Teacher Incentives and Student Achievement: Evidence from New York City Public Schools

Roland G. Fryer, *Harvard University and NBER*

As global policy makers and school leaders look for ways to improve student performance, financial incentives programs for teachers have become increasingly popular. This article describes a school-based randomized trial in over 200 New York City public schools designed to better understand the impact of teacher incentives. I find no evidence that teacher incentives increase student performance, attendance, or graduation, nor do I find evidence that these incentives change student or teacher behavior. If anything, teacher incentives may decrease student achievement, especially in larger schools. The article concludes with a speculative discussion of theories to explain these stark results.

> 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. (US Secretary of Education Arne Duncan, the National Press Club, July 27, 2010)

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

[*Journal of Labor Economics*, 2013, vol. 31, no. 2, pt. 1]

0734-306X/2013/3102-0001$10.00

I. Introduction

Human capital, especially teacher quality, is believed to be one of the most important inputs into education production. A 1 standard deviation increase in teacher quality raises math achievement by 0.15–0.24 standard deviations per year and reading achievement by 0.15–0.20 standard deviations per year (Rockoff 2004; Hanushek and Rivkin 2005; Aaronson, Barrow, and Sander 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 (Rivkin, Hanushek, and Kain 2005; Aaronson et al. 2007; 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, Evans, and Schwab 2004; Hoxby and Leigh 2004).[1]

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 three effects. If teachers lack motivation or incentive to put effort into important inputs to the education production function (e.g., lesson planning, parental engagement), financial incentives for student achievement may have a positive impact by motivating teachers to increase their effort. However, 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. Conversely, if 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 could 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).

Despite intense opposition, there has been growing enthusiasm among education reformers and policy makers around the world to link teacher

Foundation is gratefully acknowledged. Contact the author at rfryer@fas.harvard.edu. The usual caveat applies.

[1] Corcoran et al. (2004) find that in 1964–71, 20%–25% of new female teachers were ranked in the top 10% of their high school cohort, while in 2000, less than 13% 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% in 1963 to 1% in 2000, and the share in the lowest aptitude category rose from 16% to 36% in the same period.

compensation to student achievement in myriad ways.[2] 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, Tennessee, and Texas, and Washington, DC, 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 US Department of Education selected 62 programs in 27 states to receive over $1.2 billion over 5 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 India.

The empirical evidence on the efficacy of teacher incentives is ambiguous. Data from field experiments in Kenya and India yield effect sizes of approximately 0.20 standard deviations in math and reading when teachers earned average incentives of 2% and 3% of their yearly salaries, respectively (Glewwe, Ilias, and Kremer 2010; Muralidharan and Sundararaman 2011). Data from a pilot initiative in Tennessee, where the average treatment teachers earned incentives totaling 8% of their annual salary, suggest no effect of incentives on student achievement.[3]

In the 2007–8 through the 2009–10 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

[2] Merit pay faces opposition from the two major unions: the American Federation of Teachers (AFT) and the National Education Association (NEA). Although in favor of reforming teacher compensation systems, the AFT and the NEA officially object to programs that reward teachers on the basis of student test scores and principal evaluations, while favoring instead systems that reward teachers on the basis of 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.

[3] Nonexperimental analyses of teacher incentive programs in the United States have also shown no measurable success, although one should interpret these data with caution due to the lack of credible causal estimates (Vigdor 2008; Glazerman, McKie, and Carey 2009).

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 on the basis of the 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 that the average New York City (NYC) public school has roughly 60 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. In elementary and middle schools, school report card scores hinge on student performance and progress on state assessments, student attendance, and learning environment survey results. High schools are evaluated similarly, with graduation rates, regents exams, and credits earned replacing state assessment results as proxies for performance and progress.

An important feature of the experiment is that schools had discretion over their incentive plans. As mentioned above, if a participating school met 100% of the annual targets, it received a lump sum equivalent to $3,000 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, or virtually anything in between. The only restriction was that schools were not allowed to distribute rewards on the basis of 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 my 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.

[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 to agree to include random assignment to ensure a proper evaluation.

The results from this incentive experiment are informative. Providing incentives to teachers on the basis of the 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 -0.018 (0.024) standard deviations (hereafter σ) in mathematics and -0.014σ (0.020) in reading for elementary schools and -0.046σ (0.018) in math and -0.030σ (0.011) in reading for middle schools, per year. Thus, if an elementary school student attended schools that implemented the teacher incentive program for 3 years, her test scores would decline by -0.054σ in math and by -0.042σ in reading, neither of which is statistically significant. For middle school students, however, the negative impacts are more sizable: -0.138σ in math and -0.090σ in reading over a 3-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, I find no evidence that teacher incentives affect teacher behavior, measured by retention in district or school, number of personal absences, and teacher responses to the learning environment survey, which partly determined whether a school received the performance bonus.

I 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 (0.073), in global history for white students who took that regent's exam, to 0.114σ (0.091), in math state exam scores for white elementary school students.

The article concludes with a (necessarily) speculative discussion of possible explanations for 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. I argue that a more likely explanation is that, due in part to strong influence by teacher's unions, all incentive schemes piloted thus far in the United States have been unnecessarily complex, and thus teachers cannot always predict how their efforts translate to rewards; this provides teachers with less control than incentive experiments in developing countries and may lead teachers to underestimate the expected value of increased effort. This uncertainty and lack of control in American incentive schemes, relative to those attempted in developing countries, may explain my results. Other explanations suggesting that 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.

II. A Brief Literature Review

There is a nascent but growing body of literature on the role of teacher incentives on student performance (Lavy 2002, 2009; Vigdor 2008; Glazerman et al. 2009; Glewwe et al. 2010; Springer et al. 2010; Muralidharan and Sundararaman 2011), including an emerging literature on the optimal design of such incentives (Neal 2011). There are four papers, three of them outside the United States, that provide experimental estimates of the causal impact of teacher incentives on student achievement: Duflo and Hanna (2005), Glewwe et al. (2010), Springer et al. (2010), and Muralidharan and Sundararaman (2011).

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 the teacher absence rate was significantly lower in treatment schools (22%) compared to control schools (42%) and that student achievement in treatment schools was 0.17σ higher than in control schools.

Glewwe et al. (2010) report results from a randomized evaluation that provided fourth through eighth grade teachers 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 toward short-term increases in test scores but not toward long-term learning.

Muralidharan and Sundararaman (2011) 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 3-year pilot initiative on teacher incentives conducted in the Metropolitan Nashville School System from the 2006–7 school year through the 2008–9 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 95th 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, re-

spectively. 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]

The contribution of this article is threefold. 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 staff, this design is optimal. Second, this 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, the set of outcomes is expansive and includes information on student achievement, student behavior, teacher retention, and teacher effort.

III. Program Details

A. Overview

On October 17, 2007, NYC's mayor, schools chancellor, and the president of the UFT announced an initiative to provide teachers with financial incentives to improve student performance, attendance, and school culture. The initiative was conceived as a 2-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 deciding how incentives would be distributed to each teacher and other staff. Below, I describe how schools were selected and the incentive scheme, and I provide an overview of the distribution of incentive rewards to schools.

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

[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.

B. School Selection

Table 1 provides an accounting of how the experimental sample was selected. Eligible middle and high schools were selected on the basis of the average proficiency ratings on fourth and eighth grade state tests, respectively. Eligible elementary schools were selected on the basis of poverty rates and student demographic characteristics, such as the percentage of English language learners and special education students. The NYC DOE identified 438 schools that met the above-mentioned eligibility criteria. Of these schools, 34 were barred by the UFT for unknown reasons, and eight were District 75 (i.e., special education) schools. The remaining 396 comprise the experimental sample, among which 212 schools were randomly selected by the author and offered treatment. 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% of their active

Table 1
Sample Construction

	Number of Schools	Number of Observations		
		Elementary	Middle	High
Met the eligibility criteria	438	. . .	. . .	. . .
Barred by the United Federation of Teachers	34	. . .	. . .	. . .
Special district schools	8	. . .	. . .	. . .
Experimental sample	396	64,373	50,413	70,826
Treatment	233	37,791	29,857	38,861
Control	163	26,582	20,556	31,965
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 English language arts exam scores	. . .	61,829	47,473	. . .
Valid state math exam scores	. . .	62,418	48,044	. . .
Valid regents English exam scores	. . .	. . .	19,813	. . .
Valid regents math exam scores	. . .	. . .	11,219	. . .
Valid regents science exam scores	. . .	. . .	18,370	. . .
Valid regents US history exam scores	. . .	. . .	17,791	. . .
Valid regents global history exam scores	. . .	. . .	21,711	. . .

NOTE.—All statistics are reported for the 2007–8 school year (year 1), unless otherwise specified.

full-time staff represented by the UFT at the school to vote for the program.[7] Schools forwarded voting results through e-mail to the DOE by late November. Of the 212 schools randomly chosen to receive treatment, 179 garnered enough votes to participate, and 33 declined treatment.[8] To increase the number of schools eligible to participate, 21 schools were added off the wait list; 19 garnered the requisite votes. So, overall, 233 schools in the experimental sample were invited to participate in the program, and 198 actually participated. The final experimental sample in year 1 consists of the lottery sample, with 233 treatment schools and 163 control schools.

In the second year, 195 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 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.

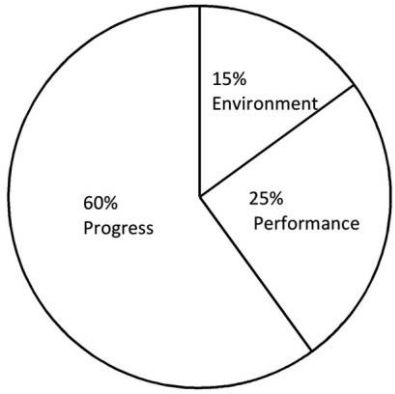
C. Incentive Scheme

Figure 1 shows how the progress report card score, which is the basis for awarding incentives, is calculated. Environment, which accounts for 15% of the progress report card score, is derived from attendance rate (5% of the overall score) and learning environment surveys administered to students, teachers, and parents in the spring semester (10%). Attendance rate is a school's average daily attendance. While environment subscores are calculated identically for all schools, student performance (25%) and student progress (60%) are calculated differently for high schools versus elementary and middle schools. In elementary and middle schools, student performance depends on the percentage of students at grade level and the median proficiency rating in English language arts (ELA) and math state tests. In high schools, student performance is measured by 4-year and 6-year graduation rates and diploma-weighted graduation rates.[9] Student progress 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 ele-

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

[8] 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.

[9] The DOE awards different levels of diplomas—local, regents, advanced regents, and advanced regents with honors—depending on the number of regents exams passed. Further details on graduation requirements and progress report card score calculation can be found on the DOE website (http://schools.nyc.gov/default.htm).



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

FIG. 1.—Progress report card metrics

mentary and middle schools. Student progress in high schools is measured by the percentage of students earning more than 10 credits and regents exam pass rates in the core subjects—English, math, science, US 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 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, 1 point is not obvious.

Table 2 shows the number of points by which schools had to increase their progress report card scores in the 2007–8 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 2006–7 academic year, it needed to increase its progress report card score by 15 points to meet the annual target.

1. An Example

Consider the following simplified example with an elementary school that ranks at about the 10th percentile citywide and at about the 25th

Table 2
Progress Report Target Points

Citywide Ranking Based on the Previous Year (Percentile)	Elementary and Middle	High
≥85th	7.5	2
≥45th and <85th	12.5	3
≥15th and <45th	15	4
≥5th and <15th	17.5	6
<5th	20	8

NOTE.—Calculated by the author.

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 its 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. However, 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.

2. 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 performance. For example, TAP in Chicago—started by Arne Duncan and described in the quote at the beginning of this article—rewarded teachers on the basis of 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 on the basis of 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 on the basis of the school's overall performance only. A teacher participating in Houston's ASPIRE program would be rewarded the predetermined 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.

D. 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 or 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 they 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 on by the compensation committee. In the first year, 104 out of 198 schools that participated met the improvement target and received the full bonus, while 18 schools met at least 75% of the target and received half of the maximum incentive payment. In total, the compensation received by participating schools totaled $22 million. In the second year, 154 out of 191 schools that participated received the full bonus, while seven schools received half the maximum compensation. The total compensation awarded to schools in the second year was $31 million. I do not have precise numbers for year 3, but the DOE claims that the total costs of the experiment was approximately $75 million.

Figure 2*A* 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 2*B* presents a histogram of the fraction of teachers receiving the same amount

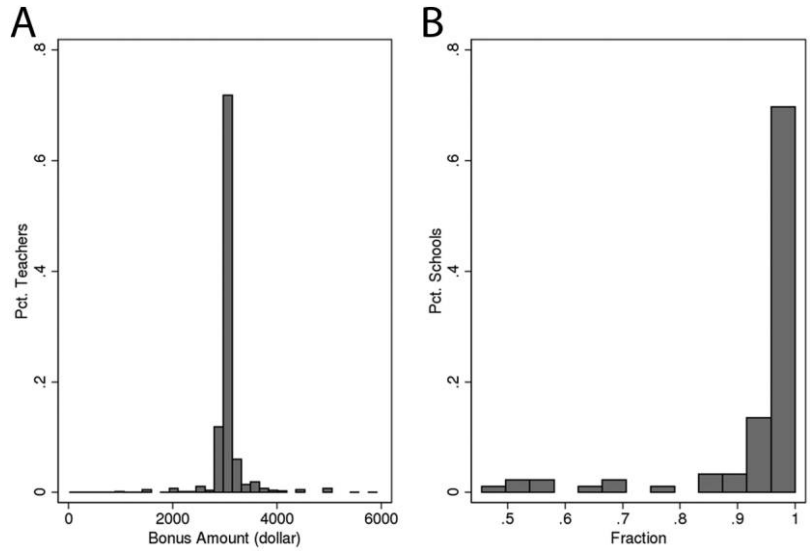


FIG. 2.—*A*, Distribution of individual bonus amount; *B*, distribution of school incentives scheme.

in each school in order to characterize how many schools decided on 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.

IV. Data and Research Design

A. Data

I 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 the 2006–7 to 2009–10 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 3–8. For high school students, the data contain regents exam scores and graduation rates. Data on attendance and behavioral incidences are available for all students.

The 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 four) on both math and ELA tests to advance to the next grade without attending summer school. Material covered in the math test is divided among five strands: (1) number sense and operations, (2) algebra, (3) geometry, (4) measurement, and (5) statistics and probability. The ELA test is designed to assess students on three learning standards: (1) information and understanding, (2) literary response and expression, and (3) critical analysis and evaluation. Both tests include multiple-choice, short-response, and extended-response questions. The ELA test also includes a brief editing task and is divided into reading and listening sections.

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 they may be granted special accommodations (additional time, translation services, etc.) at the discretion of school or state administrators. In the 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 NYC sample.

I construct measures of attendance, behavioral problems, and grade point average (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.[10] Behavioral problems are measured as the total number of behavioral incidences in record each year. Attendance, behavioral problems, and GPA were normalized to have a mean of zero and a standard deviation of one by grade level each year in the full NYC sample.

I 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 I include in all regressions. Previous year's test scores are available for most students who were in the district in the previous year.[11] I 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. See the appendix, available in the online version of *Journal of Labor Economics*, for further details.

I also construct school-level controls. To do this, I assign each student who was present at the beginning of the year, that is, in September, to the first school attended. I construct the school-level variables on the basis

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

[11] See table 1 for exact percentages of experimental group students with valid test scores from previous years.

of 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 2006–7 academic year.

I 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 nonmissing 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 2006–7 academic year until the 2008–9 academic year. I take the TVA measured in standard deviation units and standardize the number by grade level each year to have a mean of zero and a standard deviation of one in the full city sample. For teachers who taught more than one grade, I take the average of TVA across grade levels. In addition, I construct the cumulative teacher absences using HR data on teacher attendance through May of each academic year.

Table 3 provides pretreatment descriptive statistics. Columns 1–4 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 teachers, the percentage of Asian teachers, and TVA in math in the 2006–7 academic year.

B. 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 ($\text{incentive}_{i,g,s,t}$) and controls for basic student and school characteristics, X_i and X_s, respectively:

$$\text{outcome}_{i,g,s,t} = \alpha_1 + \beta_1 X_i + \gamma_1 X_s + \delta_g + \zeta_t + \pi_1 \text{incentive}_{i,g,s,t} + \varepsilon_{i,g,s,t}.$$

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, I must compare participating and nonparticipating schools that would have had the same academic outcomes had they

Table 3
Descriptive Statistics and Covariate Balance

	New York City District (1)		Experimental Sample (2)		Treatment Group (3)		Control Group (4)		*p*-Value Balance	
									(1) and (2)	(3) and (4)
Percent white	.119	(.192)	.014	(.022)	.013	(.017)	.015	(.027)	.00	.51
Percent black	.365	(.291)	.411	(.265)	.407	(.266)	.415	(.265)	.00	.76
Percent Hispanic	.405	(.256)	.552	(.266)	.556	(.267)	.546	(.265)	.00	.69
Percent Asian	.105	(.162)	.018	(.031)	.018	(.025)	.019	(.039)	.00	.72
Percent other race	.006	(.007)	.005	(.005)	.005	(.005)	.005	(.005)	.01	.98
Percent male	.502	(.080)	.515	(.071)	.514	(.080)	.517	(.056)	.00	.65
Percent female	.498	(.080)	.485	(.071)	.486	(.080)	.483	(.056)	.00	.65
Percent free lunch eligible	.858	(.181)	.959	(.038)	.960	(.040)	.959	(.036)	.00	.92
Percent special education	.083	(.056)	.110	(.051)	.108	(.050)	.113	(.053)	.00	.30
Percent English language learner	.135	(.143)	.191	(.149)	.189	(.143)	.195	(.158)	.00	.71
2006–7 ELA score	652.952	(17.199)	639.324	(9.068)	639.614	(9.214)	638.909	(8.874)	.00	.49
2006–7 math score	671.327	(20.413)	657.193	(14.880)	657.829	(15.025)	656.283	(14.680)	.00	.36
Eighth grade ELA score	661.061	(20.446)	646.489	(12.854)	647.425	(9.554)	645.143	(16.574)	.00	.44
Eighth grade math score	672.029	(21.222)	655.689	(6.750)	656.127	(6.314)	655.078	(7.372)	.00	.50
School size ÷ 100	6.832	(5.738)	6.356	(4.118)	6.257	(4.071)	6.498	(4.192)	.05	.57
2006–7 school progress report card score	53.964	(14.377)	51.680	(15.968)	52.100	(16.160)	51.084	(15.723)	.00	.53
2006–7 individual behavioral incidences	.142	(.167)	.180	(.162)	.185	(.171)	.174	(.149)	.00	.49
2006–7 school-wide behavioral incidences	98.435	(147.081)	117.699	(125.257)	120.335	(126.853)	113.933	(123.231)	.00	.62
Percent female teachers	.822	(.128)	.810	(.101)	.813	(.103)	.805	(.100)	.05	.46

Percent male teachers	.178	(.128)	.190	(.101)	.187	(.103)	.195	(.100)	.05	.46
Percent white teachers	.566	(.257)	.378	(.176)	.396	(.186)	.351	(.158)	.00	.03
Percent black teachers	.231	(.232)	.337	(.226)	.321	(.225)	.360	(.226)	.00	.14
Percent Hispanic teachers	.154	(.155)	.246	(.169)	.247	(.175)	.245	(.161)	.00	.89
Percent Asian teachers	.044	(.073)	.034	(.030)	.030	(.027)	.038	(.033)	.00	.02
Percent other race teachers	.002	(.007)	.003	(.007)	.003	(.008)	.003	(.007)	.13	.92
Teacher salary ÷ 1,000	68.048	(7.397)	66.512	(4.043)	66.430	(4.035)	66.628	(4.067)	.00	.67
Teacher experience	8.300	(2.600)	7.900	(2.109)	7.919	(2.136)	7.873	(2.078)	.00	.85
2006–7 ELA TVA	.044	(.557)	.055	(.528)	.067	(.543)	.038	(.508)	.67	.65
2006–7 math TVA	.024	(.560)	.041	(.590)	.097	(.602)	−.040	(.565)	.52	.04
Number of teachers in school	57.276	(27.485)	59.061	(21.618)	59.130	(21.768)	58.961	(21.486)	.17	.95
Percent missing 2006–7 ELA score	.506	(.262)	.501	(.258)	.503	(.259)	.498	(.256)	.70	.86
Percent missing 2006–7 math score	.496	(.267)	.489	(.265)	.491	(.266)	.486	(.263)	.58	.86
Percent missing eighth grade ELA score	.262	(.173)	.280	(.201)	.261	(.187)	.306	(.220)	.26	.32
Percent missing eighth grade math score	.218	(.147)	.223	(.183)	.211	(.163)	.240	(.208)	.67	.49
Missing 2006–7 individual behavioral incidences	.112	(.081)	.115	(.071)	.114	(.073)	.115	(.068)	.49	.93
Missing 2006–7 school-wide behavioral incidences	.030	(.170)	.000	(.000)	.000	(.000)	.000	(.000)	.00	. . .
Percent missing 2006–7 ELA TVA	.875	(.060)	.878	(.057)	.877	(.056)	.880	(.059)	.35	.70
Percent missing 2006–7 math TVA	.871	(.058)	.872	(.052)	.870	(.053)	.876	(.049)	.73	.34
N	1,417		396		233		163		. . .	. . .

NOTE.—ELA = English language arts. Each column reports summary statistics from different samples. All variables are school-level measures or averages. Teacher value added (TVA) was standardized to have a mean of zero and a standard of deviation of one by test grade level in the full city sample before the school average was taken. Means and standard deviations (in parentheses) are reported. The *p*-values of the differences between samples are reported in the last two columns.

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 the analysis to schools that were selected (by the UFT and the DOE) to be included in the experimental sample, I can estimate the causal impact of being offered a chance to participate in a 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 in 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. The specifications estimated are of the form

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

where the vector of student level controls, X_i, includes a mutually inclusive and collectively exhaustive set of race dummies, predetermined measures of the outcome variables when possible (i.e., preintervention test scores or the number of behavioral incidences), and indicators for gender, free-lunch eligibility, special education status, and English language learner status. The set of school-level controls, X_s, includes the school's prelottery number of behavioral incidences and the percentages of students at a school who are black, Hispanic, free-lunch eligible, English-language learners, and special education students. 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 program enrollment), I 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 (IV) for participation. I 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

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

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.

V. The Impact of Teacher Incentives

A. Student Achievement

Table 4 presents first-stage, ITT, and TOT estimates of the effect of teacher incentives on state math and ELA test scores. Columns 1–3 report estimates from the elementary school sample, columns 4–6 report estimates from middle schools, and columns 7–9 present results for a pooled sample of elementary and middle schools. I present both raw estimates and those that contain the parsimonious set of controls described in the previous section. Note that the coefficients in the table are normalized so that they are in standard deviation units and represent 1-year impacts.

Surprisingly, all estimates of the effect of teacher incentives on student achievement are negative in both elementary and middle school; the middle school results are significant at the 5% level. The ITT effect of the teacher incentive scheme is -0.014σ (0.020) in reading and -0.018σ (0.024) in math for elementary schools and -0.030σ (0.011) in reading and -0.046σ (0.018) in math for middle schools. The effect sizes in middle school are nontrivial—a student who attends a participating middle school for 3 years of the experiment is expected to lose 0.090σ in reading and 0.138σ in math. The TOT estimates are smaller than the ITT estimates, as the first-stage coefficients are all larger than one.

Table 5 presents results similar to table 4, 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, 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 meet the requirements for regents diploma and, in addition, score a 65 or higher in the mathematics B, life science, physical science, and foreign

Table 4
The Impact of Teacher Incentives on Student Achievement, Elementary and Middle School

	Elementary*			Middle†			Pooled Sample		
	First Stage (1)	ITT (2)	TOT (3)	First Stage (4)	ITT (5)	TOT (6)	First Stage (7)	ITT (8)	TOT (9)
ELA:									
Raw	1.319	−.013	−.010	1.137	−.031	−.027	1.236	−.020	−.017
	(.057)	(.023)	(.017)	(.045)	(.012)	(.011)	(.043)	(.014)	(.011)
Control	1.323	−.014	−.010	1.137	−.030	−.026	1.236	−.020	−.016
	(.055)	(.020)	(.015)	(.046)	(.011)	(.010)	(.043)	(.013)	(.010)
N students	175,894	175,894	175,894	147,141	147,141	147,141	323,317	323,317	323,317
N schools	230	230	230	136	136	136	323	323	323
Math:									
Raw	1.318	−.020	−.015	1.136	−.051	−.045	1.235	−.033	−.027
	(.057)	(.026)	(.020)	(.045)	(.019)	(.017)	(.043)	(.017)	(.014)
Control	1.322	−.018	−.014	1.136	−.046	−.040	1.235	−.032	−.026
	(.055)	(.024)	(.018)	(.046)	(.018)	(.016)	(.042)	(.016)	(.013)
N students	176,387	176,387	176,387	147,493	147,493	147,493	324,172	324,172	324,172
N schools	230	230	230	136	136	136	323	323	323

NOTE.—Each column reports results from separate regressions. Dependent variables are the state ELA (English language arts) and math scores standardized to have a mean of zero and a standard deviation of one by grade level each academic year in the full city sample. Scores from all 3 years of implementation are used. First-stage regression uses the years receiving treatment as the outcome variable and reports the coefficient on the dummy variable for being randomized into the treatment group. The intent-to-treat (ITT) estimates report the effect of being assigned to the treatment group using the ordinary least-squares method. The treatment-on-treated (TOT) 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 2006–7 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. Number of student observations is reported as an aggregate total of observations for all 3 years of implementation. Number of school observations is reported for the 2007–8 school year (year 1).

* Sample within schools designated as elementary in the New York City Department of Education city progress report files, as well as students in grade levels 3–5 in schools designated as K–8.

† Sample within schools designated as middle schools, as well as students in grade levels 6–8 of schools designated as K–8.

language regents exams.[12] Table 5 presents first-stage, ITT, and TOT estimates for the impact of teacher incentives on comprehensive English, mathematics, science, US history, and global history regents exam scores. All exam scores were standardized to have a mean of zero and a standard deviation of 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 insignificant. The ITT effect on the English regents exam score is -0.003σ (0.046), the effect on the integrated algebra exam score is -0.020σ (0.033), and the effect on science scores is -0.019σ (0.038). The ITT effect on the US history exam score is -0.033σ (0.057), and that on the global history exam score is -0.063σ (0.046). The TOT effect is of a comparable magnitude.

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

Table 6 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, TVA, and teacher salary. The coefficients in the table are ITT estimates with a 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. For middle school students, the only exceptions are students who are free-lunch eligible, are attending larger schools, or are taught by more experienced teachers (i.e., received higher salaries). Among high school students, the exceptions are students who are white or Asian, scored in the lowest tercile on eighth grade state tests, or are attending

[12] 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, US history, and foreign languages. In this article, I present results on comprehensive English, integrated algebra, living environment, US 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.

Table 5
The Impact of Teacher Incentives on Student Achievement, High School

	Raw				Control			
	Coefficient	Standard Error	Number of Students	Number of Schools	Coefficient	Standard Error	Number of Students	Number of Schools
Regents exam scores:								
English:								
First stage	1.082	(.095)	55,791	101	1.069	(.090)	55,791	101
ITT	.009	(.052)	55,791	101	−.003	(.046)	55,791	101
TOT	.009	(.048)	55,791	101	−.003	(.043)	55,791	101
Mathematics:								
First stage	1.111	(.058)	55,785	134	1.101	(.060)	55,785	134
ITT	−.019	(.031)	55,785	134	−.020	(.033)	55,785	134
TOT	−.017	(.028)	55,785	134	−.018	(.029)	55,785	134
Science:								
First Stage	1.042	(.066)	57,364	109	1.028	(.067)	57,364	109
ITT	−.021	(.039)	57,364	109	−.019	(.038)	57,364	109
TOT	−.020	(.037)	57,364	109	−.018	(.037)	57,364	109
US history:								
First stage	1.107	(.094)	50,315	90	1.089	(.090)	50,315	90
ITT	−.028	(.064)	50,315	90	−.033	(.057)	50,315	90
TOT	−.025	(.058)	50,315	90	−.030	(.052)	50,315	90
Global history:								
First stage	1.008	(.083)	61,892	89	.987	(.084)	61,892	89
ITT	−.057	(.045)	61,892	89	−.063	(.046)	61,892	89
TOT	−.057	(.045)	61,892	89	−.063	(.046)	61,892	89

Four-year graduation:								
Graduated:								
First stage	.840	(.078)	27,995	87	.830	(.073)	27,995	87
ITT	−.056	(.024)	27,995	87	−.044	(.021)	27,995	87
TOT	−.066	(.029)	27,995	87	−.053	(.026)	27,995	87
Regents diploma:								
First stage	.996	(.084)	15,803	85	.985	(.079)	15,803	85
ITT	−.077	(.045)	15,803	85	−.074	(.043)	15,803	85
TOT	−.078	(.046)	15,803	85	−.075	(.044)	15,803	85

NOTE.—Each row reports results from different regressions. The dependent variables are regents exam scores in comprehensive English, mathematics (integrated algebra), science (living environment), US history, and global history for students in grades 8–12 and high school graduation outcomes. Regents exam scores are standardized by exam type each academic year to have a mean of zero and a standard of deviation of one in the full city sample. Graduation outcome is coded as dummy variables that take the value one if the student graduated high school or if the student graduated with a regents diploma, respectively, and zero otherwise. First stage uses the years that the 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 (ITT) estimates report the effect of being assigned to the treatment group using the ordinary least-squares method. The treatment-on-treated (TOT) 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 eighth grade ELA (English language arts) and math state test scores for students in grades 9–12 and seventh grade ELA and math state test scores for students in eighth grade. Substituting seventh grade previous test scores for all grades does not qualitatively alter results. Raw regressions also include grade fixed effects and year fixed effects. Control regressions control for student demographics and school characteristics in addition. Standard errors are clustered at school level. Number of student observations is reported as an aggregate total of observations for all 3 years of implementation. Number of school observations is reported for the 2007–8 school year (year 1).

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	−.017	−.013	1.121	−.019	−.017	.915	−.014	−.016
	(.054)	(.021)	(.016)	(.045)	(.025)	(.022)	(.068)	(.054)	(.059)
N students	181,393	181,393	181,393	154,017	154,017	154,017	207,718	207,718	207,718
N schools	230	230	230	136	136	136	88	88	88
Behavior problems	1.308	−.000	−.000	1.121	.010	.009	.915	.061	.067
	(.054)	(.017)	(.013)	(.045)	(.022)	(.019)	(.068)	(.035)	(.039)
N students	181,393	181,393	181,393	154,017	154,017	154,017	207,718	207,718	207,718
N schools	230	230	230	136	136	136	88	88	88
GPA	1.387	.005	.004	1.153	−.003	−.003	1.161	−.004	−.003
	(.094)	(.041)	(.029)	(.050)	(.031)	(.027)	(.082)	(.033)	(.028)
N students	45,410	45,410	45,410	128,764	128,764	128,764	123,434	123,434	123,434
N schools	201	201	201	136	136	136	. . .	. . .	. . .
Predictive ELA	1.069	−.022	−.021	.947	−.021	−.022	. . .	. . .	. . .
	(.042)	(.017)	(.016)	(.042)	(.019)	(.020)	. . .	. . .	. . .
N students	108,515	108,515	108,515	49,117	49,117	49,117	. . .	. . .	. . .
N schools	230	230	230	136	136	136	. . .	. . .	. . .
Predictive math	1.063	−.026	−.025	.949	−.049	−.052	. . .	. . .	. . .
	(.042)	(.020)	(.019)	(.042)	(.022)	(.024)	. . .	. . .	. . .
N students	108,078	108,078	108,078	48,777	48,777	48,777	. . .	. . .	. . .
N schools	230	230	230	136	136	136	. . .	. . .	. . .

NOTE.—ITT = intent to treat; TOT = treatment on treated; ELA = English language arts. Each column reports results from separate regressions. The dependent variables are attendance rate, the number of behavioral problems, annual GPA (grade point average), and spring predictive state exam scores from each academic year. All outcome variables are standardized by grade level each academic year to have a mean of zero and a standard deviation of one in the full city sample. The regression specification used is the same as in tables 4 and 5. Standard errors, reported in parentheses, are clustered at school level. Number of student observations is reported as an aggregate total of observations for all 3 years of implementation. Number of school observations is reported for the 2007–8 school year (year 1).

* Sample within schools designated as elementary in the New York City Department of Education city progress report files, as well as students in grade levels 3–5 in schools designated as K–8.

† Sample within schools designated as middle schools, as well as students in grade levels 6–8 of schools designated as K–8.

‡ Students in grade levels 9–12.

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, although not shown here in tabular form, demonstrate that the coefficient on treatment is small and statistically zero.[13] This suggests that differential attrition is not likely to be a concern in interpreting the results.

B. Alternative Outcomes

Thus far, I have concentrated on student progress on state assessments, the most heavily weighted element of NYC's incentive scheme. Now, I introduce two additional measures of student performance and three measures of school environment: GPAs, predictive math and ELA exams, school environment surveys, attendance, and behavioral incidences. Many of these outcomes enter directly into the incentive scheme and may be affected by it.

Table 7 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 3–8 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 article, I analyze math and ELA test scores from the spring Acuity Predictive Assessment.[14] 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 a mean of zero and a standard deviation of one each academic year across the full city sample. Grades were extracted from files containing the transcripts of all students in the district.[15] 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 a mean of zero and a standard deviation of one among students in the same grade with the same grade scale across the school district. The number of behavioral incidences was pulled from behavior data, which record the date, level, location, and a short description of all

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

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

[15] Elementary school transcripts are not available for all schools each academic year. High school transcripts were not available until the 2008–9 academic year.

Table 7
The Impact of Teacher Incentives on Teacher Behavior

	Elementary School			Middle School			K–8		
	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT
Retention in district	1.224	.002	.002	1.233	−.006	−.005	1.290	.010	.007
	(.052)	(.006)	(.005)	(.082)	(.011)	(.009)	(.091)	(.013)	(.010)
N teachers	21,700	21,700	21,700	8,289	8,289	8,289	4,693	4,693	4,693
N schools	187	187	187	85	85	85	40	40	40
Retention in school	1.224	−.007	−.005	1.233	−.027	−.022	1.290	−.000	−.000
	(.052)	(.012)	(.010)	(.082)	(.017)	(.014)	(.091)	(.027)	(.021)
N teachers	21,700	21,700	21,700	8,289	8,289	8,289	4,693	4,693	4,693
N schools	187	187	187	85	85	85	40	40	40
Personal absences	1.220	.275	.225	1.225	−.440	−.359	1.290	.613	.475
	(.053)	(.212)	(.174)	(.083)	(.403)	(.325)	(.091)	(.496)	(.382)
N teachers	18,543	18,543	18,543	6,727	6,727	6,727	3,977	3,977	3,977
N schools	187	187	187	85	85	85	40	40	40

NOTE.—Each column reports results from different regressions. The dependent variables are retention in district and in school and personal absences in a year. Retention in district is coded as a dummy variable that takes the value one if the 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 the teacher stayed in the same school the next academic year. Teacher absences is the number of days absent from school for personal reasons in an academic year. Outcome variables from the first 2 years of implementation are used. First stage uses the number of years receiving treatment as the outcome variable and reports the coefficient on the dummy variable for being in the treatment group. The intent-to-treat (ITT) estimates report the effect of being assigned to the treatment group using the ordinary least-squares method. The treatment-on-treated (TOT) estimates report the effect of teaching at treated schools, using the random assignment into the treatment group as the instrument. Teacher demographic variables and the 2006–7 teacher value are included as controls. Standard errors, reported in parentheses, are clustered at school level. Number of teacher observations is reported as an aggregate total of observations for the first 2 years of implementation. Number of school observations is reported for the 2007–8 school year (year 1).

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 a mean of zero and a standard deviation of 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 -0.022σ (0.017) in the elementary school sample and -0.021σ (0.019) in the middle school sample. The ITT effect on predictive math exams is -0.026σ (0.020) in the elementary school sample and -0.049σ (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 state test score findings.

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 -0.017σ (0.021) in the elementary school sample, -0.019σ (0.025) in the middle school sample, and -0.014σ (0.054) in the high school sample. The effects on behavioral incidences and GPAs are similarly small and insignificant.

C. Teacher Behavior

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

I measure retention in two ways: in school and in district—both of which were constructed using HR 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 the coding of retention in school, I construct a dummy variable that takes the value one if a teacher was found in the NYC school district's 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 NYC 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, I divided the sample of teachers into elementary, middle, and K–8 school samples and estimated the treatment effects separately on each sample. To measure absences, I obtained the number of personal absences as of May for teachers who did not exit the system.

Table 8 presents results on the impact of teacher incentives on measures of teacher behavior. There is no evidence that teacher incentives affect teacher attendance or retention in either district or school. Elementary school teachers in treatment schools were 0.2% more likely to stay in the NYC school district, were 0.7% 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.

VI. Discussion

The previous sections demonstrate that the teacher incentive scheme piloted in 200 NYC 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, I 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 the analysis, along with data gleaned from other experiments, I 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 control. It is important to note that I cannot rule out the possibility that other unobservable differences between the developing countries and America (e.g., teacher motivation) produce the differences.

A. Incentives Were Not Large Enough

One potential explanation for the 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 too 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

Table 8
The Impact of Teacher Incentives on Teacher Survey Results

	Elementary			Middle			K–8			High		
	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT	First Stage	ITT	TOT
Response rate	1.667	.018	.011	1.740	.019	.011	1.883	−.011	−.006	1.559	.037	.024
	(.074)	(.023)	(.013)	(.103)	(.040)	(.022)	(.100)	(.061)	(.030)	(.136)	(.028)	(.017)
N schools	187	187	187	96	96	96	40	40	40	87	87	87
Safety/respect	1.667	−.088	−.053	1.740	.210	.121	1.884	−.558	−.296	1.559	.095	.061
	(.074)	(.126)	(.075)	(.103)	(.167)	(.091)	(.100)	(.263)	(.130)	(.135)	(.158)	(.096)
N schools	187	187	187	96	96	96	39	39	39	86	86	86
Community	1.667	.071	.042	1.740	.072	.041	1.884	−.583	−.309	1.559	−.000	−.000
	(.074)	(.133)	(.078)	(.103)	(.188)	(.102)	(.100)	(.283)	(.138)	(.135)	(.177)	(.108)
N schools	187	187	187	96	96	96	39	39	39	86	86	86
Engagement	1.667	.046	.027	1.740	.089	.051	1.884	−.403	−.214	1.559	.024	.016
	(.074)	(.129)	(.076)	(.103)	(.177)	(.097)	(.100)	(.278)	(.134)	(.135)	(.161)	(.098)
N schools	187	187	187	96	96	96	39	39	39	86	86	86
Academic expectations	1.667	−.006	−.004	1.740	.088	.050	1.884	−.506	−.268	1.559	.101	.065
	(.074)	(.129)	(.076)	(.103)	(.181)	(.099)	(.100)	(.287)	(.139)	(.135)	(.163)	(.098)
N schools	187	187	187	96	96	96	39	39	39	86	86	86

NOTE.—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 a mean of zero and a standard deviation of one. The intent-to-treat (ITT) estimates report the effect of being assigned to the treatment group using the ordinary least-squares method, and the treatment-on-treated (TOT) estimates report the effect of receiving treatment, with the random assignment as the instrument. Regressions control for student demographics and previous achievement measures. Standard errors are reported in parentheses. Number of school observations is reported for the 2007–8 school year (year 1).

in effort. Moreover, an overwhelming majority of teachers in schools that met the annual target earned an amount close to $3,000. This is less than 4.1% 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, although 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 teachers in poor-performing schools, including all treatment and control schools in the 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 that 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 solely on the basis of their students' test scores. Teachers whose performance was at lower thresholds could earn $5,000–$10,000. The maximum amount is roughly 31% 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 two or three more 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. Considering the New York results alongside those of the POINT pilot leaves open the possibility that if the magnitude of rewards is, in fact, too small to influence teacher behavior, then teachers' labor supply could be so inelastic that the investment required to influence behavior at the margin may be too large to be relevant to policy discussions.

B. Incentive Scheme Was Too Complex

In the 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 the proficiency level and the median proficiency rating in state tests. Recall that the performance score depends on how a school performs compared

to its peer schools that had a 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, 1 standard deviation.

Similarly, the POINT pilot in Tennessee, like the pilots in 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 (2011). Lack of experimentation with simple incentives schemes like those that have been successful in the developing world makes it impossible to draw definitive conclusions about their effectiveness in American public schools; however, this growing body of international evidence strongly suggests that teachers respond to simple, individualized incentives by modifying behavior in desirable ways. It is plausible that ambiguities in the incentives structures employed in New York and Tennessee may have served to flatten the function that maps effort into expected reward.

C. Group-Based Rewards Are Ineffective

Although schools were given 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 (2011) 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, with an average of 3.28 teachers in each treated school. Similar results are obtained in Glewwe et al. (2010), where the average number of teachers per school was 12. Provided that NYC public schools have 60 teachers on average, the applicability of the results from these analyses is suspect. When there are only three (or 12) teachers in a school, monitoring and penalizing those teachers who shirk their responsibility is less costly. Whatever the mechanism, the success of international incentives programs in smaller schools suggests that individual- or small-group-focused incentives may be effective modifiers of behavior.

However, Lavy (2002) also suggests that group-based incentives may be effective in larger schools. His nonexperimental 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.

D. Teachers Are Ignorant, Not Lazy

Evidence from student incentive experiments reveals that programs that directly incentivize inputs to the education production function are more likely to influence student behavior and improve student learning outcomes than those that incentivize outcome measures like test scores or grades (Fryer 2011). While economic theory predicts that outcome experiments should be more effective, this relies on the premise that students know the best way to improve their academic skills and increase effort efficiently. In reality, despite enthusiasm about the opportunity to profit from their achievement, students reveal ignorance about their own education production function that makes output incentives programs ineffective in terms of improving student achievement.

Teachers who are offered incentives for outcomes that are poorly understood and inherently high variance may face an analogous predicament. Since the teacher incentives that have been tested in American schools uniformly incentivize student learning outputs, teachers are required to know effective strategies for promoting student achievement, which, while it is reasonable to expect professional educators to hold this knowledge, is not necessarily the case. While Duflo and Hanna's (2005) input experiment revealed a profound influence of input incentives on teacher behavior, table 8 reveals no statistically significant changes in teacher behavior in NYC.

So, if teachers are indeed ignorant of the function that maps their own effort into student achievement, how might that affect their performance and, ultimately, student learning? There are two ways I might imagine a teacher to behave in the face of this type of uncertainty: (1) maintain the status quo and hope for the best or (2) invest time and effort into strategies for improving outcomes with unknown expected payoffs. On the one hand, if teachers are ignorant of the education production function, they may not see any value in expending extra effort into a production function whose output they cannot predict. If teachers choose 1 and do not respond to incentives, I would expect to see no statistically significant differences between the treatment and control groups, which is consistent with a wide range of the results.

On the other hand, there remain a surprising number of statistically significant and negative estimates in the middle school cohort that are difficult to explain. One explanation is that teachers are responding to the incentives, but in counterproductive ways. If a teacher invests excess time

into practices or interventions that are less efficient in producing student achievement than their normal practices, I would expect especially motivated and misinformed teachers to overinvest time in ineffective practices at the expense of student learning. In addition, the locus of negative results at the middle school level may suggest that employing new strategies is riskier in a middle school setting.

The most striking evidence against the hypothesis that the results are driven by teachers' lack of knowledge of the production function is driving the results is presented in table 8, which displays treatment effects on five areas of the teacher survey that partly determined 10% of the school's overall progress report score. Even if they are ignorant of the education production function, survey response rates suggest that teachers are not exhibiting behavior that is consistent with a misguided but high effort response to the incentives program.

As before, I 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: safety and respect, communication, engagement, and academic expectations. Questions in the safety and respect section ask whether a school provides a physically and emotionally secure learning environment. 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 academic expectations area measure how well a school develops rigorous academic goals for students. The scores were standardized to have a mean of zero and a standard deviation of 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 2007–8 academic year and 76% in the 2008–9 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

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25, no. 1:95–135.

Baker, George. 2002. Distortion and risk in optimal incentive contracts. *Journal of Human Resources* 37:728–51.

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, no. 2:230–35.

Duflo, Esther, and Rema Hanna. 2005. Monitoring works: Getting teachers to come to school. Working Paper no. 11880, National Bureau of Economic Research, Cambridge, MA.

Firestone, William A., and James R. Pennell. 1993. Teacher commitment, working conditions, and differential incentive policies. *Review of Educational Research* 63, no. 4:489–525.

Fryer, Roland G. 2011. Financial incentives and student achievement: Evidence from randomized trials. *Quarterly Journal of Economics* 126, no. 4:1755–98.

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 Report no. 6319-520, Washington, DC.

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

Hanushek, Eric, and Steven Rivkin. 2005. Teachers, schools and academic achievement. *Econometrica* 73, no. 2:417–58.

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.

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, no. 2:236–40.

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, no. 3:843–77.

Johnson, Susan M. 1984. Merit pay for teachers: A poor prescription for reform. *Harvard Education Review* 54, no. 2:175–85.

Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating teacher impacts on student achievement: An experimental validation. Working Paper no. 14607, National Bureau of Economic Research, Cambridge, MA.

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

———. 2009. Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review* 99, no. 5:1979–2021.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. Teacher performance pay: Experimental evidence from India. *Journal of Political Economy* 119, no. 1:39–77.

Neal, Derek. 2011. The design of performance pay systems in education. Working Paper no. 16710, National Bureau of Economic Research, Cambridge, MA.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73, no. 2:417–58.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94, no. 2:247–52.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2008. Can you recognize an effective teacher when you recruit one? Working Paper no. 14485, National Bureau of Economic Research, Cambridge, MA.

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*. Nashville: National Center on Performance Incentives.

Vigdor, Jacob L. 2008. Teacher salary bonuses in North Carolina. Working Paper no. 15, National Center for Analysis of Longitudinal Data in Education Research, Durham, NC.