tutoring Offered Before/After School | 0.148 | 0.199 |
| | (0.148) | (0.144) |
| Observations | 128 | 129 |
| *B. Teacher Survey Outcomes* | | |
| Use of Student Data | -0.052 | -0.114 |
| | (0.103) | (0.112) |
| High Expectations For Students | -0.105 | -0.077 |
| | (0.091) | (0.099) |
| Observations | 309 | 305 |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each cell denotes a separate regression; survey outcomes standardized to have mean = 0, sd = 1 across all NYC schools; robust standard errors in parentheses; all regressions control for: pre-treatment (2007) survey outcome, survey response rate, school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); regressions are weighted by the number of students or teachers surveyed; sample sizes differ across years due to the closure of four schools at the end of the 2007-2008 school year and the elimination of schools with no student survey respondents (Panel A).

Table 8: The Impact of Teacher Incentives on Turnover and the Qualifications of New Teachers

| | Year 1: 2008 | | Year 2: 2009 | |
|---|---|---|---|---|
| | Sample Mean | Treatment Effect | Sample Mean | Treatment Effect |
| *A. Teacher Turnover* | | | | |
| Fraction of teachers not retained by school district | 0.11 | 0.004<br>(0.006) | 0.08 | 0.003<br>(0.005) |
| Fraction of teachers leaving for another NYC school | 0.07 | 0.005<br>(0.005) | 0.07 | 0.009<br>(0.009) |
| Observations | | 305 | | 305 |
| *B. New Teacher Characteristics* | | | | |
| Fraction of new teachers with MA | | | 0.37 | 0.019<br>(0.038) |
| Fraction of new teachers with prior teaching experience | | | 0.28 | 0.029<br>(0.029) |
| Observations | | | | 261 |

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

Table A1: Baseline School Characteristics by Participation Vote

| | Voted "yes" | Voted "no" | Difference | p-value |
|---|---|---|---|---|
| Number of Schools | 158 | 25 | | |
| Average enrollment | 558 | 574 | -16 | 0.754 |
| Average enrollment, tested grades | 364 | 361 | 3 | 0.939 |
| Fraction elementary school | 0.61 | 0.72 | -0.11 | 0.284 |
| Fraction middle school | 0.27 | 0.20 | 0.07 | 0.487 |
| Fraction K-8 school | 0.13 | 0.08 | 0.05 | 0.508 |
| *School Accountability Outcomes* | | | | |
| Peer index (mean = 0, sd = 1) | -0.91 | -0.87 | -0.05 | 0.247 |
| Overall accountability score | 52.5 | 55.1 | -2.5 | 0.452 |
| Target score | 66.3 | 68.2 | -1.9 | 0.480 |
| *Student Characteristics* | | | | |
| Average math scale score (2007) | 655 | 661 | -6 | 0.102 |
| Change in math scale score (2006 to 2007) | 10.7 | 10.2 | 0.5 | 0.704 |
| Average reading scale score (2007) | 640 | 644 | -5 | 0.040 |
| Change in reading scale score (2006 to 2007) | 1.7 | 0.2 | 1.4 | 0.316 |
| Fraction English Language Learner | 0.20 | 0.18 | 0.02 | 0.549 |
| Fraction special education | 0.12 | 0.12 | 0.00 | 0.773 |
| Fraction free lunch | 0.88 | 0.86 | 0.02 | 0.608 |
| Fraction Hispanic | 0.56 | 0.54 | 0.03 | 0.672 |
| Fraction Black | 0.41 | 0.42 | -0.01 | 0.868 |
| Fraction White | 0.01 | 0.01 | 0.00 | 0.772 |
| *Teacher Characteristics* | | | | |
| Number of teachers | 55 | 56 | -2 | 0.707 |
| Number of teachers, tested classrooms | 15 | 17 | -2 | 0.237 |
| Average years of experience | 7.9 | 8.4 | -0.6 | 0.163 |
| Fraction with masters degree | 0.49 | 0.50 | 0.00 | 0.841 |
| Average absences (2007) | 7.1 | 7.1 | 0.0 | 0.426 |
| Average absences, tested classrooms (2007) | 7.0 | 7.2 | -0.2 | 0.775 |
| Fraction teachers not retained by DOE (2007) | 0.11 | 0.10 | 0.02 | 0.184 |
| Fraction teachers changing schools (2007) | 0.07 | 0.07 | 0.00 | 0.996 |
| Fraction of new teachers with MA | 0.38 | 0.36 | 0.03 | 0.725 |
| Fraction of new teachers with prior experience | 0.31 | 0.35 | -0.04 | 0.576 |

Notes: Characteristics measured at beginning of 2007-2008 school year unless otherwise noted; average absences per teacher include absences taken for personal or sick leave.

Table A2: Heterogeneity in the Impact of Teacher Incentives on Student Achievement by ELL/Special Education Status and Tercile of Prior Achievement, 2007-2008

| | Reading | | | Math | | |
|---|---|---|---|---|---|---|
| | (1) OLS | (2) OLS | (3) IV | (4) OLS | (5) OLS | (6) IV |
| Treatment | | | | | | |
| * ELL or Special Education | 0.115 | 0.160 | 0.194 | -0.570 | -0.327 | -0.399 |
| | (1.198) | (0.783) | (0.945) | (2.018) | (0.775) | (0.935) |
| * Bottom Third | -0.104 | -0.486 | -0.607 | -0.768 | -0.639 | -0.793 |
| | (1.043) | (0.578) | (0.710) | (1.823) | (0.633) | (0.787) |
| * Middle Third | -0.850 | -0.288 | -0.355 | -0.984 | -0.528 | -0.652 |
| | (1.040) | (0.447) | (0.541) | (1.720) | (0.484) | (0.597) |
| * Upper Third | -1.477 | -0.228 | -0.284 | -2.065 | -0.838 | -1.041 |
| | (1.402) | (0.620) | (0.762) | (1.953) | (0.643) | (0.789) |
| Test of equality (p-value) | 0.312 | 0.846 | 0.926 | 0.251 | 0.868 | 0.959 |
| Observations | 1,232 | 1,231 | 1,231 | 1,232 | 1,231 | 1,231 |
| Additional covariates | | X | X | | X | X |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; first column displays mean and sd score across all NYC students; each cell denotes a separate regression, dependent variable: school average reading or math score; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: indicators for school level, pre-treatment (2007) math or reading scale score, pre-treatment (2007) peer index and overall accountability score, and pre-treatment (fall 2007) student demographic characteristics: percentage ELL, special education, free lunch recipients, and student race (African American and Hispanic); sample sizes differ across years due to the closure of four schools at the end of the 2007-2008 school year.

Table A3: The Effect of Teacher Incentives on the Percentage and Composition of Tested Students, 2007 - 2008

| | Reading | | | Math | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | OLS | OLS | IV | OLS | OLS | IV |
| *Dependent Variable:* | | | | | | |
| Percentage of Students Tested | -0.001 | -0.003 | -0.003 | 0.001 | 0.000 | 0.000 |
| | (0.005) | (0.003) | (0.004) | (0.004) | (0.003) | (0.004) |
| Observations | 309 | 309 | 309 | 309 | 309 | 309 |
| Percentage of tested students ELL | 0.010 | 0.002 | 0.003 | 0.008 | -0.000 | -0.000 |
| | (0.016) | (0.002) | (0.002) | (0.016) | (0.001) | (0.002) |
| Observations | 257 | 257 | 257 | 260 | 260 | 260 |
| Percentage of tested students | -0.000 | 0.002 | 0.002 | -0.002 | 0.000 | 0.000 |
| special education | (0.006) | (0.003) | (0.003) | (0.006) | (0.003) | (0.003) |
| Observations | 295 | 295 | 295 | 294 | 294 | 294 |
| Additional covariates | | X | X | | 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) percentage of students tested, percentage ELL, or percentage special education; additional controls include: school level, peer index, overall accountability score, free lunch recipients, and student race (African American and Hispanic), regressions of percentage of students tested weighted by total enrollment, all other regressions weighted by number of tested students; in column (3) and (6) regressions, actual treatment status is instrumented for with original treatment assignment.

Table A4: Heterogeneity in Impact of Teacher Incentives on Student Math and Reading Achievement by Accountability Grade

| | Reading | | | Math | | |
|---|---|---|---|---|---|---|
| | (1) OLS | (2) OLS | (3) IV | (4) OLS | (5) OLS | (6) IV |
| *A. Year 1: 2007 - 2008* | | | | | | |
| Treatment*D or F | -3.719 | 0.175 | 0.188 | -5.292 | -0.329 | -0.379 |
| | (2.147)+ | (1.116) | (1.244) | (3.407) | (1.144) | (1.279) |
| Treatment*B or C | 0.454 | -0.733 | -0.892 | 0.944 | -0.400 | -0.470 |
| | (1.309) | (0.615) | (0.726) | (2.163) | (0.700) | (0.825) |
| Treatment* A | -1.856 | 0.063 | 0.091 | -3.703 | -1.542 | -2.031 |
| | (2.489) | (1.122) | (1.439) | (3.608) | (1.084) | (1.447) |
| Test A/B = C = D/F (pvalue) | 0.231 | 0.698 | 0.685 | 0.238 | 0.653 | 0.625 |
| Observations | 309 | 309 | 309 | 309 | 309 | 309 |
| *B. Year 2: 2008 - 2009* | | | | | | |
| Treatment*D or F | -3.533 | -1.924 | -3.246 | -7.273 | -3.326 | -5.582 |
| | (2.321) | (1.300) | (2.178) | (4.017)+ | (2.388) | (3.924) |
| Treatment*B or C | -0.488 | -0.379 | -0.465 | -0.686 | -0.463 | -0.574 |
| | (0.976) | (0.435) | (0.518) | (1.888) | (0.658) | (0.786) |
| Treatment* A | 1.179 | -0.091 | -0.113 | 1.511 | -0.603 | -0.749 |
| | (1.802) | (0.686) | (0.843) | (2.726) | (0.952) | (1.174) |
| Test A/B = C = D/F (pvalue) | 0.277 | 0.450 | 0.400 | 0.193 | 0.515 | 0.460 |
| Observations | 305 | 302 | 302 | 305 | 302 | 302 |
| Additional covariates | | X | X | | X | X |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column within a panel denotes a separate regression; dependent variable: school average reading or math scale score interacted with indicator for school grade; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: prior year scale score, indicators for school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); sample sizes differ across years due to the closure of four schools at the end of the 2007-2008 school year and the elimination of an additional three schools that did not receive 2008 accountability grades.

Table A5: Heterogeneity in Impact of Teacher Incentives on Student Achievement by Bonus Receipt, 2008 - 2009

| | Reading | | | Math | | |
|---|---|---|---|---|---|---|
| | (1) OLS | (2) OLS | (3) IV | (4) OLS | (5) OLS | (6) IV |
| Treatment | -0.759 | -0.431 | -0.568 | -0.804 | -0.673 | -0.684 |
| | (0.992) | (0.447) | (0.680) | (1.899) | (0.690) | (1.052) |
| * Any Bonus (predicted) | 0.944 | 0.287 | 0.386 | 0.269 | 0.047 | -0.148 |
| | (1.873) | (0.741) | (1.133) | (3.238) | (1.080) | (1.662) |
| Observations | 302 | 302 | 302 | 302 | 302 | 302 |
| Additional covariates | | X | X | | X | X |

Notes: + significant at 10%; * significant at 5%; ** significant at 1%; each column within a panel denotes a separate regression; dependent variable: school average reading or math scale score interacted with indicator for school grade; robust standard errors in parentheses; all regressions weighted by number of students tested in math or reading; additional covariates include: prior year scale score, indicators for school level, peer index, overall accountability score, percentage of students ELL, special education, free lunch recipients, and student race (African American and Hispanic); in column (3), actual treatment status is instrumented for with original treatment assignment