NBER WORKING PAPER SERIES

TEACHER INCENTIVES AND STUDENT ACHIEVEMENT: EVIDENCE FROM NEW YORK CITY PUBLIC SCHOOLS

Roland G. Fryer

Working Paper 16850
http://www.nber.org/papers/w16850

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2011

This project would not have been possible without the leadership and support of Joel Klein. I am also grateful to Jennifer Bell-Ellwanger, Joanna Cannon, and Dominique West for their cooperation in collecting the data necessary for this project, and to my colleagues Edward Glaeser, Richard Holden, and Lawrence Katz for helpful comments and discussions. Vilsa E. Curto, Meghan L. Howard, Won Hee Park, Jörg Spenkuch, David Toniatti, Rucha Vankudre, and Martha Woerner provided excellent research assistance. Financial Support from the Fisher Foundation is gratefully acknowledged. The usual caveat applies. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Teacher Incentives and Student Achievement: Evidence from New York City Public Schools
Roland G. Fryer
NBER Working Paper No. 16850
March 2011
JEL No. I0,J0

ABSTRACT

Financial incentives for teachers to increase student performance is an increasingly popular education policy around the world. This paper describes a school-based randomized trial in over two-hundred New York City public schools designed to better understand the impact of teacher incentives on student achievement. I find no evidence that teacher incentives increase student performance, attendance, or graduation, nor do I find any evidence that the incentives change student or teacher behavior. If anything, teacher incentives may decrease student achievement, especially in larger schools. The paper concludes with a speculative discussion of theories that may explain these stark results.

Roland G. Fryer
Department of Economics
Harvard University
Littauer Center 208
Cambridge, MA 02138
and NBER
rfryer@fas.harvard.edu

When I was in Chicago, our teachers designed a program for performance pay and secured a $27 million federal grant. ... In Chicago's model – every adult in the building – teachers, clerks, janitors and cafeteria workers – all were rewarded when the school improved. It builds a sense of teamwork and gives the whole school a common mission. It can transform a school culture.

- Secretary of Education Arne Duncan, The National Press Club, July 27, 2010

1 Introduction

Many educational reforms have been attempted over the past few decades with the goal of increasing academic achievement: lowering class size, increasing spending, providing incentives for teachers to obtain more education, and so on.[1] In 1961, 23.5 percent of teachers had a Master's degree or a higher degree. In 2001, 56.8 percent of teachers had at least a Master's degree. Student to teacher ratios in public schools have decreased from over 22 to 1 in 1971 to less than 16 to 1 in 2001, a decrease of 33 percent in class size in 30 years. America spends more on education than it has ever before: per-pupil spending has increased (in 2005 dollars) from approximately $4,700 per student in 1970 to over $10,000 (Snyder and Dillow, 2010). Yet, despite these reforms to increase achievement, Figure 1 demonstrates that test scores have been largely constant over the past thirty years.

Human capital, especially teacher quality, is believed to be one of the most important inputs into education production. A one-standard deviation increase in teacher quality raises math achievement by 0.15 to 0.24 standard deviations per year and reading achievement by 0.15 to 0.20 standard deviations per year (Rockoff, 2004; Hanushek and Rivkin, 2005; Aaronson et al., 2007; Kane and Staiger, 2008). The difficulty, however, is that one cannot identify *ex ante* the most productive teachers. Value added measures are not strongly correlated with observable teacher characteristics

[1] There have been many other attempts to increase achievement and close the achievement gap. Early childhood interventions such as Head Start, Nurse-Family Partnership, and the Abecedarian Project boost kindergarten readiness, but the effects on achievement often fade once children enter school (Currie and Thomas, 1995; Olds, 2006). More aggressive strategies that place disadvantaged students in better schools through busing (Angrist and Lang, 2004) or through school choice plans (Rouse, 1998; Krueger and Zhu, 2004; Cullen et al., 2005; Hastings et al., 2006) have also left the racial achievement gap essentially unchanged. School districts have been sources of innovative strategies, including smaller schools and classrooms (Fin and Achilles, 1999; Nye et al., 1995; Krueger, 1999; Krueger and Whitmore, 2001; Jepsen and Rivkin, 2002), mandatory summer school (Jacob and Lefgren, 2004), student incentives (Bettinger, 2010; Fryer, 2010), after-school programs (Lauer et al., 2006; Redd et al., 2002), budget, curricula, and assessment reorganization (Borman et al., 2003; Borman et al., 2007; Cook et al., 2000), and policies to lower the barrier to teaching via alternative paths to accreditation (Decker et al., 2004; Kane et al., 2008).

(Aaronson et al., 2007; Rivkin et al., 2005; Kane and Staiger, 2008; Rockoff et al., 2008). Some argue that this, coupled with the inherent challenges in removing low performing teachers and increased job market opportunities for women, contributes to the fact that teacher quality and aptitude has declined significantly in the past 40 years (Corcoran et al., 2004; Hoxby and Leigh, 2004).[2]

One potential method to increase student achievement and improve the quality of individuals selecting teaching as a profession is to provide teachers with financial incentives based on student achievement. Theoretically, teacher incentives could have one of the three effects. If teachers lack motivation or incentive to put effort into lesson planning, parental engagement, and so on, financial incentives for student achievement may have a positive impact by motivating teachers to increase their effort. If, however, teacher incentives have unintended consequences such as explicit cheating, teaching to the test, or focusing on specific, tested objectives at the expense of more general learning, teacher incentives can have a negative impact on student performance (Holmstrom and Milgrom, 1991; Jacob and Levitt, 2003). Similarly, some argue that teacher incentives can decrease a teacher's intrinsic motivation or lead to harmful competition between teachers in what some believe to be a collaborative environment (Johnson, 1984; Firestone and Pennell, 1993). Third, if teachers do not know how to increase student achievement, the production function has important complementarities outside their control, or the incentives are either confusing or too weak, teacher incentives may have no impact on achievement.

There has been growing enthusiasm among education reformers and policy makers around the world to link teacher compensation to student achievement in myriad ways.[3] This is due, in part, to the low correlation between a teacher's observables at the time of hiring and his value-added, and, in part, to policy makers' belief that a new payment scheme will attract more achievement-minded applicants. A number of states, including Colorado, Florida, Michigan, Minnesota, South Carolina,

[2]Corcoran et al. (2004) find that in the 1964-1971 period, 20-25 percent of new female teachers were ranked in the top 10 percent of their high school cohort, while in 2000, less than 13 percent were ranked at the top decile. Hoxby and Leigh (2004) similarly find that the share of teachers in the highest aptitude category fell from 5 percent in 1963 to 1 percent in 2000 and the share in the lowest aptitude category rose from 16 percent to 36 percent in the same period.

[3]Merit pay faces opposition from the the two major unions: The American Federation of Teachers (AFT) and the National Education Association (NEA). Though in favor of reforming teacher compensation systems, the AFT and the NEA officially object to programs that reward teachers based on student test scores and principal evaluations, while favoring instead systems that reward teachers based on additional roles and responsibilities they take within the school or certifications and qualifications they accrue. The AFT's official position cites the past underfunding of such programs, the confusing metrics by which teachers were evaluated, and the crude binary reward system in which there is no gradation of merit as the reasons for its objection. The NEA's official position maintains that any alterations in compensation should be bargained at the local level, and that a singular salary scale and a strong base salary should be the standard for compensation.

Tennessee, Texas, and Washington, D.C., have implemented statewide programs for districts and schools to provide individual and group incentives to teachers for student achievement and growth, and many more individual school districts have implemented similar policies. In 2010, the U.S. Department of Education selected 62 programs in 27 states to receive over $1.2 billion over five years from the Teacher Incentive Fund. States applying for funds from "Race to the Top," the Obama Administration's $4.4 billion initiative to reform schools, are evaluated on plans to improve teacher and principal effectiveness by linking teacher evaluations to student growth and making decisions about raises, tenure, and promotions depending on student achievement. Similar initiatives are underway in the United Kingdom, Chile, Mexico, Israel, Australia, Portugal, and parts of India.

The empirical evidence on the efficacy of teacher incentives is ambivalent. Data from field experiments in Kenya and India yield effect sizes of approximately 0.20 standard deviations in math and reading (Glewwe et al., 2010; Muralidharan and Sundararaman, forthcoming). Data from a pilot initiative in Tennessee suggests no effects of incentives on student achievement. Other, non-experimental analyses, of teacher incentive programs in the United States have also shown no measurable success, though one should interpret these data with caution due to the lack of credible causal estimates (Glazerman et al., 2009; Vigdor, 2008).

In the 2007-2008 through the 2009-2010 school year, the United Federation of Teachers (UFT) and the New York City Department of Education (DOE) implemented a teacher incentive program in over 200 high-need schools, distributing a total of roughly $75 million to over 20,000 teachers.[4] The experiment was a randomized school-based trial, with the randomization conducted by the author. Each participating school could earn $3,000 for every UFT-represented staff member, which the school could distribute at its own discretion, if the school met the annual performance target set by the DOE based on school report card scores. Each participating school was given $1,500 per UFT staff member if it met at least 75% of the target but not the full target. Note: given that the average New York City public school has roughly sixty teachers, this implies a transfer of $180,000 to schools on average if they met their annual targets and a transfer of $90,000 if they met at least 75% of, but not the full target. School report card scores hinge on student performance and progress on state test scores for elementary and middle schools, Regents exam results and graduation rates for high schools, student attendance, and learning environment survey results administered to teachers, parents and students for all schools.

[4]The details of the program were negotiated by Chancellor Joel Klein and Randi Weingarten, along with their staffs. At the time of the negotiation, I was serving as an advisor to Chancellor Klein and convinced both parties that we should include random assignment to ensure a proper evaluation.

An important feature of our experiment is that schools had discretion over their incentive plans. As mentioned above, if a participating school met one-hundred percent of the annual targets, it received a lump sum equivalent to $3000 per full-time unionized teacher. Each school had the power to decide whether all of the rewards would be given to a small subset of teachers with the highest value-added, whether the winners of the rewards would be decided by lottery, and virtually anything in-between. The only restriction was that schools were not allowed to distribute rewards based on seniority. Theoretically, it is unclear how to design optimal teacher incentives when the objective is to improve student achievement. Much depends on the characteristics of the education production function. If, for instance, the production function is additively separable, then individual incentives may dominate group incentives, as the latter encourages free-riding. If, however, the production function has important complementarities between teachers in the production of student achievement, group incentives may be more effective at increasing achievement (Baker, 2002).

To our surprise, an overwhelming majority of the schools decided on a group incentive scheme that varied the individual bonus amount only by the position held in the school. This could be because teachers have superior knowledge of education production and believe the production function to have important complementarities, because they feared retribution from other teachers if they supported individual rewards, or simply because this was as close to pay based on seniority (the UFT's official view) that they could do.

The results from our incentive experiments are informative. Providing incentives to teachers based on school's performance on metrics involving student achievement, improvement, and the learning environment did not increase student achievement in any statistically meaningful way. If anything, student achievement declined. Intent-to-treat estimates yield treatment effects of -0.015 (0.024) standard deviations (hereafter σ) in mathematics and -0.011σ (0.020) in reading for elementary schools, and -0.048σ (0.017) in math and -0.032σ (0.011) in reading for middle schools, *per year*. Thus, if an elementary school student attended schools that implemented the teacher incentive program for three years, her test scores would decline by -0.045σ in math and by -0.033σ in reading - neither of which is statistically significant. For middle school students, however, the negative impacts are more sizeable: -0.144σ in math and -0.096σ in reading over a three-year period.

The impact of teacher incentives on student attendance, behavioral incidences, and alternative achievement outcomes such as predictive state assessments, course grades, Regents exam scores, and high school graduation rates are all negligible. Furthermore, we find no evidence that teacher incentives affect teacher behavior, measured by retention in district or in school, number of per-

sonal absences, and teacher responses to the learning environment survey, which partly determined whether a school received the performance bonus.

We also investigate the treatment effects across a range of subsamples – gender, race, previous-year achievement, previous-year teacher value added, previous-year teacher salary, and school size – and find that although some subgroups seem to be affected differently by the program, none of the estimates of the treatment effect are positive and significant if one adjusts for multiple hypothesis testing. The coefficients range from -0.264σ (0.074), in global history for white high school students, to 0.120σ (0.094), in math state exam scores for white elementary school students.

The paper concludes with a (necessarily) speculative discussion about what can explain the stark results, especially when one compares them with the growing evidence from developing countries. One explanation is that incentives are simply not effective in American public schools. This could be due to a variety of reasons, including differential teacher characteristics, teacher training, or effort. We argue that a more likely explanation is that all incentive schemes piloted thus far in the US, due in part to strong influence by teacher's unions, have been more complex and provided teachers with less agency than incentive experiments in developing countries. This ambiguity and lack of agency in American incentive schemes, relative to those attempted in developing countries, may explain our results. Other explanations such as the incentives were not large enough, group-based incentives are ineffective, or teachers are ignorant of the production function all contradict the data in important ways.

The next section provides a brief review of the emerging literature on the effect of teacher incentives on student achievement. Section 3 provides details of the experiment and its implementation. Section 4 describes the data and research design used in the analysis. Section 5 presents estimates of the impact of teacher incentives on a host of student and teacher outcomes. The final section concludes. There is an online data appendix that provides details on how we construct our covariates and our sample from the school district administrative files used in our analysis.

2 A Brief Literature Review

There is a nascent but growing body of literature on the role of teacher incentives on student performance (Glazerman et al., 2009; Glewwe et al., 2010; Lavy, 2002; Lavy, 2009; Muralidharan and Sundararaman, forthcoming; Springer et al., 2010; Vigdor, 2008.), including an emerging literature on the optimal design of such incentives (Neal, 2011). There are four papers, three

of them outside the US, which provide experimental estimates of the causal impact of teacher incentives on student achievement: Duflo and Hanna (2005), Glewwe et al. (2010), Muralidharan and Sundararaman (forthcoming), and Springer et al. (2010).

Duflo and Hanna (2005) randomly sampled 60 schools in rural India, and provided them with financial incentives to reduce absenteeism. The incentive scheme was simple; teachers' pay was linear in their attendance, at the rate of Rs 50 per day, after the first 10 days of each month. They found that teacher absence rate was significantly lower in treatment schools (22 percent) compared to control schools (42 percent), and that student achievement in treatment schools were 0.17σ higher than in control schools.

Glewwe et al. (2010) report results from a randomized evaluation that provided teachers for grades 4 through 8 in Kenya with group incentives based on test scores and find that while test scores increased in program schools in the short run, students did not retain the gains after the incentive program ended. They interpret these results as being consistent with teachers expending effort towards short-term increases in test scores but not towards long-term learning.

Muralidharan and Sundararaman (forthcoming) investigate the effect of individual and group incentives in 300 schools in Andhra Pradesh, India and find that both group and individual incentives increased student achievement by 0.12σ in language and 0.16σ in math in the first year, both equally successful. In the second year, however, individual incentives are shown to be more effective with an average effect of 0.27σ across math and language performance, while group incentives had an average effect of 0.16σ.

Springer et al. (2010) evaluated a three-year pilot initiative on teacher incentives conducted in the Metropolitan Nashville School System from the 2006-07 school year through the 2008-09 school year. 296 middle school mathematics teachers who volunteered to participate in the program were randomly assigned to the treatment or the control group, and those assigned to the treatment group could earn up to $15,000 as a bonus if their students made gains in state mathematics test scores equivalent to the 95 percentile in the district. They were awarded $5,000 and $10,000 if their students made gains equivalent to the 80th and the 90th percentiles, respectively. Springer et al. (2010) found there was no significant treatment effect on student achievement and on measures of teachers' response such as teaching practices.[5]

[5]There are several non-experimental evaluations of teacher incentive programs in the US, all of which report non-significant impact of the program on student achievement. Glazerman et al. (2009), who report a non-significant effect of -0.04 standard deviations on student test scores for the Teacher Advancement Program in Chicago and Vigdor (2008), who reports a non-significant effect of the ABC School-wide Bonus Program in North Carolina. Outside the US, Lavy (2002, 2009) reports significant results for teacher incentive programs in Israel.

The contribution of our paper is three-fold. First, the incentive scheme allows for schools to choose how to allocate incentive payments. If schools have superior knowledge of their production function (relative to a social planner) or better knowledge about their staffs, this design is optimal. Second, our experiment is the largest on teacher incentives in American public schools by orders of magnitude and the incentive scheme is similar to those being implemented in school districts across the country. Third, our set of outcomes is expansive and includes information on student achievement, student behavior, teacher retention, and teacher effort.

3 Program Details

3.1 Overview

On October 17, 2007, New York City's Mayor, Schools Chancellor, and the President of the United Federation of Teachers (UFT) announced an initiative to provide teachers with financial incentives to improve student performance, attendance, and school culture. The initiative was conceived as a two-year pilot program in roughly 400 of the lowest performing public schools in NYC.[6] School performance was tied to metrics used to calculate NYC's school report card - a composite measure of school environment, student academic performance, and student academic progress. The design of the incentive scheme was left to the discretion of the school. There were three requirements: (1) incentives were not allowed to be distributed according to seniority; (2) schools had to create a compensation committee that consisted of the principal, a designee of the principal, and two UFT staff members; and (3) the committee's decision had to be unanimous. The committee had the responsibility of developing how incentives would be distributed to each teacher and other staff.

Below, we describe how schools were selected and the incentive scheme, and provide an overview of the distribution of incentive rewards to schools.

3.2 School Selection

Table 1 provides an accounting of how we selected our experimental sample. Eligible middle and high schools were selected based on the average proficiency ratings on 4th and 8th grade state tests, respectively. Eligible elementary schools were selected based on poverty rates, student demographic characteristics such as the percentage of English Language Learners and special education students.

[6] The pilot program did not expand to include more schools in the second and third years due to budget constraints, but all schools that completed the program in the first or second year were invited to participate again in the following years.

The NYC Department of Education identified 438 schools that met the above mentioned eligibility criteria. Of these schools, 34 were barred by the UFT for unknown reasons and 8 were District 75 (i.e. special education) schools. The remaining 396 comprise our experimental sample, among which 212 schools were randomly selected by the author and offered treatment.[7] In November 2007, schools in the treatment group were invited to participate in the program. To formally accept the offer, schools were required to have at least 55 percent of their active full-time staff represented by the UFT at the school to vote for the program.[8] Schools forwarded voting results through email to the DOE by late November. Of the 212 schools randomly chosen to receive treatment, 179 schools garnered enough votes to participate, 33 declined treatment.[9] To increase the number of schools eligible to participate, we added 21 schools off the wait list; 19 garnered the requisite votes. So, overall, 233 schools in our experimental sample were invited to participate in the program, and 198 schools actually participated. The final experimental sample in year one consists of the lottery sample, with 233 treatment schools and 163 control schools.[10]

In the second year, 195 schools out of the 198 schools that received treatment in the first year were invited to participate in the second year pilot program (the other three schools were closed because of low performance). Of the 195 schools offered treatment, 191 schools voted to participate in the second year. In the third year of treatment, 191 schools that received treatment in the second year were invited to participate; 189 schools voted to participate in the program.

3.3 Incentive Scheme

Figure 2 shows how the progress report card score, which is the basis for awarding incentives, is calculated. Environment, which accounts for 15 percent of the progress report card score, is derived from attendance rate (5 percent of the overall score) and learning environment surveys administered to students, teachers, and parents in the spring semester (10 percent). Attendance rate is a school's average daily attendance. Student performance (25 percent) depends on the percentage of students

[7]There were 34 schools that met the eligibility criterion, but were excluded by the UFT for unknown reasons.

[8]Repeated attempts to convince the DOE and the UFT to allow schools to opt-in to the experimental group before random assignment were unsuccessful.

[9]Anecdotal evidence suggests that schools declined treatment for a variety of reasons, including fear that more work (not outlined in the agreement) would be required for bonuses. As one teacher in a focus group put it, "money ain't free."Furthermore, some teachers in focus groups expressed resentment that anyone would believe that teachers would be motivated by money.

[10]There were 187 elementary schools, 82 middle schools, 39 K-8 schools, 73 high schools, 1 K-12 school, and 14 schools that served both middle and high school students in the sample. 68 schools in our experimental sample also participated in Fryer's (2010) student incentive program in the 2007-08 and 2008-09 academic years. Excluding these schools from the sample did not change the qualitative results.

at grade level and the median proficiency ratings in ELA and math state tests for elementary and middle schools, and the 4-year and 6-year graduation rates and diploma-weighted graduation rates for high schools.[11] Student progress, which accounts for 60 percent of the overall score, depends on the average changes in proficiency ratings among students and the percentage of students making at least a year of progress in state tests for elementary and middle schools. Student progress in high schools is measured by the percentage of students earning more than 10 credits and the Regents exam pass rates in the core subjects - English, math, science, United States history, and global history. Schools can also earn extra credit points by exemplary gains in proficiency ratings or credit accumulation and graduation rates among high-need students such as English Language Learners, special education students, or those in the lowest tercile in ELA and math test scores citywide.

In each of the three categories, learning environment, student performance and student progress, schools were evaluated by their relative performance in each metric compared to their peer schools and all schools in the city, with performance relative to peer schools weighted three times of the weight given to performance relative to all schools citywide. However, because it is calculated using many metrics and because scores in each metric are calculated relative to other schools, how much effort is needed to raise the Progress Report card score by, say, one point is not obvious.

The table below shows the number of points by which schools had to increase their progress report card scores in the 2007-08 academic year in order to be considered to have met their goal and receive their incentive payment. The table illustrates that the target depends on the citywide ranking based on the previous year's Progress Report card score. If, for example, an elementary school was ranked at the 20th percentile in the 2006-07 academic year, it needed to increase its progress report card score by 15 points to meet the annual target.

[11] The DOE awards different levels of diplomas - Local, Regents, Advanced Regents, and Advanced Regents with Honors - depending on the number of Regents exams passed. The more Regents exams a student has to take to obtain a diploma the more weight it was given. In the 2007-08 academic year, the weights given to Local, Regents, Advanced Regents, and Advanced Regents diplomas were 1.0, 2.0, 2.5, and 3.0, respectively. In order to graduate with a Local diploma, students who entered high school in September, 2004 had to receive a total of 44 credits and score 55 or above in comprehensive English, mathematics A, global history and geography, United States history and government, and any science Regents exams. To graduate with a Regents diploma, students had to score 65 or above in the same five required Regents exam areas. To graduate with an Advanced Regents diploma, students had to score 65 or above in the already mentioned Regents exam areas and in addition, in mathematics B, life science, physical science, and foreign langauge Regents exams. Further details on graduation requirement and progress report card score calculation can be found on the DOE website.

Progress Report Target Points

Citywide Ranking Based on the Previous Year	Elementary & Middle	High
$\geq$ 85th percentile	7.5	2
$\geq$ 45 and $<$ 85	12.5	3
$\geq$ 15 and $<$ 45	15	4
$\geq$ 5 and $<$ 15	17.5	6
$<$ 5th percentile	20	8

Notes: Numbers calculated by the author.

A. An Example

Consider the following simplified example with an elementary school that ranks at about the 10th percentile citywide, and at about the 25th percentile among its peer schools. This school would have to increase its total progress report card scores by 17.5 points to meet the annual target. Let's now assume that the school increased the attendance rate to be about the 30th percentile citywide and the 75th percentile in the peer group. Then, holding everything constant, the school will increase the overall score by 1 point. Similarly, if the school increased their performance to the same level, the school will increase its score by 5 points. If student progress increased to the same level, its progress report card score will increase by 12 points. Hence, if the peer group and district schools stay at the same level, a low-performing school would be able to meet the annual target only if it dramatically increased its performance in all of the subareas represented in the progress report. On the other hand, because all scores are calculated relative to other schools, some schools can reach their incentive targets if their achievement stays constant and their peer schools underperform in a given year.

B. A Brief Comparison With Other School District Incentive Schemes

Most school districts that have implemented performance pay use similar metrics to NYC to measure teacher's performance. For example, the Teacher Advancement Program (TAP) in Chicago – started by Arne Duncan and described in the quote at the beginning of this paper – rewarded teachers based on classroom observations (25%) and school-wide student growth on Illinois state exams (75%). Houston's ASPIRE program uses school value added and teacher value added in state

exams to reward the top 25% and 50% of teachers. Alaska's Public School Performance Incentive Program divides student achievement into six categories and rewards teachers based on the average movement up to higher categories. Florida's S.T.A.R used a similar approach.

A key difference between the incentive schemes piloted in America thus far and those piloted in developing countries is that those in America compare teachers' or schools' performance to the distribution in the district. That is, teachers are not rewarded unless the entire school satisfies a criterion or their performance is in the top X percentile of their district, despite how well any individual or group of teachers performs. NYC's design rewards teachers based only on school's overall performance. A teacher participating in Houston's ASPIRE program would be rewarded the pre-determined bonus amount only if his teacher value added in one subject is in the top 25% of the district, regardless of how he or his school performs. Chicago's TAP program rewards teachers similarly. This ambiguity – the likelihood of receiving an incentive depends on my effort and the effort of others – may have served to flatten the function that maps effort into output.

3.4 Incentive Distribution

The lump-sum performance bonus awarded to a school was distributed to teachers in whatever way the school's compensation committee decided. Recall that the compensation committee consisted of the principal, a designee of the principal, and two UFT staff members. The committee was not allowed to vary the bonus by seniority, but could differentiate the compensation amount by the position held at school, by the magnitude of contribution made (e.g., teacher value added), or could distribute the bonus amount equally. The committee was chosen by December of the first year, and the committee reported to the UFT and the DOE their decision on how to distribute the bonus.

School bonus results were announced in September of the following year for elementary, K-8 and middle schools and in November for high schools, shortly after the DOE released Progress Report cards. Rewards were distributed to teachers either by check or as an addition to their salary, in accordance with the distribution rule decided upon by the compensation committee. In the first year, 104 schools out of 198 schools that participated met the improvement target and received the full bonus, while 18 schools met at least 75 percent of the target and received half of the maximum incentive payment possible. In total, the compensation received by participating schools totaled $22 million. In the second year, 154 schools out of 191 schools that participated received the full bonus, while 7 schools received half the maximum compensation. The total compensation awarded to schools in the second year was $31 million. We do not have precise numbers for year three, but

the DOE claims that the total costs of the experiment was approximately $75 million.

Figure 3A shows the distribution of individual compensation in the experiment. Most teachers in the schools that received the full bonus of $3,000 per staff were rewarded an amount close to $3,000. Figure 3B presents a histogram of the fraction of teachers receiving the same amount in each school in order to characterize how many schools decided upon an egalitarian distribution rule. More than 80% of schools chose to reward the same bonus amount to at least 85% of the teaching staff each year.

4 Data and Research Design

We combined data from two sources: student-level administrative data on approximately 1.1 million students across the five boroughs of the NYC metropolitan area from 2006-2007 to 2009-2010 school year, and teacher-level human resources data on approximately 96,000 elementary and middle school teachers during this same time period. The student level data include information on student race, gender, free- and reduced-price lunch eligibility, behavior, attendance, matriculation with course grades, and state math and ELA test scores for students in grades three through eight. For high school students, our data contain Regents exam scores and graduation rates. Data on attendance and behavioral incidences are available for all students.

Our main outcome variable is an achievement test unique to New York. The state ELA and math tests, developed by McGraw-Hill, are high-stake exams administered to students in the third through the eighth grade. Students in third, fifth, and seventh grades must score at level 2 or above (out of 4) on both math and ELA tests to advance to the next grade without attending summer school. The math test includes questions on number sense and operations, algebra, geometry, measurement, and statistics. Tests in the earlier grades emphasize more basic content such as number sense and operations, while later tests focus on advanced topics such as algebra and geometry. The ELA test is designed to assess students on three learning standards - information and understanding, literary response and expression, critical analysis and evaluation, and includes multiple-choice and short-response sections based on a reading and listening section, along with a brief editing task.

All public-school students are required to take the math and ELA tests unless they are medically excused or have a severe disability. Students with moderate disabilities or limited English proficiency must take both tests, but may be granted special accommodations (additional time,

translation services, and so on) at the discretion of school or state administrators. In our analysis, test scores are normalized to have a mean of zero and a standard deviation of one for each grade and year across the entire New York City sample.

We construct measures of attendance, behavioral problems and GPA using the NYC DOE data. Attendance is measured as the number of days present divided by the number of days present plus the number of days absent.[12] Behavioral problems are measured as the total number of behavioral incidences in record each year. GPA is measured as the mean course grade each year, calculated at a 1-4 scale for elementary school students and a 1-100 scale for middle school students. Attendance, behavioral problems and GPA were normalized to have mean zero and standard deviation one by grade level each year in the full New York City sample.

We use a parsimonious set of controls to aid in precision and to correct for any potential imbalance between treatment and control groups. The most important controls are achievement test scores from previous years, which we include in all regressions. Previous year's test scores are available for most students who were in the district in the previous year.[13] We also include an indicator variable that takes on the value of one if a student is missing a test score from a previous year and zero otherwise.

Other individual-level controls include a mutually exclusive and collectively exhaustive set of race dummies, indicators for free lunch eligibility, special education status, and English language learner status. A student is income-eligible for free lunch if her family income is below 130 percent of the federal poverty guidelines, or categorically eligible if (1) the student's household receives assistance under the Food Stamp Program, the Food Distribution Program on Indian Reservations (FDPIR), or the Temporary Assistance for Needy Families Program (TANF); (2) the student was enrolled in Head Start on the basis of meeting that program's low-income criteria; (3) the student is homeless; (4) the student is a migrant child; or (5) the student is a runaway child receiving assistance from a program under the Runaway and Homeless Youth Act and is identified by the local educational liaison.

Special education status is determined through a series of observations, interviews, reviews of report cards and administration of tests after the initial referral by teachers or parents. Student classified as in need of special education drafts an Individualized Education Program (IEP) with teachers and special staff and follows it while integrating the general curriculum as much as possible.

[12]The DOE does not collect absence data from schools after the first 180 days, so attendance rate calculated is the rate in the first 180 days.

[13]See Table 2 for exact percentages of experimental group students with valid test scores from previous years.

English Language Learners are those who speak a language other than English at home and score below proficient on English assessments when they enter the school system. They receive support through bilingual programs and English as a Second Language (ESL) programs.

We include other measures of academic achievement and behavior problems as controls. The squared previous year's state test scores are included in the parsimonious set of controls, as well as the total number of behavioral incidences recorded in 2006-2007 school year.

We also construct school level controls. To do this, we assign each student who was present at the beginning of the year, i.e., in September, to the first school that they attended. We construct the school-level variables based on these school assignments by taking the mean value of student demographic variables and student test scores for each school. Variables constructed this way include percentage of black, Hispanic, special education, limited English proficiency, and free-lunch eligible students. Also constructed is the total number of behavioral incidences in a school in the 2006-07 academic year.

We construct teacher-level variables from NYC Human Resources (HR) records and Teacher Value Added data. Teacher gender and race are constructed by taking the most recent non-missing records from 2004 to 2010 HR records. Teacher experience, or years of experience as a teacher, is taken from the October, 2007 HR file. Teacher Value Added (TVA) data are available from the 2006-2007 academic year until the 2008-2009 academic year. We take the TVA measured in the standard deviation unit, and standardize the number by grade level each year to have mean zero and standard deviation zero in the full city sample. For teachers who taught more than one grade, we take the average of TVA across grade levels. In addition, we construct the cumulative teacher absences in May of each academic year.

Table 2 provides pre-treatment descriptive statistics. The first four columns show the mean and standard deviation of student and teacher characteristics in all schools in the NYC district, the experimental sample, the treatment group, and the control group. In addition, the last two columns show the p-value of the difference between the mean of the entire district and that of the experimental sample and the p-value of the difference between the treatment group and the control group. The table of summary statistics shows that most student and teacher characteristics are balanced between the treatment and the control group. The only exceptions are the percentage of white teacher, the percentage of Asian teachers, and teacher-value-added in math in the 2006-2007 academic year.

Research Design

The simplest and most direct test of any teacher incentive program would be to examine the outcome of interest (e.g., test scores) regressed on an indicator for enrollment in the teacher incentive program for grades g, in school s, in year t ($incentive_{i,g,s,t}$) and controls for basic student and school characteristics, X_i and X_s, respectively:

$$outcome_{i,g,s,t} = \alpha_1 + \beta_1 X_i + \gamma_1 X_s + \delta_g + \zeta_t + \pi_1 incentive_{i,g,s,t} + \varepsilon_{i,g,s,t}.$$

Yet, if schools select into teacher incentive programs because of important unobserved determinants of academic outcomes, estimates obtained using the above equation may be biased. To confidently identify the causal impact of incentive programs, we must compare participating and non-participating schools which would have had the same academic outcomes had they both participated in the program. By definition, this involves an unobservable counterfactual.

In the forthcoming analysis, the counterfactual is constructed by exploiting the random assignment of schools into treatment and control groups. Restricting our analysis to schools that were selected (by the UFT and the DOE) to be included in the experimental sample, we can estimate the causal impact of being offered a chance to participate in an teacher incentive program by comparing the average outcomes of schools randomly selected for treatment and the average outcomes of schools randomly selected for control. Schools that were not chosen to participate form the control group corresponding to the counterfactual state that would have occurred to treatment schools if they had not been offered a chance to participate.

Let T_s be an indicator for a treatment school. The mean difference in outcomes between treatment schools ($T_s = 1$) and control schools ($T_s = 0$) is known as the "Intent-to-Treat" (ITT) effect, and is estimated by regressing student outcomes on T_s. In theory, predetermined student school characteristics (X_i and X_s) should have the same distribution across treatment and control groups because they are statistically independent of treatment assignment. In small samples, however, more precise estimates of the ITT can often be found by controlling for these characteristics. The specifications estimated are of the form:

$$outcome_{i,g,s,t} = \alpha_2 + \beta_2 X_i + \gamma_2 X_s + \pi_2 T_s + \delta_g + \zeta_t + \varepsilon_{i,s},$$

where our vector of student level controls, X_i, includes an indicator for gender, a mutually inclusive and collectively exhaustive set of race dummies, an indicators for free-lunch eligibility, special

education status, and English language learner status, separately, and pre-determined measures of the outcome variables when possible (i.e., pre-intervention test scores and the number of behavioral incidences). The set of school-level controls, X_s, includes the percentages of students at school who are black, Hispanic, free-lunch eligible, English language learners, and special education students, and the pre-lottery number of behavioral incidences at school. The ITT is an average of the causal effects for students enrolled in treatment schools compared to those enrolled in control schools, at the time of random assignment. The ITT therefore captures the causal effect of being offered a chance of participating in the incentive program, not of actually participating.

Under several assumptions (that the treatment group assignment is random, control schools are not allowed to participate in the incentive program, and treatment assignment only affects outcomes through school enrollment), we can also estimate the causal impact of actually participating in the incentive program. This parameter, commonly known as the "Treatment-on-Treated" (TOT) effect, measures the average effect of treatment on schools that choose to participate in the merit pay program. The TOT parameter can be estimated through a two-stage least squares regression of student outcomes on participation, with original treatment assignment (T_s) as an instrumental variable for participation. We use the number of years a student spent in treated schools as the actual participation variable. The first stage equations for IV estimation take the form:

$$incentive_{i,g,s,t} = \alpha_3 + \beta_3 X_i + \gamma_3 X_s + \delta_g + \zeta_t + \pi_3 T_s + \varepsilon_{i,s,g,t},$$

where π_3 captures the effect of treatment assignment (T_s) on the average number of years a student spends in a treatment school.

The TOT is the estimated difference in outcomes between students in schools who were induced into participating through treatment assignment and those in the control group who would have enrolled if they had been offered the chance.

5 The Impact of Teacher Incentives

5.1 Student Achievement

Table 3 presents first-stage, ITT, and TOT estimates of the effect of teacher incentives on state math and ELA test scores. Columns one through three report estimates from our elementary school sample, columns four through six report estimates from middle schools, and the final three

columns present results for a pooled sample of elementary and middle schools. We present both raw estimates and those that contain the parsimonious set of controls described in the previous section. Note: the coefficients in the table are normalized so that they are in standard deviation units and represent one year impacts.

Surprisingly, all estimates of the effect of teacher incentives on student achievement are negative in both elementary and middle school – and statistically significant so in middle school. The ITT effect of the teacher incentive scheme is -0.011σ (0.020) in reading and -0.015σ (0.024) in math for elementary schools, and -0.032σ (0.011) in reading and -0.048σ (0.017) in math for middle schools. The effect sizes in middle school are non-trivial – a student who attends a participating middle school for three years of our experiment is expected to lose 0.096σ in reading and 0.144σ in math. The TOT estimates are smaller than the ITT estimates, as the first stage coefficients are all larger than one.

Table 4 presents results similar to table 3, but for high schools. High school students do not take the New York state exams. Instead, they have to take and score 55 or above in Regents exams in five key subject areas to graduate with a local diploma. To graduate with a Regents diploma or an Advanced Regents diploma, students have to score 65 or above in more subject areas. For example, students who entered high school in September 2005 had to score 65 or above in comprehensive English, integrated algebra, global history and geography, U.S. history and government, and science Regents exams to graduate with a Regents diploma. [14] Table 4 presents first-stage, ITT, and TOT estimates for the impact of teacher incentives on comprehensive English, mathematics, science, U.S. history, and global history Regents exam scores. All exam scores were standardized to have mean zero and standard deviation one in the full city sample each academic year.

Similar to the analysis of elementary and middle schools, there is no evidence that teacher incentives had a positive effect on achievement. Estimates of the effect of teacher incentives on high school achievement are all small and statistically non-significant. The ITT effect on the English Regents exam score is -0.003σ (0.044), the effect on the integrated algebra exam score is -0.011σ (0.031), and the effect on science scores is -0.016σ (0.037). The ITT effect on U.S. history exam score is -0.033σ (0.054) and that on global history exam score is -0.063σ (0.045). The TOT

[14] Regents exams are offered in January, June, and August of each academic year in the following subject areas: comprehensive English, algebra, geometry, trigonometry, chemistry, physics, biology, living environment, earth science, world history, U.S. history, and foreign languages. In this paper, we present results on comprehensive English, integrated algebra, living environment, U.S. history, and global history Regents exam scores. Among mathematics and science exam areas, integrated algebra and living environment were selected because the highest number of students took those exams. Using other exam scores gives qualitatively similar results.

effect is of a comparable magnitude.

The bottom panel of table 4 reports treatment effects on four-year graduation rates. The dependent variables are a dummy for graduating in 4 years, which takes the value one if student graduated in 4 years and zero otherwise, and a dummy for graduating in 4 years with a Regents diploma, which takes the value one if student graduated with a Regents diploma and zero otherwise. Students enrolled in treatment schools were 4.4 percent less likely to graduate in four years (which is statistically significant at 5% level) and were 7.4 percent less likely to obtain a Regent's diploma (statistically significant at 10% level). Note: during the period of the experiment, mean graduation rates fluctuated between 54% and 61%.

Table 5 explores heterogeneity in treatment effects across a variety of subsamples of the data: gender, race, free-lunch eligibility, previous years student test scores, school size, teacher value-added, and teacher salary. The coefficients in the table are ITT estimates with our parsimonious set of controls. All categories are mutually exclusive and collectively exhaustive. The effect of teacher incentives on achievement does not vary systematically across the subsamples. The only exceptions are among middle school students who are free-lunch eligible, students who are attending larger schools, and those taught by more experienced teachers (i.e., received higher salaries), and among high school students who are white or Asian, students who scored lower in 8th grade state tests, and students who are attending larger schools. Students in these subsamples seem to be affected more negatively by teacher incentives.

The estimates above use the sample of students for which I have achievement test scores. If students in treatment and control schools have different rates of selection into this sample, my results may be biased. A simple test for selection bias is to investigate the impact of the treatment offer on the probability of entering the sample. The results of this test, though not shown here in tabular form, demonstrate that the coefficient on treatment is small and statistically zero.[15] This suggests that differential attrition is not likely to be a concern in interpreting the results.

5.2 Alternative Outcomes

Thus far, we have concentrated on student progress on state assessments, the most heavily weighted element of NYC's incentive scheme. Now, we introduce two additional measures of student performance and three measures of school environment: grade point averages, predictive math and ELA exams, school environment surveys, attendance and behavioral incidences. Many of these outcomes

[15]Tabular results are available from the author upon request.

enter directly into the incentive scheme and may be affected by it.

Table 6 shows estimates of the impact of teacher incentives on this set of alternative outcomes. Predictive assessments are highly correlated with the state exams and are administered to all public school students in grades three through eight in October and May. The DOE gives several different types of predictive exams, and schools can choose to use one of the options depending on their needs. In this paper, we analyze math and ELA test scores from the spring Acuity Predictive Assessment.[16] Each student's attendance rate is calculated as the total number of days present in any school divided by the total number of days enrolled in any school. Attendance rate was standardized by grade level to have mean zero and standard deviation one each academic year across the full city sample. Grades were extracted from files containing the transcripts of all students in the district.[17] Elementary school students received letter grades, which were converted to a 4.0 scale, and middle and high school students received numeric grades that ranged from 1 to 100. Students' grades from each academic year were averaged to yield an annual GPA. As with test scores, GPAs were standardized to have mean of zero and standard deviation of one among students in the same grade with the same grade scale across the school district. Number of behavioral incidences were pulled from behavior data, which record the date, level, location and short description of all incidences. The total number of incidences attributed to a student in an academic year across all schools and grades he attended was calculated, and standardized by grade level to have mean zero and standard deviation one each academic year across the full city sample.

Results from predictive assessments provide an identical portrait to that depicted by state test scores. The effect of the teacher incentive program on predictive ELA exams is negative and statistically insignificant, with the ITT effect equal to -0.019σ (0.016) in the elementary school sample and -0.022σ (0.018) in the middle school sample. The ITT effect on predictive math exams is -0.023σ (0.020) in the elementary school sample and -0.051σ (0.022) in the middle school sample. Note that the effect of teacher incentives on middle school students' predictive math exam scores is negative and statistically significant, consistent with the findings with state test scores.

Teacher incentives have a statistically insignificant effect on other alternative student outcomes. The ITT and TOT effects on attendance rate, which enters directly in the calculation of progress report card scores are negative across all school levels. The ITT effect is estimated to be -0.018σ

[16]Eighth grade students did not take the spring predictive tests, because they did not have to take state exams in the following year.

[17]Elementary school transcripts are not available for all schools each academic year. High school transcripts were not availble until the 2008-09 academic year.

(0.020) in the elementary school sample, -0.019σ (0.022) in the middle school sample, and -0.014σ (0.050) in the high school sample. The effects on behavioral incidences and grade point averages are similarly small and insignificant.

5.3 Teacher Behavior

In this section, we estimate the impact of the teacher incentive program on two important teacher behaviors: absences and retention. We assign teachers to treatment or control groups if they were assigned to a treatment or a control school, respectively, in October of 2007. We only include teachers who were teaching at schools in the randomization sample in 2007, and ignore all who enter the system afterwards.

We measure retention in two ways: in school and in district – both of which were constructed using Human Resources data provided by DOE. Retention in school was constructed as a dummy variable that takes the value one if a teacher was associated with the same school in the following academic year, and zero otherwise. Retention in district is more complicated. Like our coding of retention in school, we construct a dummy variable that takes the value one if a teacher was found in the New York City school district's Human Resources (HR) file in the following academic year, and zero otherwise. But there are two important caveats. First, charter schools and high schools are not included in the New York City public school district's HR files and therefore, some teachers who left the district may have simply moved to teach at charter schools or high schools in the district. As the same types of teacher certificates qualify teachers to teach in both middle and high schools, it is possible that some teachers who left the district from middle schools went to teach at high schools. It is unlikely, however, that a significant number of elementary school teachers obtained new certificates to qualify for teaching in middle schools. Therefore, we divided the sample of teachers into elementary, middle and K-8 school samples and estimate the treatment effects separately on each sample. To measure absences, we were given the number of personal absences as of May for teachers who did not exit the system.

Table 7 presents results on the impact of teacher incentives on our measures of teacher behavior. There is no evidence that teacher incentives affect retention in either district or school, or teacher absences. Elementary school teachers in treatment schools were 0.2 percent more likely to stay in the NYC school district, 0.7 percent less likely to stay at the same school in the following academic year, and took 0.275 more days of personal absences. Middle school teachers exhibit similar patterns. None of these effects are statistically significant, nor are they economically meaningful.

6 Discussion

The previous sections demonstrate that the teacher incentive scheme piloted in 200 New York City public schools did not increase achievement. If anything, achievement may have declined as a result of the experiment. Yet, incentive schemes in developing countries have proven successful at increasing achievement.

In this section, we consider four explanations for these stark differences: (1) incentives may not have been large enough; (2) the incentive scheme was too complex; (3) group-based incentives may not be effective; and (4) teachers may not know how they can improve student performance. Using our analysis, along with data gleaned from other experiments, we argue that the most likely explanation is that the NYC incentive scheme, along with all other American pilot initiatives thus far, is too complex and provides teachers with too little agency. It is important to note that we cannot rule out the possibility that other unobservable differences between the developing countries and America (e.g. teacher motivation) produce the differences.

Incentives Were Not Large Enough

One potential explanation for our stark results is that the incentives simply were not large enough. There are two reasons that the incentives to increase achievement in NYC may have been small. First, although schools had discretion over how to distribute the incentives to teachers if they met their performance targets, an overwhelming majority of them chose to pay teachers equally. These types of egalitarian distribution methods can induce free-riding and undercut individual incentives to put in effort. Moreover, an overwhelming majority of teachers in schools that met the annual target earned am amount close to $3000. This is less than 4.1 percent of the average annual teacher salary in the sample. One might think that the bonus was simply not large enough for teachers to put in more effort, though similar incentive schemes in India (3%) and Kenya (2%) were relatively smaller.

Second, the measures used to calculate the progress report card scores directly influence other accountability measures such as the AYP (Adequate Yearly Progress) that determine whether a school will be subjected to regulations or even be closed, which results in all staff losing their jobs. Hence, all poor performing schools, including all treatment and control schools in our experiment, have incentives to perform well on the precise measures that were being incentivized. Thus, it is not clear whether the teacher incentive program provides additional incentives, at the margin, for teachers to behave differently.

A brief look at the results of the Project on Incentives in Teaching (POINT), a pilot initiative in Nashville, Tennessee, suggests that a larger incentive in schools which are not under pressure by AYP was still not any more effective. Teachers in POINT treatment schools were selected from the entire school district and could earn up to $15,000 in a year based solely on their students' test scores. Teachers whose performance was at lower thresholds could earn $5,000 to $10,000. The maximum amount is roughly 22% of the average teacher salary in Nashville. Springer et al. (2010) find that even though about half of the participating teachers could have reached the lowest bonus threshold if their students answered on average 2 or 3 more items out of 55 items correctly, student achievement did not increase significantly more in classrooms taught by treatment teachers. Moreover, they report that treatment teachers did not seem to change their instructional practices or effort level.

Incentive Scheme Was Too Complex

In our experiment it was difficult, if not impossible, for teachers to know how much effort they should exert or how that effort influences student achievement because of the complexity of the progress report card system used in NYC. For example, the performance score for elementary and middle schools is calculated using the percentage of students at proficiency level and the median proficiency rating in state tests. Recall, the performance score depends on how a school performs compared to its peer schools that had similar student achievement level in the previous year and compared to all schools in the district. But it is highly unlikely that teachers can predict at which percentile their school will be placed relative to the peer group and the district in these measures of performance if the school increased the overall student achievement by, for example, one standard deviation.

Moreover, the POINT pilot in Tennessee, like other American school districts, contained an incentive scheme that was dependent on the performance of others rather than simpler incentive schemes such as those in Duflo and Hanna (2005), Glewwe et al. (2010), and Muralidharan and Sundararaman (forthcoming). It is plausible that this ambiguity may have served to flatten the function that maps effort into expected reward.

Group-Based Rewards Are Ineffective

Although we gave schools the flexibility to choose their own incentive schemes, the vast majority of them settled on a group-based scheme. Group-based incentive schemes introduce the potential for free-riding and may be ineffective under certain conditions. Yet, in some contexts, they have

been shown to be effective. For example, Muralidharan and Sundararaman (forthcoming) found that the group incentive scheme in government-run schools in India had a positive and significant effect on student achievement. However, the authors stress that 92% of treatment schools had between two and five teachers. The average number of teachers in a treated school was 3.28. Similar results are obtained in Glewwe et al. (2010) where the average number of teachers per school was twelve. Provided that New York City public schools have 60 teachers on average, the applicability of the results from these analyses is suspect. When there are only 3 (or 12) teachers in a school, monitoring and imposing cost on those teachers who shark their responsibility is less costly.

On the other hand, Lavy (2002) also suggests that group-based incentives may be effective in larger schools. His non-experimental evaluation of the teacher incentives intervention in Israel, in which teachers were incentivized on the average number of credit units per student, the proportion of students receiving a matriculation certificate, and the dropout rate, reveals that the program had a positive and significant impact on the average number of credits and test scores. The average number of teachers in the treatment schools in Israel is approximately 80, closer to the average number of teachers in a school in NYC.

Teachers are ignorant, not lazy

If teachers only have a vague idea of how they could increase student achievement, then there may be little incentive to increase effort. The most striking evidence against the hypothesis that our results are driven by teachers' lack of knowledge of the production function is driving our results is presented in table 8, which displays treatment effects on five areas of the teacher survey which partly determined 10 percent of the school's overall progress report score. As before, we present first stage, ITT, and TOT estimates for each dependent variable.

The first outcome is the teachers' response rate to the learning environment survey. The next four outcomes are the teacher's average responses to four areas of the survey questions: academic expectations, communication, engagement, and safety and respect. Questions in the academic expectations area measures how well a school develops rigorous academic goals for students. The communication area examines how well a school communicates its academic goals and requirements to the community. The engagement area measures the degree to which a school involves students, parents, and educators to promote learning. Questions in the safety and respect section asks whether a school provides a physically and emotionally secure learning environment. The scores

were standardized to have mean zero and standard deviation one by school level in the full city sample.

One might predict that teachers in the incentive program would be more likely to fill out the survey and give higher scores to their schools given that they can increase the probability of receiving the performance bonus by doing so. This requires no knowledge of the production function - just an understanding of the incentive scheme. Table 8 reveals that treatment teachers were not significantly more likely to fill out school surveys. The mean response rate at treatment schools was 64% in the 2007-2008 academic year and 76% in the 2008-2009 academic year. This may indicate that teachers did not even put in the minimum effort of filling out teacher surveys in order to earn the bonus.

References

[1] Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." Journal of Labor Economics, 25(1): 95-135.

[2] Angirst, Joshua D. and Kevin Lang. 2004. "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program." American Economic Review, 94(5): 1613-1634.

[3] Baker, George. 2002. "Distortion and Risk in Optimal Incentive Contracts." Journal of Human Resources, 37: 728-751.

[4] Baker, George, Robert Gibbons, and Kevin J. Murphy. "Subjective Performance Measures in Optimal Incentive Contracts." Quarterly Journal of Economics, 109(4): 1125-1156.

[5] Bettinger, Eric. 2010. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." NBER Working Paper No. 16333.

[6] Borman, Geoffrey D., Gina M. Hewes, Laura T. Overman, and Shelly Brown. 2003. "Comprehensive School Reform and Achievement: A Meta-Analysis." Review of Educational Research, 73(2): 125-230.

[7] Borman, Geoffrey D., Robert E. Slavin, Alan C.K. Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers. 2007. "Final Reading Outcomes of the National Randomized Field Trial of Success for All." American Educational Research Journal, 44(3): 701-731.

[8] Cook, Thomas D., Robert F. Murphy, and H. David Hunt. 2000. "Comer's School Development Program in Chicago: A Theory-Based Evaluation." American Educational Research

Journal, 37(2): 535-597.

[9] Corcoran, Sean P., William N. Evans, and Robert M. Schwab. 2004. "Changing Labor-Market Opportunities for Women and the Quality of Teachers, 1957-2000." American Economic Review, 94(2): 230-235.

[10] Cullen, Julie B., Brian A. Jacob, and Steven D. Levitt. 2005. "The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools." Journal of Public Economics, 89(5-6): 729-760.

[11] Currie, Janet and Duncan Thomas. 1995. "Does Head Start Make a Difference?" American Economic Review, 85(3): 341-364.

[12] Decker, Paul T., Daniel P. Mayer, and Steven Glazerman. 2004. "The Effects of Teach for America on Students: Findings from a National Evaluation." Mathematica Policy Research Report No. 8792-750.

[13] Duflo, Esther and Rema Hanna. 2005. "Monitoring Works: Getting Teachers to Come to School." NBER Working Paper No. 11880.

[14] Finn, Jeremy D., and Charles M. Achilles. 1999. "Tennessee's Class Size Study: Findings, Implications, Misconceptions." Educational Evaluation and Policy Analysis, 21(2): 97-109.

[15] Firestone, William A., and James R. Pennell. 1993. "Teacher Commitment, Working Conditions, and Differential Incentive Policies." Review of Educational Research, 63(4): 489-525.

[16] Fryer, Roland G. 2010. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." NBER Working Paper No. 15898.

[17] Glazerman, Steven, Allison McKie, and Nancy Carey. 2009. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year One Impact Report." Mathematica Policy Research, Inc.

[18] Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." American Economic Journal, 2(3): 205-227.

[19] Hanushek, Eric and Steven Rivkin. 2005. "Teachers, Schools and Academic Achievement." Econometrica, 73(2): 417-458.

[20] Hastings, Justine S., Thomas J. Kane, and Douglas Staiger. 2006. "Preferences and Heterogenous Treatment Effects in a Public School Choice Lottery. NBER Working Paper No. 12145.

[21] Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." Journal of Law, Economics, and Organization, 7: 24-52.

[22] Hoxby, Caroline M. and Andrew Leigh. 2004. "Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States." American Economic Review, 94(2): 236-240.

[23] Jacob, Brian A., and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis."' The Review of Economics and Statistics, 86(1): 226-244.

[24] Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." Quarterly Journal of Economics, 118(3): 843-877.

[25] Jepsen, Christopher, and Steven G. Rivkin. 2002. "What is the Tradeoff between Smaller Classes and Teacher Quality?" NBER Working Paper No. 9205.

[26] Jonson, Susan M. 1984. "Merit Pay for Teachers: A Poor Prescription for Reform." Harvard Education Review, 54(2): 175-185.

[27] Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Validation." NBER Working Paper No. 14607.

[28] Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City." Economics of Education Review, 27(6): 615-631.

[29] Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." Quarterly Journal of Economics, 114(2): 497-532.

[30] Krueger, Alan B., and Diane M. Whitmore. 2001. "The Effect of Attending a Small Class in Early Grades on College-test Taking and Middle School Test Results: Evidence from Project Star." Economic Journal, 111(468): 1-28.

[31] Krueger, Alan B., and Pei Zhu. 2004. "Another Look at the New York City School Voucher Experiment." American Behavioral Scientist, 47(5): 658-698.

[32] Lauer, Patricia A., Motoko Akiba, Stephanie B. Wilkerson, Helen S. Apthorp, David Snow, and Mya L. Martin-Glenn. 2006. "Out-of-School-Time Programs: A Meta-Analysis of Effects

for At-Risk Students." Review of Educational Research, 76(2): 275-313.

[33] Lavy, Victor. 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." The Journal of Political Economy, 110(6): 1286-1317.

[34] Lavy, Victor. 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics." American Economic Review, 99(5): 1979-2021.

[35] Muralidharan, Karthik and Venkatesh Sundararaman. 2011. "Teacher Performance Pay: Experimental Evidence from India." Journal of Political Economy (forthcoming).

[36] Neal, Derek 2011. "The Design of Performance Pay Systems in Education." NBER Working Paper No. 16710.

[37] Nye, Barbara, B. DeWayne Fulton, Jayne Boyd-Zaharias, and Van A.Cain. 1995. "The Lasting Benefits Study, Eighth Grade Technical Report." Nashville, TN: Center for Excellence for Research in Basic Skills, Tennessee State University.

[38] Olds, David L. 2006. "The Nure-family Partnership: An Evidence-based Preventive Intervention." Infant Mental Health Journal, 27(1): 5-25.

[39] Redd, Zakia, Stephanie Cochran, Elizabeth Hair, and Kristin Moore. 2002. Academic Achievement Programs and Youth Development: A Synthesis. Washington, DC: Child Trends.

[40] Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." Econometrica, 73(2): 417-458.

[41] Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." American Economic Review, 94(2): 247-252.

[42] Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2008. "Can You Recognize An Effective Teacher When You Recruit One?" NBER Working Paper 14485.

[43] Rouse, Cecilia E. 1998. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." Quarterly Journal of Economics, 113(2): 553-602.

[44] Snyder, Thomas D., and Sally A. Dillow. 2010. Digest of Education Statistics 2009 (NCES 2010-013). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

[45] Springer, Matthew G., Dale Ballou, Laura S. Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. "Teacher Pay for Performance:

Experimental Evidence from the Project on Incentives in Teaching." Conference paper, National Center on Performance Incentives.

[46] Vigdor, Jacob L. 2008. "Teacher Salary Bonuses in North Carolina." Conference paper, National Center on Performance Incentives.

7 Online Data Appendix

7.1 Student Level Variables

Assignment to School and Grade

A non-negligible number of students attended more than one schools and grades in each academic year. In order to ensure that results are not driven by students' self-selection into treatment schools, we assigned students to the first school and grade they were present in each academic year. If there are multiple schools and grades with the same attendance patterns, we assigned students to the school with the lowest alphanumeric order dbn and the the lowest grade.

Assignment to Treatment or Control

Students were assigned to treatment and control group if they attended treatment or control school, respectively, in the 2007-08 academic year. If students did not attend any school in New York City in September, they were not included in the sample. If there were more than one schools with the same attendance patterns, students were dropped from the sample.

Treatment Variable in TOT regressions

The fraction of an academic year spent in any treated school was calculated for each student, by taking the number of days spent in treated schools and dividing the number by the maximum number of days student could be present at a school. The treatment variable in TOT regressions was constructed by taking the cumulative years spent in treated schools. For example, if a student spent 0.9 year in treated schools in year 1, 0.8 in year 2, and 0.5 in year 3, his treatment variable would be 0.9 in year 1, 1.7 in year 2, and 2.2 in year 3.

Demographic Variables

DOE provided Edlabs with enrollment files that contained student sex, race, free lunch status, English language learner status, and special education status for academic years 2003-04 to 2009-10. Demographic variables that should be constant over time, such as sex and race, were constructed by taking the most recent non-missing records from the enrollment files. Other student characteristics - free lunch eligibility status, English language learner status, and special education

status - were constructed from each year's enrollment file. All demographic variables were coded as dummy variables that take the value one if student belongs to the demographic group and zero otherwise.

State Test Scores

NYC DOE administers state ELA and math assessment tests in January and March of each academic year to students in grades three through eight. All students in public schools must take the tests unless they are medically excused or have severe disabilities. Students with limited English proficiency or mild disabilities have to take the tests, but are granted special accomodations. ELA and math assessments are developed by McGraw-Hill and include multiple-choice sessions and short and extended response questions, as well as editing tasks in ELA in some grade levels. ELA assessments ask students to draw conclusions, compare and contrast information and ideas, interpret meaning and explain cause-and-effect relationships. Math assessments ask students to demonstrate the knowledge of and apply facts and definitions, read and interpret graphs and tables, and represent concepts with mathematical signs and symbols. Students in the third, fifth, and seventh grades must score proficient or above to advance to the next grade.

NYC DOE provided test score data that contained grade level, proficiency level, and scale test scores for each subject for academic years 2006-07 to 2009-10. We standardized scale scores to have mean zero and standard deviation one by grade level and by academic year over the full city sample. The 2006-07 academic year's ELA and math scale scores were included as control variables in all raw regressions, and the squared 2006-07 ELA and math scale scores were used in addition in all control regressions on the elementary and the middle school samples. The most recent 8th grade ELA and Math test scores taken before the 2007-08 academic year were constructed for high school students and similarly used as control variables in regressions on the high school sample.

Regents Exam Scores

New York State Department of Education requires high school students to pass a certain number of Regents examinations in core subjects to graduate and awards different levels of diplomas depending on the number of exams passed. For example, in order to earn a local diploma, a student who entered high school in 2007 would have to obtain 55 points or above in comprehensive English, mathematics, global history, U.S. history, and science Regents exams. To earn a Regents diploma, he would have to obtain 65 points or above in all subject areas mentioned above.

Regents exams are administered in five subject areas - ELA, math, science, foreign language, and social studies - three times every year in January, June, and August. Students can choose when to take each subject exam, and they can also choose to satisfy regents requirements with approved alternative assessments such as the SAT subject tests.

DOE provided Regents exam data for academic years 2007-08 to 2009-10. Each data file contained Regents exam codes and scores. There are multiple exams offered under each subject area. We restricted our attention to one exam per subject areas taken by the highest number of students. For example, we used integrated algebra scores among mathematics exams and living environment exam scores among science exams. We standardized exam scores to have mean zero and standard deviation one by exam codes over the full city sample. We calculated the average of all standardized scores of regents exams taken in the subject in a given academic year, and used the resulting value as the standardized Regents score.

In this paper, we use comprehensive English, mathematics, science, U.S. history, and global history Regents exam scores as high school outcome variables.

Predictive Exam Scores

Every NYC school uses periodic assessment to monitor students' progress towards the state learning standards. DOE provides several options including Acuity Predictive Assessments, Acuity Instructionally Targeted Assessments, Performance Series Computer Adaptive Assessments, etc. Schools choose from these options depending on their needs.

Of these period assessments, DOE provided Acuity Predictive Assessment results for all students who took the exams in 2007-08 and 2008-09. Acuity Predictive Assessments closely mirror state tests, and are designed to predict student performance level at state tests so that students and teachers can learn how much work is needed and where they should focus in order to pass the state learning standards. Acuity Predictive Assessments are administered to students in grades three through eight once in the fall, and once in the spring. Predictive ELA exams were administered in late October and late May and predictive math exams were administered in January and late May in the 2007-08 and 2008-09 academic years. We took percentage points from the data files, and standardized the scores by grade and by the test period to have mean zero and standard deviation one over the full city sample of students with valid scores.

Attendance Rate

DOE provided monthly attendance data files that record the number of days present and the number of days absent in each school and grade students were enrolled in. Attendance rate was calculated by first taking the sum of the number of days present across all schools and grades students attended in a year and dividing the number by the total number of days students were enrolled in any school and grade. Then, attendance rate was standardized by the assigned grade level to have mean zero and standard deviation one over the full city sample.

Behavioral Incidences

DOE provided incidence-level behavioral records of students for academic years 2006-07 to 2009-10. Records included information on the place, time and the severity level of incidences. We assumed that students who attended school in NYC in the academic year and were not found in the behavioral files did not have any behavioral incidence that year. We calculated the total number of behavioral incidences at all levels attributed to students each year. We standardized the number of incidences by assigned grade to have mean zero and standard deviation one over the full city sample.

We also constructed school-level total number of behavioral incidences as the total number of incidences committed at school each academic year. If school was open in the academic year and was not found in the behavioral files, it was assumed that there was no behavioral incidence in that school.

GPA

GPA was constructed from course grade data files that contained the grades received in all courses taken in each academic year. Elementary and middle school grades were available for all academic years from 2006-07 to 2009-10, and high school grades were available from 2008-09. Elementary school grades were calculated in the 1-4 scale, and middle and high school grades were calculated in the 1-100 scale. I dropped duplicate records that have the same course code and grade within an academic year, and calculated the yearly GPA as an average grade across all the courses taken that year. GPA was standardized to have mean zero and standard deviation one by grade and by academic year over the full city sample.

Graduation

DOE provided four-year graduation results for cohorts who entered high school in 2004 and

2005. Graduation status is coded as a dummy variable that takes the value one if student graduted in four years, and zero otherwise. We also constructed dummy variables for graduating with Regents diploma or higher for students who graduated high school in four years.

7.2 Teacher Level Variables

Demographic Variables

DOE provided Human Resources data, which contained teacher sex, race, date of birth, experience, and salary step, from academic years 2004-05 to 2009-10. Demographic variables, such as sex and race, were constructed and coded in the same way as student demographic variables. Teacher salary and years of teaching experience before the 2007-08 academic year were taken from the Human Resources data, and used as control variables in teacher level regressions, along with gender and race dummies.

Teacher Value Added

DOE provided multi-year teacher value added data, which contained TVA measured in standard deviation units in every classroom. Teacher value added measures the changes in student achievement attributable to classrooms controlling for a large set of student and classroom level variables[18]. We dropped TVA values in classrooms taught by more than one teachers or by teachers with blank IDs. TVA measures were standardized to have mean zero and standard deviation one by grade level over all classrooms in the city. If teachers taught multiple classrooms, the average of standardized teacher value added measures was taken across all classrooms taught in the academic year. The 2006-07 ELA and Math TVA values were used as control variables in all teacher level regressions.

Teacher Absences

DOE provided a supplement to teacher value added data containing teacher identifier, demographic characteristics such as gender, race, experience, and salary step, and personal or DOE-related cumulative absences in May of the academic year. We took the cumulative number of personal absences in May for 2007-08 and 2008-09 academic years. Teachers who left the system before May were set to have missing absences.

[18]For a detailed description of the model used to calculate teacher value added, please visit the NYC DOE's website.

7.3 School Level Variables

Student Characteristics

We constructed school-level measures of student characteristics by taking the average of student demographic dummy variables and test scores over students assigned to each school by the rule described previously. Among the variables constructed, the percentages of blacks, Hispanics, special education students, limited English proficiency students, and free-lunch eligible students were included in the parsimonious set of controls in student-level control regressions. The percentages of blacks, Hispanics, Asians, other race students, males, free-lunch eligible students, special education students, English language learners, and students missing information on gender, race, free-lunch eligibility or special education status in the enrollment data were used as control variables in school-level regressions, along with average 2006-07 ELA and math state test scores for elementary or middle schools and average 8th grade ELA and math test scores for high schools.

Teacher Characteristics

School-level measures of teacher characteristics are constructed similarly, by taking the average of teacher demographic dummy variables and standardized value-added measures over teachers who were at each school in October of the academic year.

Progress Report

Progress Report overall scores and sub-scores were obtained from the data files released on the DOE website. The overall scores and sub-scores were standardized by school type - Elementary, Middle, and High School - to have mean zero and standard deviation one each academic year. The progress report overall score and the squared overall score from the 2006-07 academic year was used as a control variable in all school-level regressions.

Survey Results

DOE posts school-level summary results for teacher, parent, and student learning environment surveys on its website. We took teacher survey response rates and subscores in the areas of academic expectations, communication, engagement, and safety and respect from the data files. The subscores were standardized by school level and by academic year to have mean zero and standard deviation one.

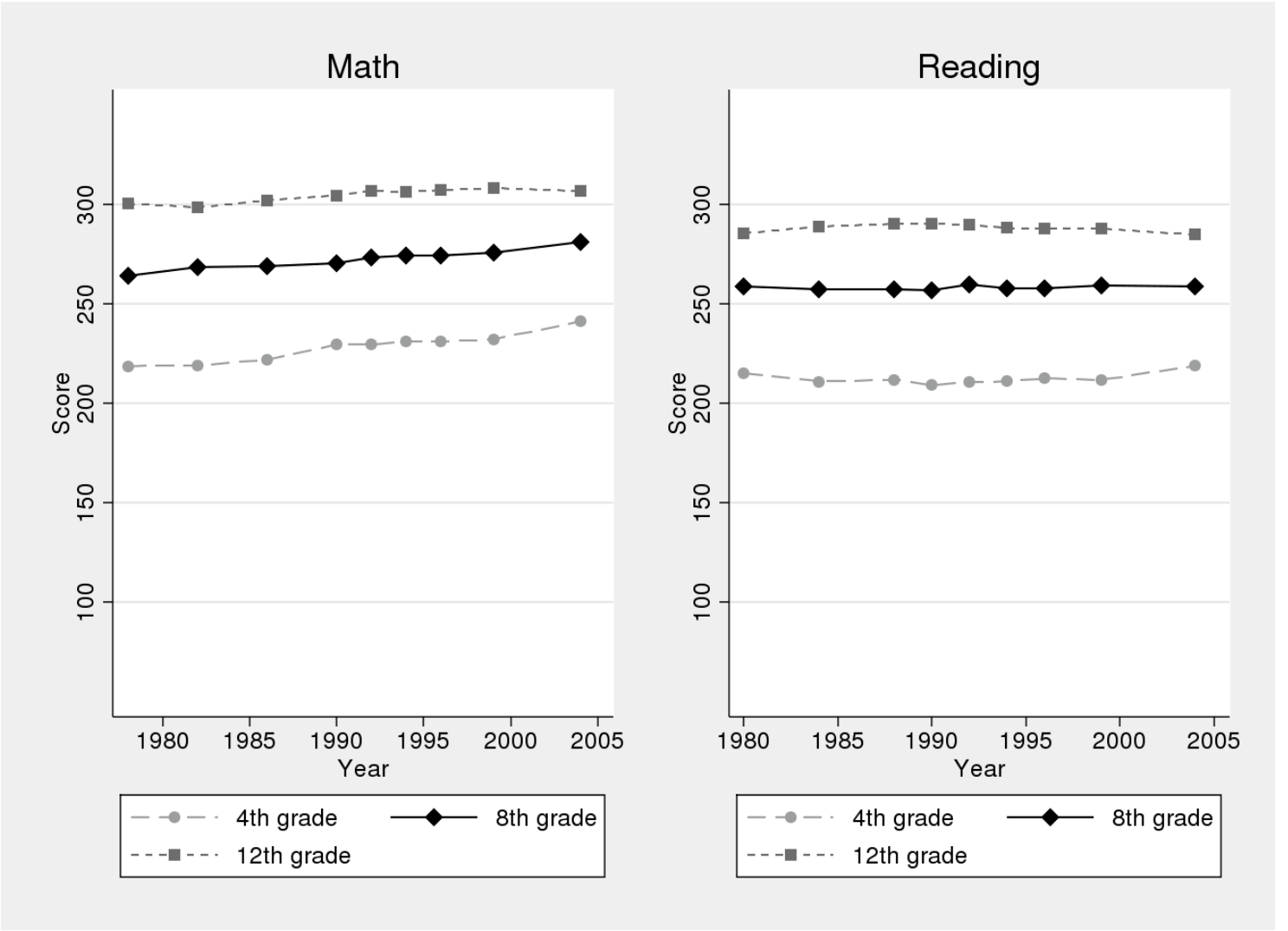
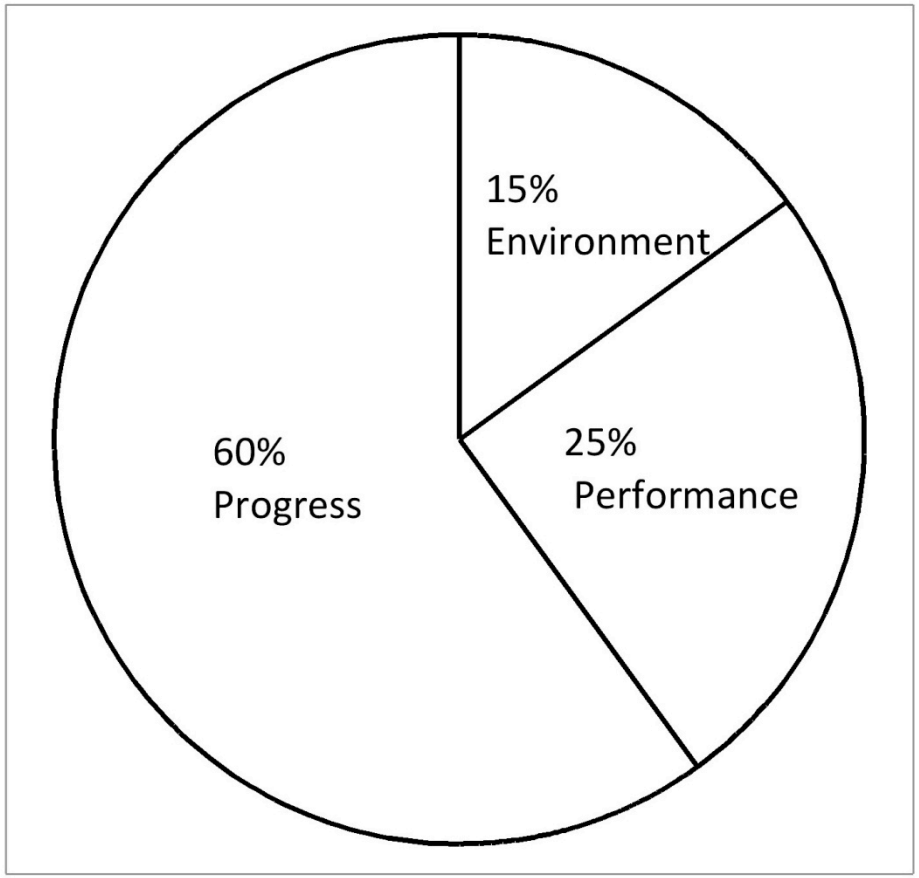


Figure 1: NAEP Achievement Scores, 1978-2004



Subscore	Criteria
Environment	5% Attendance, 10% Learning Environment Survey results
Progress	Elementary/Middle schools: Average change in state exam proficiency ratings among level 1 and 2 students, average change in state exam proficiency ratings among level 3 and 4 students, percentage of students making a year of progress among the bottom third High schools: Percentage of students earning more than 10 credits among the bottom third, weighted Regents pass rates, average completion rates of remaining Regents
Performance	Elementary/middle schools: Proportion of students at state ELA and math exam performance level 3 or 4, state exam median proficiency ratings High schools: 4-and 6-year graduation rates, diploma-weighted graduation rates

Figure 2: Progress Report Card Metrics

Figure 3A: The Distribution of Individual Bonus Amount

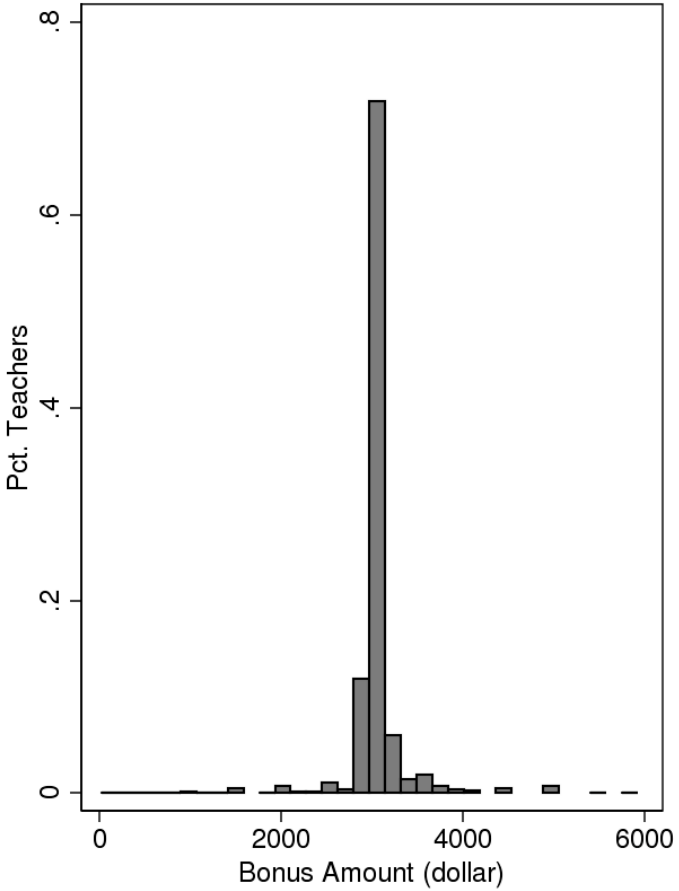
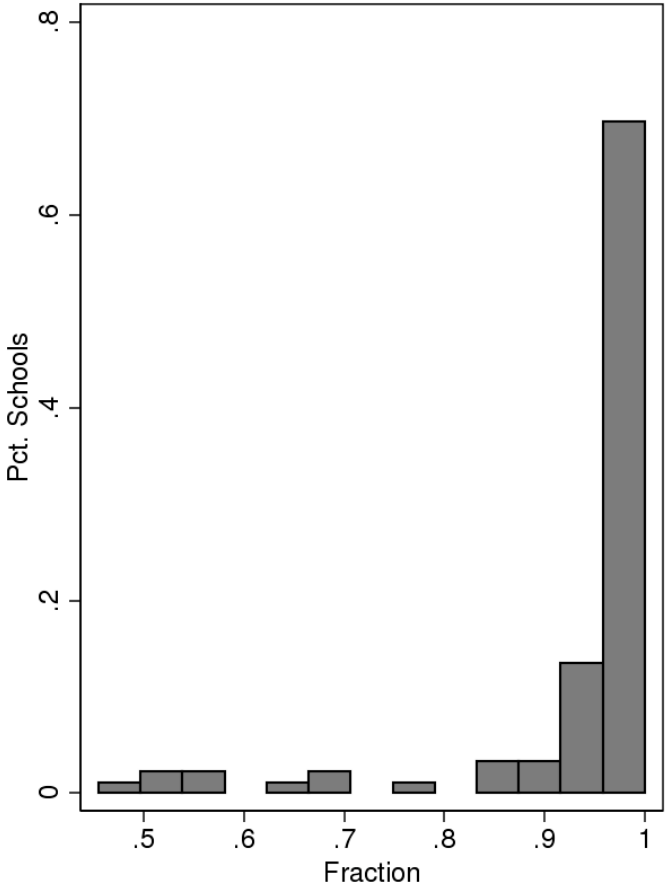
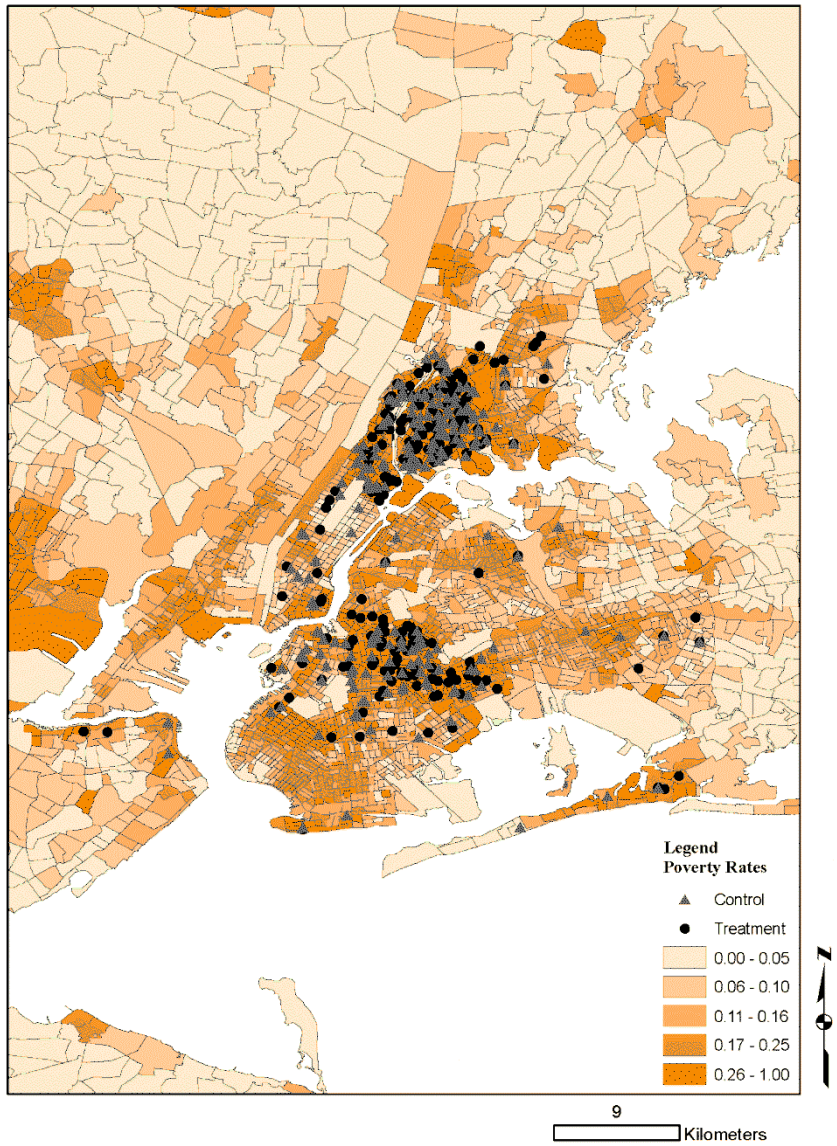


Figure 3B: The Distribution of School Incentives Scheme





Appendix Figure 1: A Map of Treatment and Control Schools

Table 1: Sample Accounting

	Number of Schools	Number of Observations		
		Elementary	Middle	High
Met the eligibility criteria	438			
Barred by the UFT	34			
Special district schools	8			
Experimental sample	396	177,861	157,616	207,741
Treatment group	233	104,823	93,472	115,518
Control group	163	73,038	64,144	92,223
Offered treatment in year 1	233			
Treated in year 1	198			
Offered treatment in year 2	195			
Treated in year 2	191			
Offered treatment in year 3	191			
Treated in year 3	189			
Valid state ELA exam scores		172,660	151,059	-
Valid state math exam scores		173,170	151,446	-
Valid Regents English exam scores		-		55,754
Valid Regents math exam scores		-		53,657
Valid Regents science exam scores		-		56,789
Valid Regents U.S. history exam scores		-		50,160
Valid Regents global history exam scores		-		61,881

NOTES: The number of observations is defined as the number of records of student and academic year pairs in the sample. The rows denoted valid state or Regents exam scores report the number of observations with non-missing state or Regents exam scores, respectively, and non-missing grade levels.

Table 2: Descriptive Statistics and Covariate Balance

	NYC District	Experimental Sample	Treatment Group	Control Group	P-value Balance	P-value Balance
	(1)	(2)	(3)	(4)	(1) & (2)	(3) & (4)
% White	0.119 (0.192)	0.014 (0.022)	0.013 (0.017)	0.015 (0.027)	0.00	0.51
% Black	0.365 (0.291)	0.411 (0.265)	0.407 (0.266)	0.415 (0.265)	0.00	0.76
% Hispanic	0.405 (0.256)	0.552 (0.266)	0.557 (0.267)	0.546 (0.265)	0.00	0.69
% Asian	0.105 (0.162)	0.018 (0.031)	0.018 (0.025)	0.019 (0.039)	0.00	0.72
% Other Race	0.006 (0.007)	0.005 (0.005)	0.005 (0.005)	0.005 (0.005)	0.01	0.98
% Male	0.502 (0.080)	0.515 (0.071)	0.514 (0.080)	0.517 (0.056)	0.00	0.65
% Female	0.498 (0.080)	0.485 (0.071)	0.486 (0.080)	0.483 (0.056)	0.00	0.65
% Free Lunch Eligible	0.858 (0.181)	0.959 (0.038)	0.960 (0.040)	0.959 (0.036)	0.00	0.92
% Special Education	0.083 (0.056)	0.110 (0.051)	0.108 (0.050)	0.113 (0.052)	0.00	0.31
% English Language Learner	0.135 (0.143)	0.191 (0.149)	0.189 (0.143)	0.195 (0.158)	0.00	0.71
2006-07 ELA Score	652.950 (17.201)	639.315 (9.070)	639.602 (9.220)	638.906 (8.869)	0.00	0.50
2006-07 Math Score	671.321 (20.417)	657.173 (14.880)	657.806 (15.029)	656.269 (14.675)	0.00	0.36
% Missing 06-07 ELA Score	0.506 (0.262)	0.501 (0.257)	0.503 (0.259)	0.498 (0.256)	0.70	0.86
% Missing 06-07 Math Score	0.496 (0.267)	0.489 (0.264)	0.491 (0.266)	0.486 (0.263)	0.58	0.86
8th Grade ELA Score	661.061 (20.446)	646.488 (12.854)	647.425 (9.554)	645.142 (16.574)	0.00	0.44
8th Grade Math Score	672.029 (21.222)	655.688 (6.750)	656.127 (6.314)	655.076 (7.372)	0.00	0.50
% Missing 8th Grade ELA Score	0.262 (0.173)	0.280 (0.201)	0.261 (0.187)	0.306 (0.220)	0.26	0.32
% Missing 8th Grade Math Score	0.218 (0.147)	0.223 (0.183)	0.211 (0.163)	0.240 (0.208)	0.67	0.49
School Size ÷ 100	6.832 (5.738)	6.357 (4.118)	6.258 (4.071)	6.499 (4.192)	0.05	0.57
% Female Teachers	0.822 (0.128)	0.810 (0.101)	0.813 (0.103)	0.805 (0.100)	0.05	0.46
% Male Teachers	0.178 (0.128)	0.190 (0.101)	0.187 (0.103)	0.195 (0.100)	0.05	0.46
% White Teachers	0.566 (0.257)	0.378 (0.176)	0.396 (0.186)	0.351 (0.158)	0.00	0.03
% Black Teachers	0.231 (0.232)	0.337 (0.226)	0.321 (0.225)	0.360 (0.226)	0.00	0.14
% Hispanic Teachers	0.154 (0.155)	0.246 (0.169)	0.247 (0.175)	0.245 (0.161)	0.00	0.89
% Asian Teachers	0.044 (0.073)	0.034 (0.030)	0.030 (0.027)	0.038 (0.033)	0.00	0.02
% Other Race Teachers	0.002 (0.007)	0.003 (0.007)	0.003 (0.008)	0.003 (0.007)	0.13	0.92
Teacher Salary ÷ 1000	68.048 (7.397)	66.512 (4.043)	66.430 (4.035)	66.628 (4.067)	0.00	0.67
Teacher Experience	8.300 (2.600)	7.900 (2.109)	7.919 (2.136)	7.873 (2.078)	0.00	0.85
2006-07 ELA TVA	0.044 (0.557)	0.055 (0.528)	0.067 (0.543)	0.038 (0.508)	0.67	0.65
2006-07 Math TVA	0.024 (0.560)	0.041 (0.590)	0.097 (0.602)	-0.040 (0.565)	0.52	0.04
% Missing 06-07 ELA TVA	0.875 (0.060)	0.878 (0.057)	0.877 (0.056)	0.880 (0.059)	0.35	0.70
% Missing 06-07 Math TVA	0.871 (0.058)	0.872 (0.052)	0.870 (0.053)	0.876 (0.049)	0.73	0.34
Number of Teachers in School	57.276 (27.485)	59.061 (21.618)	59.130 (21.768)	58.961 (21.486)	0.17	0.95
N	1417	396	233	163		

NOTES: Each column reports summary statistics from different samples, as indicated in the column header. All variables are school level measures or averages. Teacher value added was standardized to have mean zero and standard deviation one by test grade level in the full city sample before the school average was taken. The mean and the standard deviation of the variables indicated in the far left column are reported in the table. The p-value of the difference between samples are reported in the last two columns. The number of observations in each sample is reported at the bottom of the table.

Table 3: The Impact of Teacher Incentives on Student Achievement, Elementary and Middle School

	Elementary			Middle			Pooled		
	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT
ELA: Raw	1.319**	-0.010	-0.008	1.139**	-0.033**	-0.029**	1.236**	-0.020	-0.017
	(0.052)	(0.022)	(0.017)	(0.042)	(0.012)	(0.011)	(0.036)	(0.014)	(0.011)
N	172660	172660	172660	151059	151059	151059	323719	323719	323719
ELA: Controls	1.323**	-0.011	-0.008	1.140**	-0.032**	-0.028**	1.236**	-0.020	-0.016
	(0.050)	(0.020)	(0.015)	(0.043)	(0.011)	(0.010)	(0.035)	(0.013)	(0.010)
N	172660	172660	172660	151059	151059	151059	323719	323719	323719
Math: Raw	1.318**	-0.017	-0.013	1.138**	-0.053**	-0.047**	1.235**	-0.033*	-0.027*
	(0.052)	(0.025)	(0.019)	(0.042)	(0.018)	(0.016)	(0.036)	(0.017)	(0.014)
N	173170	173170	173170	151446	151446	151446	324616	324616	324616
Math: Controls	1.321**	-0.015	-0.012	1.139**	-0.048**	-0.042**	1.235**	-0.031*	-0.025*
	(0.050)	(0.024)	(0.018)	(0.043)	(0.017)	(0.015)	(0.035)	(0.016)	(0.013)
N	173170	173170	173170	151446	151446	151446	324616	324616	324616

NOTES: Each column reports results from separate regressions. The first three columns report results on the sample with test grade levels 3 to 5 and the next three columns on the sample with grade levels 6 to 8. The last three columns report results on the pooled sample. The dependent variables are the state ELA and math scores standardized to have mean zero and standard deviation one by grade level each academic year in the full city sample. Scores from all three years of implementation are used. First-stage regression uses the number of years that student received treatment as the outcome variable and reports the coefficient on the dummy variable for being randomized into the treatment group. The Intent-to-Treat estimates report the effect of being assigned to the treatment group. The Treatment-on-Treated estimates report the effect of spending time in treated schools, using the random assignment into the treatment group as the instrument. Raw regressions control for 2006-07 state test scores, test grade level dummies, and year fixed effects. Control regressions include student demographic variables and school characteristics as additional control variables. Standard errors, reported in parentheses, are clustered at school level. The number of observations, denoted N and defined in Table 1, is also reported. * denotes significance at 10% level and ** at 5% level.

Table 4: The Impact of Teacher Incentives on Student Achievement, High School

Panel A: Regents Exam Scores

	English			Mathematics			Science			U.S. History			Global History		
	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT
Raw	1.082**	0.009	0.009	1.087**	-0.011	-0.010	1.035**	-0.018	-0.017	1.106**	-0.028	-0.025	1.008**	-0.058	-0.057
	(0.090)	(0.051)	(0.047)	(0.051)	(0.028)	(0.026)	(0.061)	(0.037)	(0.036)	(0.089)	(0.061)	(0.055)	(0.079)	(0.045)	(0.045)
N	55754	55754	55754	53657	53657	53657	56789	56789	56789	50160	50160	50160	61881	61881	61881
Controls	1.069**	-0.003	-0.003	1.076**	-0.011	-0.010	1.020**	-0.016	-0.015	1.088**	-0.033	-0.030	0.987**	-0.063	-0.064
	(0.086)	(0.044)	(0.041)	(0.053)	(0.031)	(0.029)	(0.062)	(0.037)	(0.036)	(0.086)	(0.054)	(0.050)	(0.080)	(0.045)	(0.045)
N	55754	55754	55754	53657	53657	53657	56789	56789	56789	50160	50160	50160	61881	61881	61881

Panel B: 4-year Graduation

	Graduated			Regents Diploma		
	First Stage	ITT	TOT	First Stage	ITT	TOT
Raw	0.840**	-0.056**	-0.066**	0.996**	-0.077*	-0.078*
	(0.075)	(0.023)	(0.029)	(0.083)	(0.044)	(0.046)
N	27995	27995	27995	15803	15803	15803
Controls	0.830**	-0.044**	-0.053**	0.985**	-0.074*	-0.075*
	(0.070)	(0.021)	(0.026)	(0.078)	(0.042)	(0.044)
N	27995	27995	27995	15803	15803	15803

NOTES: Each column reports results from separate regressions. The dependent variables are Regents exam scores in comprehensive English, mathematics (integrated algebra), science (living environment), U.S. history, and global history, and graduation outcomes. Regents exam scores are standardized by exam type each academic year to have mean zero and standard deviation one in the full city sample. Graduation outcome is coded as dummy variables that take the value one if student graduated, or if student graduated with Regents diploma, respectively, and zero otherwise. First-stage uses the number of years that student received treatment as the outcome variable and reports the coefficient on the dummy variable for being in the treatment group. The Intent-to-Treat estimates report the effect of being assigned to the treatment group. The Treatment-on-Treated estimates report the effect of spending time in treated schools, using the random assignment into the treatment group as the instrument. Raw regressions control for 8th grade ELA and math state test scores, grade fixed effects, and year fixed effects. Control regressions control for student demographics and school characteristics in addition. Standard errors, reported in parentheses, are clustered at school level. The number of observations, denoted N and defined in Table1, is also reported. * denotes significance at 10% level and ** at 5% level.

Table 5: An Analysis of Subsamples

School Level	Subject	Full Sample	Male	Female	White	Black	Hispanic	Asian	Free Lunch	No Free Lunch
Elementary	ELA	-0.011	-0.009	-0.013	0.061	-0.019	-0.006	0.003	-0.009	0.032
		(0.020)	(0.020)	(0.020)	(0.077)	(0.023)	(0.022)	(0.055)	(0.020)	(0.033)
		172660	87599	85024	1690	65898	101722	2637	147003	5513
	Math	-0.015	-0.018	-0.013	0.120	-0.030	-0.006	-0.065	-0.011	-0.007
		(0.024)	(0.024)	(0.024)	(0.094)	(0.028)	(0.026)	(0.076)	(0.024)	(0.040)
		173170	87887	85254	1706	65877	102222	2658	147768	5537
Middle	ELA	-0.032**	-0.035**	-0.028**	-0.077*	-0.029**	-0.030**	-0.050	-0.030**	0.004
		(0.011)	(0.012)	(0.012)	(0.045)	(0.012)	(0.013)	(0.038)	(0.011)	(0.020)
		151059	78093	72932	1697	62201	83553	2926	125371	6135
	Math	-0.048**	-0.048**	-0.048**	-0.035	-0.053**	-0.042**	-0.063	-0.043**	0.002
		(0.017)	(0.017)	(0.018)	(0.045)	(0.020)	(0.020)	(0.051)	(0.017)	(0.024)
		151446	78327	73097	1736	62095	83969	2978	126065	6170
High	English	-0.003	-0.017	0.012	-0.101	-0.014	0.010	0.019	0.019	-0.019
		(0.044)	(0.048)	(0.042)	(0.082)	(0.039)	(0.052)	(0.083)	(0.046)	(0.061)
		55754	28708	27021	1155	22702	29330	2215	38284	3471
	Mathematics	-0.011	-0.012	-0.010	-0.056	0.010	-0.016	-0.109	-0.007	0.005
		(0.031)	(0.032)	(0.032)	(0.061)	(0.028)	(0.035)	(0.070)	(0.026)	(0.039)
		53657	27706	25946	910	23933	27121	1363	40537	3259
	Science	-0.016	-0.019	-0.012	-0.199**	0.038	-0.029	-0.153**	0.002	0.066
		(0.037)	(0.041)	(0.038)	(0.067)	(0.042)	(0.034)	(0.072)	(0.036)	(0.051)
		56789	28007	28771	1041	23865	29581	1963	41251	3473
	U.S. History	-0.033	-0.023	-0.040	-0.211**	-0.020	-0.029	-0.082	-0.014	-0.040
		(0.054)	(0.058)	(0.053)	(0.089)	(0.043)	(0.062)	(0.066)	(0.055)	(0.055)
		50160	25536	24608	1051	21285	25615	1874	34331	3188
	Global History	-0.063	-0.059	-0.065	-0.264**	-0.031	-0.062	-0.157**	-0.039	-0.056
		(0.045)	(0.051)	(0.042)	(0.074)	(0.044)	(0.046)	(0.068)	(0.040)	(0.051)
		61881	31585	30283	1234	26574	31649	2049	43310	3827

NOTES: Each column reports regression results on the subsample indicated on the column header. The dependent variables are standardized state or Regents exam scores, as denoted in the far left column. Estimation is done using the same regression specifications as Tables 3 and 4. From each regression, the ITT estimate is reported with standard errors in parentheses. The number of observations, defined in Table 1, is also reported. * denotes significance at 10% level and ** at 5% level.

Table 5: An Analysis of Subsamples (Continued)

School Level	Subject	Full Sample	Score Low	Score High	Schl Size Low	Schl Size High	TVA Low	TVA High	T. Salary Low	T. Salary High
Elementary	ELA	-0.011	-0.014	0.003	0.021	-0.053**	0.006	-0.025	0.003	-0.029
		(0.020)	(0.018)	(0.019)	(0.027)	(0.026)	(0.023)	(0.031)	(0.024)	(0.023)
		172660	30075	29487	86779	85788	26605	26583	72664	70377
	Math	-0.015	-0.022	-0.009	0.008	-0.050	-0.020	-0.052*	0.005	-0.034
		(0.024)	(0.021)	(0.021)	(0.032)	(0.031)	(0.027)	(0.031)	(0.028)	(0.027)
		173170	29978	29422	86980	86074	28865	28816	72555	70730
Middle	ELA	-0.032**	-0.030**	-0.034**	-0.012	-0.055**	-0.045**	-0.043**	-0.028**	-0.036**
		(0.011)	(0.011)	(0.012)	(0.013)	(0.015)	(0.014)	(0.021)	(0.012)	(0.015)
		151059	71093	69798	75830	75175	32963	32782	58479	58014
	Math	-0.048**	-0.049**	-0.053**	-0.013	-0.091**	-0.045**	-0.079**	-0.024	-0.063**
		(0.017)	(0.017)	(0.019)	(0.019)	(0.026)	(0.022)	(0.026)	(0.022)	(0.019)
		151446	70844	69607	76063	75314	36271	36239	60712	57406
High	English	-0.003	-0.023	0.019	-0.002	-0.007				
		(0.044)	(0.049)	(0.039)	(0.049)	(0.047)				
		55754	19261	18952	28406	27348				
	Mathematics	-0.011	0.022	0.007	0.020	-0.046				
		(0.031)	(0.039)	(0.052)	(0.024)	(0.046)				
		53657	13043	12648	27087	26570				
	Science	-0.016	0.014	0.033	0.048	-0.089**				
		(0.037)	(0.050)	(0.045)	(0.045)	(0.044)				
		56789	14894	14615	28516	28273				
	U.S. History	-0.033	-0.018	-0.018	0.023	-0.074				
		(0.054)	(0.062)	(0.049)	(0.054)	(0.066)				
		50160	17429	17155	25275	24885				
	Global History	-0.063	-0.088*	-0.036	0.033	-0.138**				
		(0.045)	(0.052)	(0.048)	(0.044)	(0.061)				
		61881	19899	19757	31590	30291				

NOTES: The same condition as described in the previous page applies. For the second and the third columns, the experimental sample was divided into two subsamples – elementary or middle (high) school students with 2006-07 average state test scores (8th grade test scores) above the median and below the median, and the ITT effect was estimated in each subsample. The same procedure was done using school size, and the 2006-07 teacher value added and the salary in 2007 of the teacher who taught student each year. Information on teacher value added and salary was available only for elementary and middle school students.

Table 6: The Impact of Teacher Incentives on Alternative Outcomes

	Elementary			Middle			High		
	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT
Attendance Rate	1.308**	-0.018	-0.013	1.125**	-0.019	-0.017	0.915**	-0.014	-0.016
	(0.049)	(0.020)	(0.015)	(0.042)	(0.022)	(0.020)	(0.063)	(0.050)	(0.055)
N	177861	177861	177861	157616	157616	157616	207741	207741	207741
Behavior Problems	1.308**	-0.000	-0.000	1.125**	0.012	0.011	0.915**	0.061*	0.066*
	(0.049)	(0.017)	(0.013)	(0.042)	(0.020)	(0.018)	(0.063)	(0.033)	(0.037)
N	177861	177861	177861	157616	157616	157616	207741	207741	207741
GPA	1.395**	-0.001	-0.000	1.152**	0.001	0.001	1.161**	-0.004	-0.003
	(0.084)	(0.040)	(0.029)	(0.047)	(0.031)	(0.027)	(0.075)	(0.029)	(0.025)
N	44454	44454	44454	129759	129759	129759	123441	123441	123441
Predictive ELA	1.069**	-0.019	-0.018	0.955**	-0.022	-0.023			
	(0.040)	(0.016)	(0.015)	(0.040)	(0.018)	(0.019)			
N	106382	106382	106382	51272	51272	51272			
Predictive Math	1.063**	-0.023	-0.022	0.956**	-0.051**	-0.054**			
	(0.040)	(0.020)	(0.019)	(0.040)	(0.022)	(0.023)			
N	105968	105968	105968	50915	50915	50915			

NOTES: Each column reports results from separate regressions. The dependent variables are attendance rate, the number of behavioral problems, annual GPA, and spring predictive state exam scores from each academic year. All outcome variables are standardized by grade level each academic year to have mean zero and standard deviation one in the full city sample. The first three columns report results on the sample of students in grade levels 3 to 5, the next three grade levels 6 to 8, and the last three grade levels 9 to 12. The regression specifications used are the same as in Tables 3 and 4. Standard errors, reported in parentheses, are clustered at school level. The number of observations, denoted N and defined in Table 1, is also reported. * denotes significance at 10% level and ** at 5% level.

Table 7: The Impact of Teacher Incentives on Teacher Behavior

	Elementary School			Middle School			K-8 School		
	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT
Retention in District	1.224**	0.002	0.002	1.233**	-0.006	-0.005	1.290**	0.010	0.007
	(0.052)	(0.006)	(0.005)	(0.082)	(0.011)	(0.009)	(0.091)	(0.013)	(0.010)
N	21700	21700	21700	8289	8289	8289	4693	4693	4693
Retention in School	1.224**	-0.007	-0.005	1.233**	-0.027	-0.022	1.290**	-0.000	-0.000
	(0.052)	(0.012)	(0.010)	(0.082)	(0.017)	(0.014)	(0.091)	(0.027)	(0.021)
N	21700	21700	21700	8289	8289	8289	4693	4693	4693
Personal Absences	1.220**	0.275	0.225	1.225**	-0.440	-0.359	1.290**	0.613	0.475
	(0.053)	(0.212)	(0.174)	(0.083)	(0.403)	(0.325)	(0.091)	(0.496)	(0.382)
N	18543	18543	18543	6727	6727	6727	3977	3977	3977

NOTES: Each column reports results from separate regressions. The dependent variables are retention in district and in school and personal absences in an academic year. Retention in district is coded as a dummy variable that takes the value one if teacher stays in the New York City public school district in the next academic year and zero otherwise. Retention in school is coded similarly, as a dummy variable that takes the value one if teacher stayed in the same school the next academic year. Teacher absences is the number of days teacher was absent from school for personal reasons in an academic year. Outcome variables from the first two years of implementation are used. First-stage uses the number of years that teacher received treatment as the outcome variable and reports the coefficient on the dummy variable for being in the treatment group. The Intent-to-Treat estimates report the effect of being assigned to the treatment group. The Treatment-on-Treated estimates report the effect of teaching at treated schools, using the random assignment into the treatment group as the instrument. Teacher demogrpaphic variables and the 2006-07 teacher value added are included as controls. Standard errors, reported in parentheses, are clustered at school level. The number of observations, denoted N and defined as the number of records at teacher-academic year level, is also reported. * denotes significance at 10% level and ** at 5% level.

Table 8: The Impact of Teacher Incentives on Teacher Survey Results

	Elementary			Middle			High		
	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT
Response Rate	1.667**	0.017	0.010	1.714**	0.009	0.005	1.640**	0.034	0.021
	(0.074)	(0.023)	(0.013)	(0.107)	(0.043)	(0.024)	(0.129)	(0.030)	(0.017)
N	557	557	557	240	240	240	219	219	219
Safety/Respect	1.667**	-0.089	-0.053	1.714**	0.167	0.098	1.640**	0.093	0.057
	(0.074)	(0.126)	(0.075)	(0.107)	(0.164)	(0.090)	(0.129)	(0.170)	(0.098)
N	557	557	557	240	240	240	217	217	217
Communication	1.667**	0.070	0.042	1.714**	0.035	0.020	1.640**	0.003	0.002
	(0.074)	(0.133)	(0.078)	(0.107)	(0.194)	(0.108)	(0.129)	(0.188)	(0.108)
N	557	557	557	240	240	240	217	217	217
Engagement	1.667**	0.045	0.027	1.714**	0.068	0.039	1.640**	0.035	0.022
	(0.074)	(0.129)	(0.076)	(0.107)	(0.182)	(0.101)	(0.129)	(0.175)	(0.100)
N	557	557	557	240	240	240	217	217	217
Academic Expectations	1.667**	-0.007	-0.004	1.714**	0.035	0.020	1.640**	0.093	0.057
	(0.074)	(0.129)	(0.076)	(0.107)	(0.185)	(0.103)	(0.129)	(0.173)	(0.099)
N	557	557	557	240	240	240	217	217	217

NOTES: Each column reports results from separate regressions. The dependent variables are teacher survey response rates and subscores, standardized by school level (Elementary, Middle, and High) to have mean zero and standard deviation one. First-stage uses the number of years that school received treatment as the outcome variable and reports the coefficient on the dummy variable for being in the treatment group. The Intent-to-Treat estimates report the effect of being assigned to the treatment group. The Treatment-on-Treated estimates report the effect of receiving treatment, with the random assignment into the treatment group as the instrument. All regressions control for student demographics and previous achievement measures. Standard errors are reported in parentheses. The number of observations, denoted N and defined as the number of records at school-academic year level, is also reported. * denotes significance at 10% level and ** at 5% level.