Financial Incentives and Student Achievement: Evidence from Randomized Trials

Roland G. Fryer, Jr.*
Harvard University, EdLabs, and NBER

April 8, 2010

*This project would not have been possible without the leadership and support of Eli Broad, Arne Duncan, Steven Hyman, Joel Klein, Thelma Morris-Lindsey, Michelle Rhee, and Lawrence Summers. I am also grateful to our district partners: Jennifer Bell-Ellwanger, Joanna Cannon, Dominique West (NYC); Erin Goldstein, Abigail Smith, Hella Bel Hadj Amor (Washington, DC); Amy Nowell, Asher Karp, John Jablonski (Chicago); and David Vines and Jane Didear (Dallas), for their endless cooperation in collecting the data necessary for this project. I am indebted to my colleagues Josh Angrist, Michael Anderson, Paul Attewell, Roland Benabou, David Card, Raj Chetty, Andrew Foster, Edward Glaeser, Richard Holden, Lawrence Katz, Gary King, Nonie Lesaux, Steven Levitt, John List, Glenn Loury, Franziska Michor, Peter Michor, Richard Murnane, Derek Neal, Ariel Pakes, Eldar Shafir, Andrei Shleifer, Chad Syverson, Petra Todd, Kenneth Wolpin, and Nancy Zimmerman, along with seminar participants at Brown, CIFAR, Harvard (Economics and Applied Statistics), Oxford, and University of Pennsylvania for helpful comments, discussions, advice, and support during this research experiment. The seeds of this project were planted in 2003 in a joint venture between the author, Alexander Gelber, and Richard Freeman in P.S. 70 in the Bronx, NY. Brad Allan, Austin Blackmon, Charles Campbell, Melody Casagrande, Theodore Chang, Vilsa E. Curto, Nancy Cyr, Will Dobbie, Katherine Ellis, Peter Evangelakis, Richard Hagey, Meghan L. Howard, Lindsey Mathews, Kenneth Mirkin, Eric Nadelstern, Aparna Prasad, Gavin Samms, Evan Smith, Jörg Spenkuch, Zachary D. Tanjeloff, David Toniatti, Rucha Vankudre, and Carmita Vaughn provided brilliant research assistance and project management and implementation support. Financial Support from the Broad Foundation, District of Columbia Public Schools, Harvard University, Joyce Foundation, Mayor's Fund to Advance New York City, Pritzker Foundation, Rauner Foundation, Smith Richardson Foundation, and Steans Foundation is gratefully acknowledged. Many thanks to our bank partners Chase, Sun Trust and Washington Mutual for their support and collaboration on this project. Correspondence can be addressed to the author by mail: Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge, MA, 02138; or by email: rfryer@fas.harvard.edu. The usual caveat applies.

Abstract

This paper describes a series of school-based randomized trials in over 250 urban schools designed to test the impact of financial incentives on student achievement. In stark contrast to simple economic models, our results suggest that student incentives increase achievement when the rewards are given for inputs to the educational production function, but incentives tied to output are not effective. Relative to popular education reforms of the past few decades, student incentives based on inputs produce similar gains in achievement at lower costs. Qualitative data suggest that incentives for inputs may be more effective because students do not know the educational production function, and thus have little clue how to turn their excitement about rewards into achievement. Several other models, including lack of self-control, complementary inputs in production, or the unpredictability of outputs, are also consistent with the experimental data.

1 Introduction

The United States is the richest country in the world, but American ninth graders rank 28th in math, 22nd in science, and 18th in reading achievement.[1] Seventy percent of American students graduate from high school, which ranks the United States in the bottom quartile of OECD countries (Education at a Glance 2007). In large urban areas with high concentrations of blacks and Hispanics, educational attainment and achievement are even more bleak, with graduation rates as low as thirty-one percent in cities like Indianapolis (Swanson, 2009). The performance of black and Hispanic students on international assessments is roughly equal to national scores in Mexico and Turkey – two of the lowest performing OECD countries.

In an effort to increase achievement and narrow differences between racial groups, school districts have become laboratories of innovative reforms.[2] One potentially cost-effective strategy, which has yet to be tested in urban public schools, is providing short-term financial incentives for students to achieve or exhibit certain behaviors correlated with student achievement.[3] Theoretically, providing such incentives could have one of three possible effects. If students lack sufficient motivation, dramatically discount the future, or lack accurate information on the returns to schooling to exert optimal effort, providing incentives for achievement will yield increases in student performance.[4] If

[1] Author's calculations based on data from the 2003 Program for International Student Assessment, which contains data on forty-one countries including all OECD countries.

[2] These reforms include smaller schools and classrooms (Nye et al., 1995; Krueger, 1999), mandatory summer school (Jacob and Lefgren, 2004), merit pay for principals and teachers (Podgursky and Springer, 2007), after-school programs (Lauer et al., 2006), budget, curricula, and assessment reorganization (Borman et al., 2007), policies to lower the barrier to teaching via alternative paths to accreditation (Decker, Mayer, and Glazerman, 2004; Kane, Rockoff, and Staiger, 2008), single-sex education (Shapka and Keating, 2003), data-driven instruction (Datnow, Park, and Kennedy, 2008), ending social promotion (Greene and Winters, 2006), mayoral/state control of schools (Wong and Shen, 2002, 2005; Henig and Rich, 2004), instructional coaching (Knight, 2009), local school councils (Easton et al., 1993), reallocating per-pupil spending (Marlow, 2000; Guryan, 2001), providing more culturally sensitive curricula (Protheroe and Barsdate, 1991; Thernstrom, 1992; Banks, 2001, 2006), renovated and more technologically savvy classrooms (Rouse and Krueger, 2004; Goolsbee and Guryan, 2006), professional development for teachers and other key staff (Boyd et al., 2008; Rockoff, 2008), and increasing parental involvement (Domina, 2005).

[3] Many parents, teachers, public schools, and high-achieving charter schools [Knowledge is Power Program (KIPP) and Harlem Children's Zone, for example] use some form of incentive program in their schools. This paper is the first large scale intervention project designed to test the effect of student incentives on achievement in urban public schools in America.

[4] Economists estimate that the return to an additional year of schooling is roughly ten percent and, if anything,

students lack the structural resources or knowledge to convert effort to measurable achievement or if the production function has important complementarities out of their control (effective teachers, engaged parents, or peer dynamics, e.g.) then incentives will have very little impact. Third, some argue that financial rewards for students (or any type of external reward or incentive) will undermine intrinsic motivation and lead to negative outcomes.[5] Which one of the above effects – investment incentives, structural inequalities, or intrinsic motivation – will dominate is unknown. The experimental estimates obtained will combine elements from these and other potential channels.

In the 2007-2008 and 2008-2009 school years, we conducted incentive experiments in public schools in Chicago, Dallas, New York City, and Washington, DC – four prototypically low performing urban school districts – distributing a total of $6.3 million to roughly 38,000 students in 261 schools.[6] All experiments were school-based randomized trials. The experiments varied from city to city on several dimensions: what was rewarded, how often students were given incentives, the grade levels that participated, and the magnitude of the rewards.[7] The key features of each experiment consisted of monetary payments to students (directly deposited into bank accounts opened for each student or paid by check to the student) for performance in school according to a simple incentive scheme. In all cities except Dallas, where students were paid three times a year, payments were disseminated to students within days of verifying their achievement.[8]

is higher for black students relative to whites (Card, 1999; Neal and Johnson, 1996; Neal, 2005). Short-term financial incentives may be a way to straddle the perceived cost of investing in human capital now with the future benefit of investment.

[5]There is an active debate in psychology as to whether extrinsic rewards crowd out intrinsic motivation. See, for instance, Deci (1972), Deci (1975), Kohn (1993), Kohn (1996), Gneezy and Rustichini (2000), or Cameron and Pierce (1994) for differing views on the subject.

[6]Throughout the text, I depart from custom by using the terms "we," "our," and so on. While this is a sole-authored work, it took a large team of people to implement the experiments. Using "I" seems disingenuous.

[7]There are approximately four to five articles per day in major newspapers written about NYC public schools. Given this media scrutiny and the sensitive nature of paying students to learn, we were unable to design more elaborate experiments with many treatment arms within a single city. This approach is possible in development economics [see Bertrand et al. (2009) for a good example]. Thus, to obtain important variation, we designed experiments that spanned multiple U.S. cities. The natural desire of local governments to tweak experiments being planned in other cities and to "own" a unique twist led to our pseudo-planned variation. This is not ideal. Future experiments may be able to provide important treatment variation within a city.

[8]There was a vast and coordinated implementation effort among twenty project managers to ensure that students, parents, teachers, and key school staff understood the particulars of each program; that the program was implemented

Traditional price theory, under a simple set of assumptions, predicts that providing incentives based on output is socially optimal.[9] The key idea is that students know the mapping from the vector of inputs to output and differ in their marginal returns across inputs. Incentives for inputs operate as price subsidies for those particular inputs. Incentives for output also operate as a price subsidy, but allow each student to decide which input from their production function to subsidize. Since students are assumed to have superior knowledge about how they learn, it is socially optimal to allow them to allocate their time across inputs. However, if this simple set of assumptions is violated (risk aversion, noisy output, or if students only have a vague idea of how to produce output, e.g.), then it can be more effective to provide incentives for inputs. Understanding whether incentives for inputs or outputs are more effective in increasing student achievement is of great importance to education policy makers and researchers as they build a framework to understand the economics of incentive-based education reform. This is the spirit in which we designed our set of experiments. The programs in Chicago and New York City are "output" experiments, while the programs in Dallas and Washington, DC, are "input" experiments.

In NYC, we paid fourth and seventh grade students for performance on a series of ten interim assessments currently administered by the NYC Department of Education to all students. In Chicago, we paid ninth graders every five weeks for grades in five core courses. In Dallas, we paid second graders $2 per book to read and pass a short quiz to confirm they read it. In the District of Columbia, we provided incentives for sixth, seventh, and eighth grade students on a series of five metrics that included attendance, behavior, and three inputs to the production function chosen by each school individually.

The results from our incentive experiments are interesting and in some cases quite surprising. Remarkably, incentives for output did not increase achievement. Paying students for performance on standardized tests yielded treatment effects for seventh graders between -.018 (.035) and -.030 (.063) standard deviations in mathematics and .018 (.018) and .033 (.032) standard deviations in reading. The programs in which fourth graders were paid for their test scores exhibited similar results. Rewarding ninth graders for their grades yielded increases in their grade point average of .093 (.057) and .131 (.078), but had no effect on achievement test scores in math or reading.

with high fidelity; and that payments were distributed on time and accurately.

[9] In the classic principal-agent framework, it is assumed that the agents' actions are not contractible, rendering moot the decision between inputs and outputs (Mirrlees, 1974; Holmstrom, 1979; Grossman and Hart, 1983).

Conversely, incentives can be a cost-effective strategy to raise achievement among even the poorest minority students in the lowest performing schools if the incentives are given for certain inputs to the educational production function. Paying students to read books yields a large and statistically significant increase in reading comprehension between .180 (.075) and .249 (.103) standard deviations, increases vocabulary between .051 (.068) and .071 (.093) standard deviations, and increases language between .136 (.080) and .186 (.107) standard deviations. The estimated impacts on vocabulary scores are not significant; increases in language are marginally significant. Similarly, paying students for attendance, good behavior, wearing their uniforms, and turning in their homework increases reading achievement between .152 (.092) and .179 (.106) standard deviations, and increases mathematics achievement between .114 (.106) and .134 (.122) standard deviations. The point estimates are moderate in size, but we do not have enough statistical power to make confident conclusions. The effects of incentives in Washington, DC, on reading achievement are marginally significant in reading and statistically insignificant in math.

A central question in the study of incentives is what happens when the incentives are taken away. Many believe that students will have decreased intrinsic motivation and that their achievement will be negative once the incentives are discontinued (Kohn, 1993 and references therein). Contrary to this view, the point estimate a year after the Dallas experiment is roughly half of the original effect in reading and larger in math, but not statistically significant. The finding for reading is similar to the classic "fade out" effect which has been documented in other successful interventions, such as Head Start, a high quality teacher for one year, or a smaller class size (Nye, Hedges, and Konstantopoulos, 1999; Puma et al., 2010).

We also investigate treatment effects across a range of predetermined subsamples – gender, race, previous year's performance and behavior, and an income proxy. In cities where incentives increased student achievement – Dallas (reading books) and Washington, DC (attendance, behavior, etc) – boys gained more from the experiment than girls. Partitioning the data by race shows that Hispanics gained substantially throughout the input experiments. Neither Asians nor whites did especially well, though the effects on these racial groups are measured imprecisely due to small numbers. Students eligible for free lunch, a typical proxy for poverty, gained less than students not on free lunch. Splitting the data by previous year's achievement shows no particular patterns. Dividing the sample by previous year's behavioral incidents reveals that students in the Washington, DC,

treatment with previously bad behavior show large treatment effects [.400 (.235) standard deviations in reading and .164 (.274) standard deviations in math], but these are measured with considerable error. The program in Washington, DC, was the only treatment that contained incentives for good behavior.

We conclude our statistical analysis by estimating how incentives for student achievement affect alternative outcomes, effort, and intrinsic motivation. Paying students to read books has positive spillovers on their course grades and a positive but statistically insignificant effect on their math test scores. Incentives for grades in core courses cause an increase in attendance and students pass, on average, almost one more course during their freshman year. Providing incentives for achievement test scores has no effect on any form of achievement we can measure. Across all cities, there is scant evidence that total effort increased in response to the programs, though there may be substitution between tasks. Finally, using the Intrinsic Motivation Inventory developed in Ryan (1982), we find no evidence that incentives decrease intrinsic motivation. The signs on the coefficients are seductive – input experiments seem positively associated with motivation and output experiments seem negative. However, the point estimates are too small and the standard errors are too large to conclude anything other than a null effect.

In summary, we find that relative to achievement-increasing education reform in the past few decades – Head Start, lowering class size, bonuses for effective teachers to teach in high need schools – student incentives for certain inputs provide similar results at lower cost. Yet, incentives alone, like these other reforms, are not powerful enough to close the achievement gap.

Finding the correct interpretation for our set of experiments is difficult. Much depends on the interpretation of the results from Washington, DC. The leading theory is that students do not understand the educational production function and, thus, lack the know-how to translate their excitement about the incentive structure into measurable output.[10] Students who were paid to read books, attend class, or behave well did not need to know how the vector of potential inputs relates to output, they simply needed to know how to read, make it to class, or sit still long enough to collect their short-term incentive.

There are three pieces of evidence that support this theory. First, evidence from our qualitative

[10]We characterize this theory as "leading" because it is the only theory confirmed by significant qualitative observations.

team found consistent narratives suggesting that the typical student was elated by the incentive but did not know how to turn that excitement into achievement.[11] Second, focus groups in Chicago confirmed this result; students had only a vague idea how to increase their grades. Third, there is evidence to suggest that some students – especially those who are in the bottom of the performance distribution – do not understand the production function well enough to properly assess their own performance, let alone know how to improve it (Kruger and Dunning, 1999).

Three other theories are also consistent with the experimental data. It is plausible that students know the production function, but that they lack self-control or have other behavioral tendencies that prevent them from planning ahead and taking the intermediate steps necessary to increase the likelihood of a high test score in the future. A second competing theory is that the educational production function is very noisy and students are sufficiently risk averse to make the investment not worthwhile. A final theory that fits our set of facts is one in which complementary inputs (effective parents, e.g.) are responsible for the differences across experiments.

Though incentives for student performance are considered questionable by many, there is a nascent but growing body of scholarship on the role of incentives in primary, secondary, and post-secondary education around the globe (Angrist et al., 2002; Angrist and Lavy, 2009; Kremer, Miguel, and Thornton, 2004; Behrman, Sengupta, and Todd, 2005; Angrist, Bettinger, and Kremer, 2006; Angrist, Lang, and Oreopoulos, 2006; Barrera-Osorio et al., 2008; Bettinger, 2008; Hahn, Leavitt, and Aaron, 1994; Jackson 2009).

The paper is structured as follows. Section 2 provides some details of our experiments and their implementation in each city. Section 3 describes our data, research design, and econometric framework. Section 4 presents estimates of the impact of financial incentives on student achievement. Section 5 presents estimates of the impact of financial incentives on alternative forms of achievement, effort, and intrinsic motivation. Section 6 interprets the results through the lens of economic theory. Section 7 concludes. There are two appendices. Appendix A is an implementation supplement that provides details on the timing of our experimental roll-out and critical milestones reached. Appendix B is a data appendix that provides details on how we construct our covariates and our samples from the school district administrative files used in our analysis.

[11] The qualitative team was led by Paul Attewell and consisted of seven full-time qualitative researchers who observed twelve students and their families, as well as ten classrooms in NYC.

2 Program Details

Table 1 provides an overview of each experiment and specifies conditions for each site. See Appendix A for further implementation and program details.

In total, experiments were conducted in 261 schools across four cities, distributing $6.3 million to 38,419 students.[12] All experiments had a similar roadmap to launch. First, we garnered support from the district superintendent. Second, a letter was sent to principals of schools that served the desired grade levels. Third, we met with principals to discuss the details of the programs. In New York, these meetings largely took place one school at a time; in the other three cities large meetings were assembled at central locations. After principals were given information about the experiment, there was a sign-up period. Schools that signed up to participate serve as the basis for our randomization. All randomization was done at the school level. After treatment and control schools were chosen, treatment schools were alerted that they would participate and control schools were informed that they were first in line if the program was deemed successful and continued beyond the experimental years. In each school year, students received their first payments the second week of October and their last payment was disseminated over the summer. All experiments lasted at least one full school year.

Dallas

Dallas Independent School District (DISD) is the 14th largest school district in the nation with 159,144 students. Over 90 percent of DISD students are Hispanic or black. Roughly 80 percent of all students are eligible for free or reduced lunch and roughly 25 percent of students have limited English proficiency.

Forty-three schools signed up to participate in the Dallas experiment, and we randomly chose twenty-two of those schools to be treated (more on our randomization procedure below). The treatment sample was comprised of 3,788 second grade students. To participate, students were required to have a parental consent form signed; eighty percent of students in the treatment sample signed up to participate. Participating schools received $1500 to lower the cost of implementation.

Students were paid $2 per book read for up to 20 books per semester. Upon finishing a book, students took an Accelerated Reader (AR) computer-based comprehension quiz which provided

[12] Roughly half the students and half the schools were assigned to treatment and the other half to control.

evidence as to whether the student read the book. A score of eighty percent or better on each book quiz earned each student a $2 reward. Quizzes were available on 80,000 trade books, all major reading textbooks, and the leading children's magazines. Students were allowed to select and read books of their choice at the appropriate reading level and at their leisure, not as a classroom assignment. The books came from the existing stock available at their school (in the library or in the classroom). Three times a year (twice in the fall and once in the spring) teachers in the program tallied the total amount of incentive dollars earned by each student based on the number of passing quiz scores. A check was then written to each student for the total amount of incentive dollars earned. The average student received $13.81, the maximum $80 – with a total of $42,800 distributed to students.

New York City

New York City is the largest school district in the United States and one of the largest school districts in the world – serving 1.1 million students in 1,429 schools. Over seventy percent of NYC students are black or Hispanic, fifteen percent are English language learners, and over seventy percent are eligible for free lunch.

One hundred and forty-three schools signed up to participate in the New York City experiment, and we randomly chose sixty-three schools (thirty-three fourth grades and thirty-one seventh grades) to be treated.[13] The treatment sample consisted of a total of 8,176 total students. Participating schools received $2500 if eighty percent of eligible students were signed up to participate and if the school had administered the first four assessments. The school received another $2500 later in the year if eighty percent of students were signed up and if the school had administered six assessments.

Students in the New York City experiment were given incentives for their performance on six computerized exams (three in reading and three in math) as well as four predictive assessments that were pencil and paper tests.[14] For each test, fourth graders earned $5 for completing the exam and $25 for a perfect score. The incentive scheme was strictly linear – each marginal increase in score was associated with a constant marginal benefit. A fourth grader could make up to $250 in a school

[13] Grades and schools do not add up because there is one treatment schools that contained both fourth and seventh grades and both grades participated.

[14] All schools had a computer version of the predictive assessments, but very few schools exercised that option because of the burden of moving classes in and out of relatively small computer labs or slow Ethernet connections.

year. The magnitude of the incentive was doubled for seventh graders: $10 for completing each exam and $50 for a perfect score – yielding the potential to earn $500 in a school year. To participate, students were required to turn in a signed parental consent form; eighty-two percent signed up to participate. The average fourth grader earned $139.43 and the highest earner garnered $244. The average seventh grader earned $231.55 and the maximum earned was $495. Approximately sixty-six percent of students opened a student savings account with Washington Mutual as part of the experiment and money was directly deposited into these accounts. Certificates were distributed in school to make the earnings public. Students who did not participate because they did not return a consent form took identical exams but were not paid. To assess the quality of our implementation, schools were instructed to administer a short quiz to students that tested their knowledge of the experiment; ninety percent of students understood the basic structure of the incentive program. See Appendix A for more details.

Washington, DC

The third experiment on financial incentives took place in Washington, DC – the school district with the second lowest overall achievement in the country on the National Association of Education Progress (NAEP) assessments. According to NAEP, 4.5 percent of Washington, DC, middle school students score at or above proficient in math and 7.6 percent score at or above proficient in reading. The district is comprised of 92.4 percent blacks and Hispanics; 70 percent of students are eligible for free or reduced lunch.

Washington, DC, is a relatively small school district, containing only thirty-five schools with middle school grades. Thirty-four schools signed up to participate in the experiment and we randomly selected seventeen of them to be treated. The remaining seventeen schools served as control schools. Students in treatment schools were given incentives for five inputs to the educational production function. We mandated that schools include attendance and behavior as two of the five metrics. Each school was allowed to pick the remaining three, with substantial input from our implementation team – we directed them to concentrate on interim achievement metrics.[15] Finalized metrics differed from school to school but a typical scheme included metrics for attendance,

[15]The intuition of Chancellor Rhee and several school principals suggested that schools possessed asymmetric information on what should be incentivized so we wanted to provide some freedom in choosing metrics.

behavior, wearing a school uniform, homework, and classwork.[16]

Incentives were given on a point system – students were given one point every day for satisfying each of the five metrics. At the end of each two-week pay period, students could earn up to fifty points (five metrics, 10 school days). Students earned $2 per point and the money was distributed into Sun Trust Bank Accounts or paid by check. Sixty-six percent of kids opened up an account as part of the experiment and the remaining one-third received checks in intervals left up to the school's discretion.[17] The average student earned approximately $40 every two weeks; $532.85 for the year. The highest amount received was $1,322. Each participating school received stipends for participation in the program based on the number of students in their school. Amounts were determined by current negotiated overtime rates and ranged from $2,200 in small schools to $13,000 in the largest schools. The incentive schemes in Washington, DC, were the most complicated, but 86.2 percent of students scored ninety percent or higher on a test administered to assess their understanding of the basic structure of the program.

Chicago

The Chicago experiment took place in twenty low-performing Chicago Public High Schools. Chicago is the third largest school district in the US with over 400,000 students, 88.3 percent of whom are black or Hispanic. Seventy-five percent of students in Chicago are eligible for free or reduced lunch, and 13.3 percent are English language learners.

Seventy schools signed up to participate in the Chicago experiment. To control costs, we selected forty of the smallest schools out of the seventy who wanted to participate and then randomly selected twenty to treat within this smaller set. Once a school was selected, students were required to return a signed parental consent form to participate. The set of students eligible to be treated consisted of 4,396 ninth graders. Ninety-one percent of eligible students signed up. Participating schools received up to $1,500 to provide a bonus for the school liaison who served as the main contact for our implementation team.

Students in Chicago were given incentives for their grades in five core courses: English, mathe-

[16] Detailed metrics for each school are available from the author upon request.

[17] Everyone received checks for the first two payments because SunTrust was still in the process of setting up bank accounts. After that point, it was up to schools to pick up and distribute checks every two weeks and they had the discretion to give out checks later to encourage students to open bank accounts. Checks were processed every two weeks to coincide with direct deposits.

matics, science, social science, and gym.[18] We rewarded each student with $50 for each A, $35 for each B, $20 for each C, and $0 for each D. If a student failed a core course, she received $0 for that course and temporarily "lost" all other monies earned from other courses in the grading period. Once the student made up the failing grade through credit recovery, night school, or summer school all the money "lost" was reimbursed. Students could earn $250 every five weeks and $2,000 per year. Half of the rewards were given immediately after the five-week grading periods ended and the other half is being held in an account and will be given in a lump sum conditional on high school graduation. The average student earned $695.61, the highest achiever earned $1,875.

3 Data, Research Design, and Econometric Model

We collected both administrative and survey data. The richness of the administrative data varies by school district, but includes information on each student's first and last name, birth date, address, race, gender, free lunch eligibility, behavioral incidents, attendance, matriculation with course grades, special education status, and English Language Learner (ELL) status. In Dallas and New York, we are able to link students to their classroom teachers. New York City administrative files contain teacher value-added data for teachers in grades four through eight.

Our main outcome variable is an achievement test unique to each city. We did not provide incentives of any form for these assessments.[19] In May of every school year, English-speaking students in Dallas public elementary schools take the Iowa Tests of Basic Skills (ITBS) if they are in kindergarten, first, or second grade. Spanish-speaking students in Dallas take a different exam labeled Logramos.[20] New York City administers mathematics and English Language Arts tests, developed by McGraw-Hill, in the winter for students in third through eighth grade. The Washington, DC Comprehensive Assessment System (DC-CAS) is administered each April to students in grades three through eight and ten. All Chicago tenth graders take the PLAN assessment, an ACT college readiness exam, in October. See Appendix B for more details.

We use a parsimonious set of controls to aid in precision and to correct for any potential im-

[18] Gym may seem like an odd core course to provide incentives for achievement, but roughly twenty-two percent of ninth grade students failed their gym course in the year prior to our experiment.

[19] We wanted an objective measure of achievement without the influence of incentives.

[20] The Spanish test is not a translation of ITBS. Throughout the text, we present estimates for these tests separately. The score distributions are too different to consider these results together.

balance between treatment and control. The most important controls are test scores from previous years, which we include in all regressions along with their squares. Previous year's test scores are available for most students who were in the district in the previous year (see Appendix Tables 1A through 1F for exact percentages of experimental group students with valid test scores from previous years). We include the previous two years of achievement tests in math and reading.[21] 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 if not.

Other individual-level controls include a mutually exclusive and collectively exhaustive set of race dummies pulled from each school district's administrative files, indicators for free lunch eligibility, special education status, and whether a student is an English language learner. 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. Determination of special education and ELL status varies by district. For example, in Washington, DC, special education status is determined through a series of observations, interviews, reviews of report cards, and administration of tests. In Dallas, any student who reports that his or her home language is not English is administered a test and ELL status is based on the student's score on that test.

We also construct three school-level control variables: the percentage black, percentage Hispanic, and percentage of students eligible for free lunch of the school's student body. To construct school-level variables, we construct demographic variables for every student in the district enrollment file in the experimental year and then take the mean value of these variables for each school. In Dallas, New York, and Washington, DC, we assign each student who was present at the beginning of the year, i.e. before October 1, to the first school that they attended. We assign anyone

[21] For fourth graders in New York, we only include one previous year test score because New York State assessments begin in third grade.

who moved into the school district at a later date to the school that they attended for the longest period of time. For Chicago, we are unable to determine exactly when students move into the district. Therefore, we assign each student in the experimental sample to the school that they attended first, and we assign everyone else to the school that they attended for the longest period of time. We construct the school-level variables for each city based on these school assignments.

To supplement each district's administrative data, we administered a survey in each of the four school districts. The data include basic demographics of each student such as family structure and parental education, time-use, effort and behavior in school, and (most importantly) the Intrinsic Motivation Inventory described in Ryan (1982).

Survey administration in Dallas and Washington, DC, went relatively smoothly. We offered up to $2,000 (pro-rated by size) for schools in which ninety percent or more of the surveys were completed. Eighty percent of surveys were returned in Dallas treatment schools and eighty-nine percent were returned in control schools. In Washington, DC, seventy-three percent of surveys were returned in treatment schools and seventy-one percent in control schools. For surveys administered in urban environments, these response rates are high (Parks, Housemann, and Brownson, 2003; Guite, Clark, and Ackrill, 2006).

In the two other cities, survey responses were low. In Chicago, despite offering $1,000 per school to schools for collecting ninety percent of the surveys, only thirty-five percent of surveys were returned in treatment schools and thirty-nine percent in control schools.[22] In New York City, survey administration was the most difficult. The New York City Institutional Review Board did not allow us to provide any sort of incentives for students or schools to turn in surveys. We were able to offer $500 to schools to administer the survey and could not condition the payment on survey response rate. Further, the review board insisted that students turn in an additional parental consent form for the surveys only. Only fifty-eight percent of surveys were returned in the treatment group and twenty-seven percent were returned by students in control schools.

Appendix Tables 1A through 1F provide pre-intervention descriptive statistics for Dallas (A and B), New York City fourth graders (C), New York City seventh graders (D), Washington, DC

[22] Months after Arne Duncan resigned as the CEO of Chicago Public Schools to become Secretary of Education in the Obama administration and the two-year experiment was shortened to one year, support for the program dwindled. Half of the surveys were lost. We do not report survey results from Chicago.

(E), and Chicago (F). In each table, the first three columns provide the mean, standard deviation, and number of observations for each variable used in our analysis for the entire treatment and control sample. See Appendix B for details on how each variable was constructed. The second set of numbers provides the mean, standard deviation, and number of observations for the same set of variables for our set of treatment schools. The last set of numbers provides identical data for our set of control schools.

Research Design

In designing a randomized procedure to partition our sets of interested schools into treatment and control schools, our main constraints were political. For instance, one of the reasons we randomized at the school level in every city was the political sensitivity of rewarding some students in a grade for their achievement and not others.[23] We were also asked to not implement our program in schools that were mayoral priorities for other initiatives. We used the same procedure in each city to randomly partition the set of interested schools into treatment and control schools.

The goal of any randomization is to have the most balanced sample possible across treatment and control schools on observables and unobservables. The standard method to check whether a randomization was successful is to estimate regressions such as:

$$Treatment_s = \alpha + X_s\beta + \varepsilon_s \tag{1}$$

where s represents data measured at the school level. The dependent variable takes on the value of one for all individuals in treatment schools. The number of treatment and control schools range from 34 (Washington, DC) to 143 (New York City), which is consistent with the number of groups in typical Group Randomized Trials (Donner, Brown, and Brasher, 1990; Feng et al., 2001; Angrist and Lavy, 2009), though Washington, DC, is smaller than usual.

Recall that we randomized among all schools that previously expressed interest in participating. Suppose there are X schools who are interested in participating and we aim to have a treatment group of size Y. Then, there are X choose Y potential permutations. From this enormous set of possibilities – 2.3 billion in Washington, DC, and 2.113×10^{41} in New York – we randomly selected 10,000 treatment-control designations and estimated equation (1) in each city for each possible

[23] We were also concerned that randomizing within schools could prompt some teachers to provide alternative non-monetary incentives to control students (unobservable to us) that would undermine the experiment.

randomization.[24] We then selected the randomization that minimized the z-scores from the probit regression.[25]

Appendix Table 2 present the results of our school-based randomization from each city. The column under each city includes all variables we have collected and which we will use as controls in the forthcoming analysis. Data vary by city according to availability. The model estimated is a linear regression identical to equation (1).

In Dallas, treatment and control schools are fairly balanced. In New York, we had a larger number of treatment and control schools than in other city and the samples are also very balanced. In fourth grade, there are negligible differences between treatment and control in the percent of special education students [.030 (.014)] and previous year's reading score [.021 (.010)]. In seventh grade, the sample is similarly balanced. Treatment schools in Chicago have slightly higher previous year's math scores [.100 (.041)], and slightly lower reading scores than control schools [-.115 (.053)]. Washington, DC, the smallest of the four school districts, had only thirty-four (out of thirty-five possible) schools with grades six through eight that signed up to participate. As a result, our independent variables are not as balanced across treatment and control as the other cities. Schools in the treatment group have significantly higher proportions of blacks, Hispanics, and students who report to be "other race," and are much less likely to be elementary schools. The latter may be particularly important, as there is evidence that kindergarten through eighth grade schools are more effective at increasing the test scores of their students when they enter high school compared to similar middle school students (Offenberg, 2001).[26]

To complement the linear regressions described above, Appendix Figures 1A and 1B show the geographic distributions of treatment and control schools and their census tract poverty rates across our respective cities. These maps confirm that our schools are similarly distributed across space and are more likely to be in higher poverty areas of a city.

[24]There is an active debate on which randomization procedures have the best properties. Karlan and Valdivia (2006) prefer the method adopted here. Kling, Liebman, and Katz (2007) suggest matched pairs. See Bruhn and McKenzie (2009) for a review of the issues.

[25]In NYC, we had an additional constraint that no treatment schools could contain participants in Opportunity NYC, an incentive program implemented at the same time that provided rewards for various activities (http://opportunitynyc.org/), which forced us to choose the sixth best randomization.

[26]A joint significance test for each city yields an F-statistic of 3.57 in Dallas, 1.84 in NYC fourth grade, 1.01 in NYC seventh grade, 2.88 in Chicago, and 7.24 in Washington, DC.

Econometric Models

To estimate the causal impact of providing student incentives on outcomes, we use three statistical models. We begin by estimating intent-to-treat (ITT) effects, i.e. differences between treatment and control group means. Let Z_s be an indicator of treatment school assignment, let X_i be a vector of baseline covariates measured at the individual level, and let X_s denote school-level variables; X_i and X_s comprise our parsimonious set of controls. The ITT effect, π_1, is estimated from the equation below:

$$achievement_{i,s} = \alpha_1 + Z_s\pi_1 + X_i\beta_1 + X_s\gamma_1 + \varepsilon_{1i,s} \tag{2}$$

The ITT is an average of the causal effects for students in schools who were randomly selected for treatment at the beginning of the year and students in schools that signed up for treatment but were not chosen. In other words, ITT provides an estimate of the impact of being offered a chance to participate in a financial incentive program. All student mobility between schools after random assignment is ignored. We only include students who were in treatment and control schools as of October 1 in the year of treatment.[27] For most districts, school begins in early September; the first student payments were distributed mid-October. All standard errors, throughout, are clustered at the school level.

Under several assumptions (that the treatment group assignment is random, control schools are not allowed to enroll in the incentive program, and that being selected for the incentive program only affects outcomes through the use of incentives), we can also estimate the causal impact of actually participating in the incentive program. This parameter, commonly known as the "Treatment-on-the-Treated" (TOT) effect, measures the average effect of being in a school which was assigned treatment for participating students. The TOT parameter can be estimated through a two-stage least squares regression of student achievement on fraction of the school year that the student is signed up to participate in a treatment school (*Fraction of Year in Treatment*$_i$) using initial random assignment, (Z_s), as an instrumental variable for the fraction of the school year treated:

$$achievement_{i,s} = \alpha_2 + Fraction\ of\ Year\ in\ Treatment_i \cdot \pi_2 + X_i\beta_2 + X_s\gamma_2 + \varepsilon_{2i,s}. \tag{3}$$

[27] This is due to a limitation of the attendance data files in Chicago. In other cities, the data are fine enough to only include students who were in treatment on the first day of school. Using the first day of school or October 1 does not alter the results (data not shown).

The TOT is the estimated difference in outcomes between students who participate in treatment schools and those in the control group who would have participated if given the chance.

Mobility is common in poorly performing urban schools. A key concern in estimating the TOT is how to account for students who enter or exit a treatment school during the school year. If attrition is non-random, the TOT estimates may be biased if we estimate the effect on current participants. To sidestep issues concerning endogenous mobility, we restrict our sample to students who were in treatment and control schools when the randomization took place, ignore all students who enter treatment and control after that date, and follow experimental students who leave experimental schools but remain in the school district. This approach ensures that our ITT and TOT samples are identical.[28]

In New York City, seven schools were switched from control to treatment (three in fourth grade and four in seventh grade), and thus violate the technical assumptions needed to credibly estimate TOT.[29] Thus, in lieu of TOT estimates in New York, we will provide local average treatment effects (LATE) – our third statistical model. LATE is the estimated difference in outcomes between students who participate in participating schools and those in non-participating schools who would have participated if given the chance. That is, we switch those seven schools from control to treatment when we estimate LATE.

4 The Impact of Financial Incentives on Student Achievement

Table 2 presents ITT, TOT, and LATE estimates for our output experiments. The first four columns in Table 2 present estimates from our experiments in New York; the final two columns provides results from Chicago. The odd-numbered columns in Table 2 report ITT estimates; even numbered columns report LATEs (in New York) and TOTs (in Chicago).

The impact of offering incentives to students for test scores or grades is statistically zero and substantively small in all specifications. For fourth graders, the ITT estimate of the effect of being

[28] If we allow entry and exit into treatment schools (some students enter and exit treatment schools multiple times a year), the estimates are .04 to .07 larger than those reported.

[29] Six of these cases were due to the fact that these schools contained grades kindergarten through eight. The principals insisted on having both grades in treatment if the other was in control. The remaining school was moved from control to treatment for political reasons.

offered an incentive program which pays students for test scores is -.023 (.034) standard deviations without controls and -.021 (.033) standard deviations including our set of controls for reading. For math, the results are .052 (.046) standard deviations without controls and .067 (.046) standard deviations with controls for math. The estimates of the local average treatment effect (LATE) are similar. We find a treatment effect of -.036 (.051) standard deviations in reading and .092 (.070) standard deviations in mathematics with our standard set of controls.

The results are similar for seventh graders. The raw ITT for mathematics is .008 (.048) standard deviations and -.018 (.035) standard deviations with controls. The raw ITT for reading is .040 (.036) standard deviations and .018 (.018) standard deviations with controls. Our LATE estimates are similarly small, -.030 (.063) standard deviations for math and .033 (.032) standard deviations for ELA with controls.

The final two columns in Table 2 provide estimates of the causal impact of providing incentives for course grades on achievement in Chicago. The estimates are statistically zero and substantively small in all specifications. The raw ITT for mathematics is -.030 (.031) standard deviations and -.010 (.023) standard deviations with controls. TOT estimates are very similar. The raw ITT for reading is -.027 (.044) standard deviations and -.006 (.027) standard deviations with controls. Our TOT estimates are also small, -.035 (.056) standard deviations without controls and -.008 (.035) standard deviations with controls. Paying students for grades does not increase their achievement. This may be an unfair conclusion for two reasons. First, the assessment in Chicago is created by the makers of the American College Test (ACT) and is designed to prepare students for the ACT and measure college readiness. It is not directly tied to what is being learned in the classroom. Second, we paid for grades, not test scores. The estimate of the impact of incentives on grades is .093 (.057) [ITT] and .131 (.078) [TOT]. We use test scores as our main outcome, however, in lieu of grades because of the relative subjectivity of grades.

Table 3 presents estimates of the causal effect of input incentives on student achievement. The first four columns provide results from second grade students in Dallas (separated by whether they took an English or Spanish test). The final two columns provide results for six through eighth grade students in Washington, DC. Reading achievement in Dallas is split into three mutually exclusive categories: reading comprehension, reading vocabulary, and language.

Offering students the chance to participate in a program that pays them to read books yielded

a .182 (.071) standard deviation increase in reading comprehension skills, a .045 (.068) standard deviation increase in vocabulary scores, and a .150 (.079) standard deviation increase in language skills without controls. Adding our parsimonious set of controls changes these estimates to .180 (.075), .051 (.068), and .136 (.080) standard deviations, respectively. Our set of controls does not significantly alter the point estimates.

Our estimate of the effect of actually participating in a program that pays students to read books yielded a .253 (.097) standard deviation increase in reading comprehension skills, a .062 (.093) standard deviation increase in vocabulary scores, and a .207 (.105) standard deviation increase in language skills. Adding our set of controls adjusts these estimates to .249 (.103), .071 (.093), and .186 (.107) standard deviations, respectively. Hence, paying second grade students to read books has a relatively large effect on their reading comprehension, a more modest effect on their language scores, and a small, statistically insignificant effect on vocabulary.

The juxtaposition of the results from New York and Dallas – large achievement gains when providing incentives to read books for second graders and no improvement when paying fourth graders for test performance – emphasizes the key theme from our experiments. Providing incentives for inputs, not outputs, seems to spur achievement.

Columns (3) and (4) in Table 3 presents similar estimates for Spanish-speaking students in Dallas who took the Logramos test. In this case, our ITT estimates show a .199 (.096) standard deviation decrease in reading comprehension skills, a .256 (.101) standard deviation decrease in vocabulary scores, and a .054 (.114) standard deviation decrease in language skills. Adding controls does not significantly alter the results. The TOT estimates with controls are similar, yielding decreases of .200 (.108), .281 (.119), and .073 (.149), respectively.

A straightforward explanation for these results is that incentives have a negative impact on Spanish speakers, but the intensity of the negative coefficients begs for more exploration. Students who took their tests in Spanish read roughly forty percent of their books for rewards in English, introducing the potential that our program crowded-out investment in Spanish. There are four pieces of evidence which, taken together, suggest that the crowd-out hypothesis may have merit; however, we do not have a definitive test for this theory. First, as we show later, the negative results are entirely driven by the lowest performing students on the Logramos test. These are the students who are likely most susceptible to crowd-out. Second, all bilingual students in Dallas

receive ninety percent of their instruction in Spanish, but poorly performing students are provided more intense Spanish instruction. If intense Spanish instruction is correlated with higher marginal cost of introducing English, this too is consistent with crowd-out. Third, research on bilingual education and language development suggests that introducing English to students who are struggling with native Spanish can cause their "academic Spanish" (but not their conversational skills) to decrease (Mancilla-Martinez and Lesaux, 2010). Thus, our experiment may have had the unintended consequence of crowding out (or confusing) the lowest performing Spanish-speaking students who were being provided intense Spanish remediation. Ultimately, proof of this hypothesis requires an additional experiment in which students are paid to read books in Spanish.

Columns (5) and (6) in Table 3 display the results from our Washington, DC experiments where students were rewarded for several inputs to the educational production function. Offering students a chance to participate in a program that pays them for attendance, behavior, wearing a uniform, and turning in their homework yielded a .041 (.223) standard deviation decrease in reading and a .055 (.217) standard deviation decrease in mathematics. Adding our set of controls changes these estimates to .152 (.092) and .114 (.106) standard deviations, respectively. In this case, our set of controls alters the results considerably, which is troubling. Recall that the randomization left the experimental group unbalanced on school size and fraction minority in the school.

Our estimate of the effect of participating in the Washington, DC, experiment is -.053 (.282) standard deviations for reading and -.071 (.274) standard deviations for math without controls. Adding controls changes the estimates to .179 (.106) and .134 (.122), respectively. Hence, paying middle school students for various inputs to the educational production function such as attendance and behavior has a positive, though marginally significant, effect on reading scores and a slightly smaller (not significant) effect on math scores once one accounts for our set of controls. While these estimates are modest in size and similar in magnitude to the Dallas estimates, we do not have enough statistical power to make more confident conclusions given only thirty-four schools in the experimental group (fifteen of which administered the treatment).

Let us put the magnitude of our estimates in perspective. Jacob and Ludwig (2008), in a survey of programs and policies designed to increase achievement among poor children, report that only three often practiced educational policies pass a simple cost-benefit analysis: lowering class size, bonuses for teachers for teaching in hard-to-staff schools, and early childhood programs. The effect

of lowering class size from 24 to 16 students per teacher is approximately 0.22 (0.05) standard deviations on combined math and reading scores (Krueger, 1999). While 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; Kane and Staiger, 2008), value-added measures are not strongly correlated with observable characteristics of teachers, making it difficult to ex ante identify the best teachers. The effect of Teach for America, an attempt to bring more skilled teachers into struggling schools, is 0.15 standard deviations in math and 0.03 in reading (Decker, Mayer, and Glazerman, 2004). The effect of Head Start is 0.147 (0.103) deviations in applied problems and 0.319 (0.147) in letter identification on the Woodcock-Johnson exam (Currie and Thomas, 2000; Ludwig and Phillips, 2007). An average charter school in New York City raises math scores by 0.09 (0.01) standard deviations per year and ELA scores by 0.04 (0.01) standard deviations per year (Hoxby and Muraka, 2009).[30] Thus, our estimates of the effect of participating in an incentive program based on inputs are consistent with successful reforms in recent decades, but have considerably lower cost. The most expensive of the incentive experiments tested here cost less than $600 per student. Paying students to read books was an order of magnitude less expensive. Krueger and Whitmore (2001) estimate the cost of reducing class size from 22 to 15 students to be $7,502 per student per year (in 1997-98 dollars). Early childhood investments are typically more expensive.

Tables 4A and 4B investigate treatment effects for subsamples that we deemed important before any analysis was conducted – gender, race, previous year's test score and behavioral incidence, and an income proxy.[31] Gender was divided into two categories; the 119 students in Dallas, 284 in NYC, and 24 students in Dallas with missing gender information were not included in the gender subsample estimates. We divide race into five categories: non-Hispanic whites, non-Hispanic blacks, non-black Hispanics, Asians and other race. We only include race in our analysis if there are more than one hundred students of that race within our experimental group. This restriction eliminates whites and Asians in Dallas and a small "other race" category in other cities. Previous year's test scores are partitioned into four groups – evenly distributed terciles for students with a valid pre-treatment test score and a missing category for students without a valid pre-treatment score.

[30] Very successful charter schools can generate achievement gains of .3 to .4 standard deviations per year (Dobbie and Fryer, 2009; Abdulkadiroglu et al., 2009; Angrist et al., 2010), but these are not representative.

[31] These subsamples are also standard in the literature (see Kling, Liebman, and Katz, 2007).

Because of frequent mobility in and out of urban school districts, an average of eighteen percent of students in Dallas, sixteen percent in Chicago, and thirty-one percent in Washington, DC, were missing a previous year's test score. Eligibility for free lunch is used as an income proxy.[32] The final distinction is between students with at least one behavioral incidence in the previous year that led to a suspension and those with none.[33]

Table 4A presents TOT and LATE estimates for gender and race subsamples.[34] The first column provides estimates on the full sample (from Tables 2 and 3) for comparison. There are no gender differences in the two output experiments. Seventh grade boys gain .037 (.040) standard deviations in reading and -.011 (.070) standard deviations in math. In New York, seventh grade girls are similar: .027 (.035) standard deviations and -.046 (.064) standard deviations, respectively. The same pattern holds for fourth graders. In Chicago, the treatment effect on reading scores was .018 (.038) standard deviations for boys and -.034 (.040) standard deviations for girls. Estimates for math are -.028 (.035) and -.002 (.034) standard deviations, respectively.

In sites where incentives were relatively effective, boys seem to gain more from the experiment than girls. In the book reading experiment, the TOT estimate is .319 (.110) standard deviations for boys and .178 (.106) standard deviations for girls for reading comprehension, .148 (.106) standard deviations for boys and -.012 (.095) standard deviations for girls for vocabulary, and .241 (.101) standard deviations for boys and .127 (.144) standard deviations for girls for language total. Similarly, in the attendance/behavior experiment, the TOT estimate in reading is .267 (.132) standard deviations for boys and .091 (.081) standard deviations for girls. The TOT estimate in math is .188 (.136) standard deviations for boys and .076 (.114) standard deviations for girls. This finding is surprising, given results from previous demonstration projects that seem to favor girls (Anderson, 2008; Angrist and Lavy, 2009; Sanbonmatsu et al., 2006; Kling, Liebman, and Katz, 2007).

One of the motivations to perform our experiment was to test whether financial incentives are a potentially cost-effective strategy to decrease racial and socioeconomic achievement gaps. Recall

[32]Using the home addresses in our files and GIS software, we also calculated block-group income. Results are similar and available from the author upon request.

[33]These are relatively major infractions. Five randomly chosen descriptions of behavioral incidence from Washington, DC, include: (1) wrapped teacher up in tape on arm; (2) caught stealing juices from the after school program snack. Juices were found in his book bag and he was selling them at lunch-time; (3) exited the school building without permission; (4) fighting; and (5) student was throwing objects during instruction time.

[34]Appendix Table 3 provides ITT estimates.

that the TOT estimates for the full sample in Dallas were .249 (.103) standard deviations in reading comprehension, .071 (.093) standard deviations in vocabulary, and .186 (.107) standard deviations in language. We cannot reject the null hypothesis that blacks and Hispanics gained equally from the Dallas experiment. Across the three components of the test, both racial groups scored similarly. The estimated treatment effects are .169 (.123) standard deviations for blacks and .314 (.122) standard deviations for Hispanics in reading comprehension, and .179 (.102) standard deviations for blacks and .197 (.135) standard deviations for Hispanics on language. Reading vocabulary scores show a bigger increase for blacks than Hispanics, but are very imprecise. Whites and Asians made up a small portion of the Dallas sample and were not included in the subsample analysis for that site. In Washington, DC, Hispanics benefited from the experiment: a .302 (.116) standard deviation increase in reading scores and a .168 (.132) standard deviations increase in mathematics. All racial groups in New York show either small or inconsistent gains. There were no differences among racial groups in Chicago.

Table 4B presents TOT and LATE estimates from our remaining subsamples. Partitioning by previous year's achievement shows at least one important pattern. For students taking the Logramos test in Dallas, the negative results at the mean are all driven by students in the bottom tercile.[35] The treatment effects for the bottom third of the students taking the Logramos exam are very large: -.370 (.156) in reading comprehension, -.559 (.137) in vocabulary, and -.294 (.199) in language. The treatment effect for Logramos students in the middle and top terciles is statistically zero and substantively small. All other coefficients are roughly equal across terciles. Vocabulary scores in Dallas seem to increase among weaker students. Splitting the sample by whether or not a student had a behavioral incident in the previous year shows no difference in New York. Estimates from Washington, DC, the only experiment that provided incentives for good behavior, suggest that students who had a behavioral incident leading to suspension in the previous year had a relatively large increase in their scores [.400 (.235) standard deviations in reading and .164 (.274) standard deviations in math], though large standard errors make definitive conclusions difficult. The corresponding estimates for students who were not suspended in the previous year are .175 (.097) and .139 (.115) standard deviations.

[35] This holds whether you split pre-treatment test scores in 5 or 10 categories. In the latter case, negative results are concentrated in the bottom three deciles.

Perhaps the most interesting and informative comparison is between students eligible for free lunch and their counterparts who are not eligible for free lunch.[36] The point estimates for students eligible for free lunch are lower than those not eligible for free lunch in the input experiments, but one cannot reject the null hypothesis that they are the same.[37] In our output experiments, students not eligible for free lunch show modest gains from the experiment on three out of four assessments. The point estimates for fourth graders in reading are -.006 (.057) standard deviations for students eligible for free lunch and -.141 (.112) standard deviations for those not eligible for free lunch. One cannot reject the null hypothesis that these coefficients are the same. All other coefficients for students not eligible for free lunch are positive, but measured imprecisely. The treatment effect for fourth graders in math is .179 (.143) standard deviations for students not eligible for free lunch and .056 (.072) standard deviations for those eligible for free lunch. In seventh grade, the differences are similar. The impact on achievement of participating in an incentive program is .116 (.086) standard deviations in math and .229 (.091) standard deviations in reading for seventh graders who are not eligible for free lunch. The corresponding treatment effects for students who qualify for free lunch are -.052 (.061) standard deviations and .013 (.036) standard deviations. This pattern was not seen in Chicago, where the treatment effect for students not eligible for free lunch is .025 (.077) standard deviations in English and -.075 (.052) standard deviations in math. The impact of incentives on achievement for students eligible for free lunch is -.006 (.036) standard deviations in reading and -.009 (.030) standard deviations in math in Chicago.

Taken together, our analysis of heterogeneous treatment effects adds some nuance to our main results. The bulk of the evidence suggests that input experiments are effective for minorities and especially for minority boys, while incentives for output may benefit those who do not qualify for free lunch. Providing incentives for behavior seems to be effective at increasing reading achievement among students who had at least one behavioral incident that led to a suspension in the year before treatment.

[36] In the Early Childhood Longitudinal Study, a recent and large nationally representative sample of students from kindergarten through eighth grade collected by the Department of Education, students on free lunch are more likely to be concentrated in single parent households, have more siblings, and have parents with less education than students who are not eligible for free lunch.

[37] This does not include the Logramos students.

5 Alternative Outcomes

Thus far, we concentrated on student achievement as measured by statewide assessments. Alternative measures of achievement might also be informative. For example, in NYC, the state assessments are given in late January (reading) and early March (math). This reduced calendar severely truncates the length of the experiment.[38] A predictive assessment, given to every student in the district, is highly correlated with the state exam and is given in October and June. Students were also given incentives for these predictive exams (but not the state assessments) as part of the ten total tests that were rewarded, which provides further variation. In addition, financial incentives may also affect outcomes such as behavior, daily attendance, report card grades, or overall effort.

Alternative Outcomes

Table 5 shows estimates of the impact of incentives on alternative outcomes. All variables are taken from each district's administrative files and, hence, differ slightly from city to city. In each district, we have data on attendance rates and report card grades. 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, according to each district's attendance file. Grades were pulled from files containing the transcripts for all students in each district. Letter grades were converted to a 4.0 scale. Student's grades from each semester (including the summer when applicable) were averaged to yield a GPA for the year. As with test scores, GPAs were standardized to have mean of zero and standard deviation of one among students in the same grade across the school district. In addition, Dallas contains math scores from the Iowa Tests of Basic Skills; Washington, DC, has behavioral incidence data; Chicago has the number of total credits earned; and New York City collects data on behavioral incidence and scores on additional math and ELA assessments.

Each panel in Table 5 represents a different school district. The estimates in Panel A, for Dallas, show that there is no treatment effect of incentives on attendance rates. This finding is not surprising, given that the average second grader in Dallas attends ninety-seven percent of school days. Incentives had a large impact, however, on report card grades; students who participated

[38] In the year after the experiment ended, the NYC Department of Education moved the state assessments to later in the academic year.

in the incentive program gained .311 (.142) standard deviations. Math test scores also slightly increased but the treatment effect is not statistically significant. There is no evidence that students who took the Logramos exam increased or decreased their performance on any of these alternative outcomes.

Panels B and C present estimates of the impact of incentives on alternative forms of achievement in Washington, DC, and Chicago, respectively. Students in the Washington, DC, treatment had higher attendance rates [.171 (.235)] and fewer behavioral incidents [-.323 (.245)], but because of lack of statistical power we cannot make confident conclusions. Report card grades did not increase [.049 (.148)]. Conversely, our experiment in Chicago demonstrates that paying students for better course grades has a modest impact on their grades – an increase of .131 (.078) standard deviations – along with a larger increase in attendance .214 (.113) and an increase of 2.697 (1.556) credits earned. All the estimates on the above alternative outcomes for Chicago are marginally significant. The typical course in Chicago is worth four credits. Treatment students, therefore, passed approximately one-half more course on average during their freshman year than control students. This evidence from Chicago suggests that incentives for grades may be an effective dropout prevention strategy, though it does not increase human capital in a measurable way after one year.

The final panel in Table 5 provides LATE estimates of our intervention on alternative outcomes for fourth and seventh graders in New York. Providing incentives for students to earn higher test scores did not cause students to attend school more often, behave better, earn better grades, or perform better on predictive exams for which they were incentivized. More succinctly, we cannot find any evidence on any quantifiable dimension that paying students for higher test scores changed their behavior.

Effort

Along with the outcomes described in Table 5, it is potentially important to understand how incentives altered different forms of effort in school. Unfortunately, data on student effort is not collected by school districts, so we turn to our survey data. On the survey, we asked nine questions to serve as proxies for effort, which included: (1) how often a student is late for school; (2) whether a student asks for teacher help if she needs it; (3) how much of her assigned homework she completes; (4) whether she works very hard at school; (5) whether she cares if she arrives on-time to class; (6) if her behavior is a problem for teachers; (7) if she is satisfied with their achievement; (8) whether

she pushes herself hard at school; and (9) how many hours per week she spends on homework.[39] See Appendix B for further details.

Our results, shown in Table 6, indicate that there are few differences between students who received treatment and those who did not on the dimensions of effort described above. We caution against over-interpreting any given coefficient. In Dallas, treated students report that they are -.293 (.130) standard deviations less likely to ask a teacher for help, but also report that they work .216 (.106) standard deviations harder on their schoolwork, relative to the control group. It is plausible that reading books and taking computerized exams at their own pace made students feel more independent from their teachers. Working harder on their schoolwork is consistent with the increases shown in their grade point average. In Washington, DC, students reported that they are .318 (.050) standard deviations more likely to complete their homework and have .144 (.052) standard deviations fewer behavioral problems in school. All other dimensions of effort show no differences. We cannot reject the null hypothesis of no effect for any of the effort variables in our New York City experiment.

Under some assumptions, providing incentives for a particular activity would have spillover effects on many other activities. For instance, paying students to read books might make them excited about math as well. Further, paying students for attendance and behavior might provide such enthusiasm for school that students engage differently with their teachers. Tables 5 and 6 provide evidence that such effects are not likely present, as the impacts of incentive programs seem relatively localized. Many qualitative observations confirm general excitement by students about earning rewards, but they seem to focus their behavioral changes on precisely those elements that are incentivized.

Intrinsic Motivation

One of the major criticisms of the use of incentives to boost student achievement is that the incentives may destroy the "love of learning." In other words, providing extrinsic rewards can crowd out intrinsic motivation in some situations. There is a debate in social psychology on this issue – see Cameron and Pierce (1994) for a meta-analysis of the literature.

To test the impact of our incentive experiments on intrinsic motivation, we disseminated the

[39] Because participating students in Dallas are only in second grade, many of the questions were not asked of them. Students in Dallas were only asked questions (2) and (4).

Intrinsic Motivation Inventory, developed by Ryan (1982), to students in our experimental group in all cities.[40] The instrument assesses participants' interest/enjoyment, perceived competence, effort, value/usefulness, pressure and tension, and perceived choice while performing a given activity. There are subscale scores for each of those six categories. We only include the interest/enjoyment subscale in our surveys as it is considered the self-report measure of intrinsic motivation. The interest/enjoyment subscale consists of seven statements on the survey: (1) I enjoyed doing this activity very much; (2) this activity was fun to do; (3) I thought this was a boring activity; (4) this activity did not hold my attention at all; (5) I would describe this activity as very interesting; (6) I thought this activity was quite enjoyable; and (7) while I was doing this activity, I was thinking about how much I enjoyed it. Respondents are asked how much they agree with each of the above statements on a seven-point Likert scale ranging from "not at all true" to "very true." To get an overall intrinsic motivation score, one adds up the values on each statement (reversing the sign on statements (3) and (4)). Only students with valid responses on each statement are included in our analysis of the overall score, as non-response may be confused with low intrinsic-motivation. In addition, we also estimate treatment effects on each statement independently, which allows us to use the maximum number of observations.

Table 7 provides estimates of the impact of our incentive programs on the overall intrinsic motivation score as well as each survey statement independently. Coefficients reported are TOT and LATE estimates (see Appendix Table 7 for ITT estimates). Students in Dallas who took the English exam had a small and insignificant increase in their motivation, .611 (.816) on a mean of 23.517. The intrinsic motivation of students who took the Spanish exam decreased by .902 (.643) on a mean of 24.223. In New York, intrinsic motivation decreased by 1.277 (1.064) on a mean of 25.520 . Finally, our experiment in Washington, DC, resulted in a small and insignificant increase in intrinsic motivation. Put together, these results show that our experiments had very little impact on intrinsic motivation. By these measures, the concern of some educators and social psychologists that rewarding students will negatively impact their "love of learning" seems unwarranted in this context.

A second major criticism, related to concerns over intrinsic motivation, concerns the effects on

[40]The inventory has been used in several experiments related to intrinsic motivation and self-regulation [e.g., Ryan, Koestner, and Deci (1991) and Deci et al. (1994)].

students when the incentives are discontinued. Many believe that students will experience decreased motivation and thus decreases in achievement measures when the incentives are removed following the experiment (Kohn 1993 and references therein). Thus far, we can only answer this question for the experiments implemented in Dallas. A year after paying students $2 per book to read and pass a short quiz, students who received treatment are still significantly outperforming students who were in the control group. More precisely, the coefficient on treatment in our estimating equation, one year after the experiment ended, is .088 (.080) [ITT] and .120 (.107) [TOT] and math .150 (.108) [ITT] and .206 (.145) [TOT] [not shown in tabular form]. In other words, the point estimate a year after the experiment is roughly half of the original effect in reading and not statistically significant. This is similar to the fade out effect which has been documented from Head Start, having a high quality teacher in one year, or a lower class size (Nye, Hedges, and Konstantopoulos, 1999; Puma, 2010). For students who took the Logramos exam in Dallas, the effects of the program a year after being discontinued are small and statistically insignificant [.002 (.064) [ITT] and .003 (.075) [TOT] in reading and -.046 (.080) [ITT] and -.055 (.093) [TOT] in math]. Hence, either the negative impacts observed in the year of the treatment did not persist or the negative impacts can be explained by crowd-out.

6 Interpretation

Our experiments have generated a rich set of new facts. Paying second grade students to read books significantly increases reading achievement and these effects are still present a year after the incentives are discontinued. Paying fourth grade students for test scores has little effect. Paying middle school students for attendance, behavior, wearing their uniforms, and turning in their homework has a marginally significant increase on reading achievement and has a similar but statistically insignificant impact on math achievement. Paying seventh grade students for test scores does not increase their achievement. Paying high school freshmen for their grades in core courses leads to modest increases on their overall grades, attendance, and the number of courses they pass, but has no effect on standardized test scores.

Moreover, incentives for inputs seem particularly effective for boys, blacks, and Hispanics. Students who are not eligible for free lunch and Asians tend to benefit from all incentive programs

– input or output. The only treatment that provided incentives for good behavior showed gains in achievement for students who had had behavioral problems in the previous year. Finally, there was scant evidence that general effort increased or that intrinsic motivation decreased during any of our incentive treatments.

A possible interpretation of our results is that all the estimates are essentially zero and the effects in Dallas and Washington, DC, were observed by chance alone. Yet, the size of the results and the consistency with past research cast doubt on this as an explanation (Kim, 2007). A second, more reasonable, interpretation is that the only meaningful effects stem from the Dallas experiment. This finding could either be due to the fact that the Dallas experiment targeted younger students or because reading books is a more important input into education production. Arguing against the former explanation is the fact that the fourth graders in New York demonstrated null results.

A third interpretation of the results is that incentives are effective if tailored to appropriate inputs to the educational production function. Note that, in order to make this interpretation, we either have to depend on marginally significant point estimates in the Washington, DC, experiment, or believe that the inputs rewarded were not as important for achievement as reading books. This is our leading interpretation.

Recall that the traditional economic model with a simple set of assumptions predicts that incentives for output are socially optimal. In what follows, we discuss four alterations to the simple set of assumptions that can explain the results of our incentive experiments.

Model 1: Lack of Knowledge of the Education Production Function

The standard economic model implicitly assumes that students know their production function – that is, the precise relationship between the vector of inputs and corresponding output.[41] If students only have a vague idea of how to increase output, then there may be little incentive to increase effort. In Dallas and Washington, DC, students were not required to know how to increase their test scores; they only needed to know how to read books on their grade level, attend class, behave well, wear their uniforms, and so on. In New York, students were required either to know how to produce test scores or to know someone who could help them with the task. In Chicago, students faced a similar challenge.

The best evidence for a model in which students lack knowledge of the education production

[41] Technically, students are only assumed to have more knowledge of their production function than the principal.

function lies in our qualitative data. Seven full-time qualitative researchers observed twelve students and their families, along with ten classrooms, in New York during the 2008-2009 school year. From detailed interview notes, we gather that students were uniformly excited about the incentives and the prospect of earning money for school performance.[42] Despite showing that students are excited about the incentive programs, the qualitative data also demonstrates that students had little idea about how to translate their enthusiasm into tangible action steps designed to increase their achievement. After each of the ten exams administered in New York, our qualitative team asked students how they felt about the rewards and what they could do to earn more money on the next test. Every student found the question about how to increase his or her scores difficult to answer. Students answering this question stated thinking about test-taking strategies rather than salient inputs into the education production function or improving their general understanding of a subject area.[43] For instance, many of the students expressed the importance of "reading the test questions more carefully," "not racing to see who could finish first," or "re-reading their answers to make sure they entered them correctly." Not a single student mentioned: reading the textbook, studying harder, completing their homework, or asking teachers or other adults about confusing topics.

Two focus groups in Chicago confirmed the more systematically collected qualitative data from New York. The focus groups contained a total of thirteen students, evenly split between blacks and Hispanics, males and females. Again, students reported excitement about receiving financial incentives for their grades. Students also reported that they attended school more, turned in more homework and listened more in class.[44] This finding is consistent with the empirical data in Table 6.

Yet when probed why more inputs to the educational production function were not utilized – reading books, staying after school to work on more problems, asking teachers for help when they were confused, reviewing homework before tests, or doing practice problems available in textbooks

[42]In a particularly illuminating example, one of the treatment schools asked their students to elect a new "law"' for the school, a pedagogical tool to teach students how bills make their way through Congress. The winner, by a nearly unanimous vote, was a proposal to take incentive tests every day.

[43]The only slight exception to this rule was a young girl who exclaimed "it sure would be nice to have a tutor or something."

[44]These strategies may be effective at turning failing grades into marginally passing grades, which would explain our treatment effects in Table 5, but are not likely to result in test score growth.

– one female student remarked "I never thought about it."[45]

The basic messages from students in Chicago centered on the excitement generated at the beginning of the year by the program. They responded with more effort – coming to school, paying attention in class, and so on – but students indicated that they did not notice any change on their performance on quizzes or tests, so they eventually stopped trying. As one student expressed, "classes were still hard after I tried doing my homework."

Model 2: Self-Control Problems

Another model consistent with the data is that students know the production function, but either have self-control problems or are sufficiently myopic that they cannot make themselves do the intermediate steps necessary to produce higher test scores. In other words, if students know that they will be rewarded for an exam that takes place in five weeks, they cannot commit to daily reading, paying attention in class, and doing homework even if they know it will eventually increase their achievement. Technically, students should calculate the net present value of future rewards and defer other near-term rewards of lesser value. Extensive research has shown that this is not the case in many economic applications (Laibson, 1997). Similar ideas were presented by the social psychology experiments discussed in Mischel, Shoda, and Rodriguez (1989).

Reading books provided feedback and affirmation anytime a student took a computerized test. Teachers in Chicago likely provided daily feedback on student progress in class and via homework, quizzes, chapter tests, and so on. Students in Washington, DC, were often reminded how their attendance, behavior, etc. affected their pay.

The challenge with this model is to identify ways to adequately test it. Two ideas seem promising. First, one could collect information before the experiment initiated on the discount rates of all students in treatment and control schools and then test for heterogeneous treatment effects between students with relatively high discount rates and those with low discount rates. If the theory is correct, the difference in treatment effects (between input and output experiments) should be significantly smaller for the subset of students who have low discount factors. A potential limitation of this approach is that it critically depends on the metric for deciphering high and low discount

[45]The rest of the focus group participants offered blank stares and shrugs, before settling into a more defiant mode. "I did some homework, [expletive], what else they want?" The vast majority of students in our sample considered doing their homework as maximal effort.

rates and its ability to detect other behavioral phenomena that might produce similar self-control problems. Second, one might design an intervention that assesses students every day and provides immediate incentives based on these daily assessments. If students do not significantly increase their achievement with daily assessments, it provides good evidence that self-control cannot explain our findings. A potential roadblock for this approach is the burden it would put on schools to implement it as a true field experiment for a reasonable period of time.

Model 3: Complementary Inputs

The third model that can explain our findings is that the educational production function has important complementarities that are out of the student's control. For instance, incentives may need to be coupled with good teachers, an engaging curriculum, effective parents, or other inputs to produce output. In Dallas, students could read books independently and at their own pace. In Washington, DC, we provided incentives for several inputs – many of which may be complementary. It is plausible that increased student effort, parental support and guidance, and high quality schools were necessary and sufficient conditions for test scores to increase during our Chicago or New York experiments.

There are several (albeit weak) tests of elements of this model that are possible with our administrative data. If effective teachers are an important complementary input to student incentives in producing test scores, we should notice a correlation between the value-added of a student's teacher and the impact of incentives on achievement. To test this idea we linked every student in our experimental schools in New York to their homeroom teachers for fourth grade and subject teachers (math and ELA) in seventh grade. Using data on the "value-added" of each teacher from New York City, we divided students in treatment and control schools into two groups based on high or low value-added of their teacher.[46]

Table 8 shows the results of this exercise. The first column reports LATE estimates for the New York sample for all students in treatment and control whose teachers have valid value-added

[46]Value-added estimates for New York City were produced by the Battelle Institute (http://www.battelleforkids.org/). To determine a teacher's effect, Battelle predicted achievement of a teacher's students controlling for student, classroom, and school factors they deemed outside of a teacher's control (e.g., student's prior achievement, class size). A teacher's value added score is assumed to be the difference between the predicted and actual gains of his/her students.

data. This subset represents approximately 47 percent of the full sample. The results from this subset of students are similar to the full sample, save that in fourth grade math we observe a sizable treatment effect. The next two columns divide students into those whose are assigned a teacher who is above and below the median value-added of teachers in New York City, respectively. Across these two groups, there is very little predictable heterogeneity in treatment effects. The best argument for teachers as a complementary input in production is given by fourth grade math. Students with below-the-median quality teachers gain .103 (.130) standard deviations and those with above-the-median teachers gain .223 (.111) standard deviations. The exact opposite pattern is observed for seventh grade math.

The best evidence in favor of the importance of complements in production are the differences between students who were not eligible for free lunch (and likely have intact, more educated families who are more engaged in their schooling) and those who are eligible for free lunch, though the differences are not statistically significant. The social ills that are correlated with eligibility for free lunch may be important limitations in production of achievement. An anecdote from our qualitative interviews illustrates the potential power of parental involvement and expectations coupled with student incentives to drive achievement. Our interviewers followed a high performing Chinese immigrant student home when she told an illiterate grandmother that she had earned $30 from her performance at school. The grandmother immediately retorted, "But Jimmy next door won more than you!"

We will not even hazard a guess as to whether or not complementary inputs can explain our set of results.[47]

Model 4: Unpredictability of Outputs

A classic result in price theory is that incentives should be provided for inputs when the production technology is sufficiently noisy. It is quite possible that students perceive (perhaps correctly) that test scores are very noisy and determined by factors outside their control. Thus, incentives based on these tests do not truly provide incentives to invest in inputs to the educational production function because students believe there is too much luck involved. Indeed, if one were to rank our incentive experiments in order of least to most noise associated with obtaining the incentive, a likely

[47] Ongoing work by Petra Todd and Kenneth Wolpin at the University of Pennsylvania in which they provide overarching incentives for teachers, students, and parents in Mexico City may hold valuable clues.

order would be: (1) reading books, (2) attending class and exhibiting good behavior (attendance is straightforward, but behavior depends, in part, on other students' behavior), (3) course grades, and (4) test scores. Consistent with the theory of unpredictability of outputs, this order is identical to that observed if the experiments are ranked according to the magnitude of their treatment effects.

It is important to remember that our incentive tests in New York were adaptive tests. These exams can quickly move students outside their comfort zone and into material that was not covered in class – especially if they are answering questions correctly. The qualitative team noted several instances in which students complained to their teachers when they were taken aback by questions asked on the exams or surprised by their test results. To these students – and likely many more – the tests felt arbitrary.

The challenge for this theory is that even with the inherent unpredictability of test scores, students do not invest in activities that have a high likelihood of increasing achievement (reading books, e.g.). That is, assuming students understand that reading books, doing problem sets, and so on will increase test scores (in expectation), it is puzzling why they do not take the risk.[48]

***

Deciphering which model is most responsible for our set of facts is beyond the scope of this paper. One or several combinations of the above models may ultimately be the correct framework. Future experimentation and modeling is needed. Indeed, the results suggest a theory of decision making in which agents do not know the production function. This is different from the typical uncertainty in principal-agent models.

7 Conclusion

School districts have become important engines of innovation in American education. A strategy hitherto untested in urban public schools is to provide financial incentives for student achievement. In partnership with four school districts, we conducted school-based randomized trials in 261 urban schools, distributing $6.3 million to roughly 20,000 students, designed to test the impact of incentives on student achievement.

[48] If students do not know how noisy tests are or what influences them, the model is equivalent to Model 1.

Our results show that incentives can raise achievement among even the poorest minority students in the lowest performing schools if the incentives are given for certain inputs to the educational production function. Incentives for output are much less effective. The magnitudes of the increases in achievement are similar to successful reforms in the past few decades, and obtained at lower cost. Yet incentives are by no means a silver bullet. They pass a simple cost-benefit analysis, but are not powerful enough to overcome the racial achievement gap alone. High performing charter organizations such as the Knowledge is Power Program (KIPP) and Harlem Children's Zone routinely use input incentives as an integral part of a broader whole-school strategy.[49]

The leading theory to explain our results is that students do not know the educational production function, and thus lack the know-how to transform excitement about rewards into tangible investment choices that lead to increases in achievement. Several qualitative observations support this theory, but other models such as lack of self-control, complementary inputs in production, or the unpredictability of tests, are also consistent with the experimental data.

Taken together, our experiments and economic model provide the beginnings of a theory of incentives in urban education. This theory has the potential for use in many other applications. For instance, it might be less effective to give teachers incentive pay based on outputs (test scores of their students) relative to inputs (staying after school to tutor their students). Evidence from developing countries suggests that inputs may be important in this context as well (Duflo and Hanna, 2006; Muralidharan and Sundararaman, 2009). A complete understanding will require more experimentation and constant refining of the theoretical assumptions.

References

[1] Abdulkadiroglu, Atila, Joshua Angrist, Susan Dynarski, Thomas Kane and Parag Pathak. 2009. "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." NBER Working Paper No. 15549.

[2] Anderson, Michael. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training

[49] KIPP pays students on a point system, similar to our Washington, DC, treatment, in which students can earn or lose rewards that can be redeemed at their school store. Students in the Harlem Children's Zone earn up to $120 a month for attending school and making good grades.

Projects." Journal of the American Statistical Association, 103(484): 1481-1495.

[3] Angrist, Joshua D., Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." The American Economic Review, 92(5): 1535-1558.

[4] Angrist, Joshua D., Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." The American Economic Review, 96(3): 847-862.

[5] Angrist, Joshua D., Daniel Lang, and Philip Oreopoulos. 2006. "Lead Them to Water and Pay Them to Drink: An Experiment with Services and Incentives for College Achievement." NBER Working Paper No. 12790.

[6] Angrist, Josh D., and Victor Lavy. 2009. "The Effect of High-Stakes High School Achievement Awards: Evidence from a Group-Randomized Trial." American Economic Review, 99(4): 1384-1414.

[7] Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters. 2010. "Who Benefits from KIPP?" NBER Working Paper No. 15740.

[8] Banks, James A. 2001. "Approaches to Multicultural Curriculum Reform." In Multicultural Education: Issues and Perspectives, 4th Edition, ed. James A. Banks and Cherry A.M. Banks. New York: John Wiley & Sons, Inc.

[9] Banks, James A. 2006. Cultural Diversity and Education: Foundations, Curriculum, and Teaching. Boston: Pearson Education, Inc.

[10] Barrera-Osorio, Felipe, Marianne Bertrand, Leigh L. Linden, and Francisco Perez-Calle. 2008. "Conditional Cash Transfers in Education: Design Features, Peer and Sibling Effects: Evidence from Randomized Experiment in Colombia." NBER Working Paper No. 13890.

[11] Behrman, Jere R., Piyali Sengupta and Petra Todd. 2005. "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico." Economic Development and Cultural Change, 54: 237-275.

[12] Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Edlar Shafir, and Jonathan Zinman. 2009. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing

Field Experiment." Yale University Economic Growth Center Discussion Paper No. 968.

[13] Bettinger, Eric. 2008. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores." CESifo/PEPG Conference on Economic Incentives: Do They Work in Education? Insights and Findings from Behavioral Research.

[14] 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.

[15] Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2008. "Teacher Preparation and Student Achievement." NBER Working Paper No. 14314.

[16] Bruhn, Miriam and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." American Economic Journal: Applied Economics, 1(4): 200-232.

[17] Cameron, Judy and W. David Pierce. 1994. "Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis." Review of Educational Research, 64(3): 363-423.

[18] Card, David. 1999. "The Causal Effect of Education on Earnings." In Handbook of Labor Economics Vol. 3, ed. David Card and Orley Ashenfelter, 1802-1859. Amsterdam: North Holland.

[19] Currie, Janet and Duncan Thomas. 2000. "School Quality and The Longer-Term Effects Of Head Start." Journal of Human Resources, 35(4): 755-774.

[20] Datnow, Amanda, Vicki Park, and Brianna Kennedy. 2008. Acting on Data: How urban high schools use data to improve instruction. Los Angeles: Center on Educational Governance, USC Rossier School of Education.

[21] Deci, Edward L. 1972. "The Effects of Contingent and Noncontingent Rewards and Controls on Intrinsic Motivation." Organizational Behavior and Human Performance, 8: 217-229.

[22] Deci, Edward L. 1975. Intrinsic Motivation. New York: Plenum.

[23] Deci, Edward L., Haleh Eghrari, Brian C. Patrick and Dean R. Leone. 1994. "Facilitating Internalization: The Self-Determination Theory Perspective." Journal of Personality, 62(1): 119-142.

[24] 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.

[25] Dobbie, Will and Roland G. Fryer. 2009. "Are High-Quality Schools Enough to Close the Achievement Gap? Evidence from a Bold Social Experiment in Harlem." NBER Working Paper No. 15473.

[26] Domina, Thurston. 2005. "Leveling the Home Advantage: Assessing the Effectiveness of Parental Involvement in Elementary School." Sociology of Education, 78(3): 233-249.

[27] Donner, Allan, K. Stephen Brown, and Penny Brasher. 1990. "A Methodological Review of Non-Therapeutic Intervention Trials Employing Cluster Randomization, 1979-1989." International Journal of Epidemiology, 19(4): 795-800.

[28] Duflo, Esther and Rema Hanna. 2006. "Monitoring Works: Getting Teachers to Come to School." NBER Working Paper No. 11880.

[29] Easton, John Q., Susan Leigh Flinspach, Carla O'Connor, Mark Paul, Jesse Qualls, and Susan P. Ryan. 1993. Local School Council Governance: The Third Year of Chicago School Reform. Chicago: Chicago Panel on Public School Policy and Finance.

[30] Feng, Ziding, Paula Diehr, Arthur Peterson, and Dale McLerran. 2001. "Selected Statistical Issues in Group Randomized Trials." Annual Review of Public Health, 22: 167-187.

[31] Gneezy, Uri and Aldo Rustichini. 2000. "Pay Enough or Don't Pay at All." The Quarterly Journal of Economics, 115(3): 791-810.

[32] Greene, Jay P. and Marcus A. Winters. 2006. "Getting Ahead by Staying Behind: An Evaluation of Florida's Program to End Social Promotion." Education Next, 6(2): 65-69.

[33] Grossman, Sanford J. and Oliver D. Hart. 1983. "An Analysis of the Principal-Agent Problem." Econometrica, 51(1): 7-45.

[34] Goolsbee, Austan and Jonathan Guryan. 2006. "The Impact of Internet Subsidies in Public Schools." The Review of Economics and Statistics, 88(2): 336-347.

[35] Guite, H., C. Clark, and G. Ackrill. 2006. "The Impact of Physical and Urban Environment on Mental Well-Being." Public Health, 120(12): 1117-1126.

[36] Guryan, Jonathan. 2001. "Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts." NBER Working Paper No. 8269.

[37] Hahn, A., T. Leavitt, and P. Aaron. 1994. "Evaluation of the Quantum Opportunities Program (QOP). Did the program work? A report on the post secondary outcomes and cost-effectiveness of the QOP program (1989-1993)." Massachusetts. (ERIC Document Reproduction Service No. ED 385 621).

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

[39] Henig, Jeffrey R., and Wilbur C. Rich. 2004. Mayors in the Middle: Politics, Race, and Mayoral Control of Urban Schools. Princeton, NJ: Princeton University Press.

[40] Holmstrom, Bengt. 1979. "Moral Hazard and Observability." Bell Journal of Economics, 10(1): 74-91.

[41] Hoxby, Caroline, and Sonali Murarka. 2009. "Charter Schools In New York City: Who Enrolls and How They Affect Their Students' Achievement." NBER Working Paper No. 14852.

[42] Jackson, Clement K. 2009. "A Stitch in Time: The Effects of a Novel Incentive-Based High-School Intervention on College Outcomes." http://works.bepress.com/c_kirabo_jackson/5.

[43] 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.

[44] Jacob, Brian A., and Jens Ludwig. 2008. "Improving Educational Outcomes for Poor Children." NBER Working Paper No. 14550.

[45] 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.

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

[47] Karlan, Dean S. and Martin Valdivia. 2006. "Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions." Center Discussion Paper No. 941, Yale

University Economic Growth Center.

[48] Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz, 2007. "Experimental Analysis of Neighborhood Effects." Econometrica, 75(1): 83-119.

[49] Kim, J. S. 2007. "The Effects of a Voluntary Summer Reading Intervention on Reading Activities and Reading Achievement." Journal of Educational Psychology,99(3): 505-515.

[50] Knight, Jim, ed. 2009. Coaching: Approaches and Perspectives. Thousand Oaks, CA: Corwin Press.

[51] Kohn, Alfie. 1993. Punished by Rewards. Boston: Houghton Mifflin Company.

[52] Kohn, Alfie. 1996. "By All Available Means: Cameron and Pierce's Defense of Extrinsic Motivators." Review of Educational Research, 66(1): 1-4.

[53] Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2004. "Incentives to Learn." NBER Working Paper No. 10971.

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

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

[56] Kruger, Justin, and David Dunning. 1999. "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments." Journal of Personality and Social Psychology, 77(6): 1121-1134.

[57] Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." The Quarterly Journal of Economics, 112(2): 443-477.

[58] 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.

[59] Ludwig, Jens, and Deborah Phillips, 2007. "The Benefits and Costs of Head Start." NBER Working Paper No. 12973.

[60] Mancilla-Martinez, Jeannette and Nonie K. Lesaux. 2010. "The Gap Between Spanish-speakers' Word Reading and Word Knowledge: A Longitudinal Study." Unpublished paper, Harvard University.

[61] Marlow, Michael L. 2000. "Spending, School Structure, and Public Education Quality: Evidence from California." Economics of Education Review, 19(1): 89-106.

[62] Mirrlees, J. A. 1974. "Notes on Welfare Economics, Information and Uncertainty." In Essays on Equilibrium Behavior under Uncertainty, ed. M. Balch, D. McFadden, and S. Wu. Amsterdam: North Holland.

[63] Mischel, Walter, Yuichi Shoda, and Monica L. Rodriguez. 1989. "Delay of Gratification in Children." Science, New Series 244(4907): 933-938.

[64] Muralidharan, Karthik and Venkatesh Sundararaman. 2009. "Teacher Performance Pay: Experimental Evidence from India." NBER Working Paper No. 15323.

[65] Neal, Derek A., and William Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differentials." Journal of Political Economy, 104: 869-95.

[66] Neal, Derek A. 2005. "Why Has Black-White Skill Convergence Stopped?". NBER Working Paper No. 11090.

[67] 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.

[68] Nye, Barbara, Larry V. Hedges, and Spyros Konstantopoulos. 1999. "The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment." Educational Evaluation and Policy Analysis, 21(2): 127-142.

[69] Offenberg, Robert M. 2001. "The Efficacy of Philadelphia's K-to-8 Schools Compared to Middle Grade Schools." Middle School Journal, 32(4): 23-39.

[70] OECD. 2007. Education at a Glance 2007: OECD Indicators. http://www.oecd.org/document/30/0,3343,en_2649_39263238_39251550_1_1_1_1,00.html.

[71] Parks, S. E., R. A. Housemann, and R. C. Brownson. 2003."Differential Correlates of Physical Activity in Urban and Rural Adults of Various Socioeconomic Backgrounds in the United

States." Journal of Epidemiology and Community Health, 57(1): 29-35.

[72] Podgursky, Michael J., and Matthew G. Springer. 2007. "Credentials Versus Performance: Review of the Teacher Performance Pay Research. Peabody Journal of Education, 82(4): 551-573.

[73] Protheroe, Nancy J., and Kelly J. Barsdate. 1991. Culturally Sensitive Instruction and Student Learning. Arlington, VA: Educational Research Center.

[74] Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, et al. 2010. "Head Start Impact Study: Final Report." U.S. Department of Health and Human Services, Washington, DC.

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

[76] Rockoff, Jonah E. 2008. "Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City." NBER Working Paper No. 13868.

[77] Rouse, Cecilia E., and Alan B. Krueger. 2004. "Putting Computerized Instruction to the Test: A Randomized Evaluation of a 'Scientifically Based' Reading Program." Economics of Education Review, 23(4): 323-338.

[78] Ryan, Richard M. 1982. "Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory." Journal of Personality and Social Psychology, 63: 397-427.

[79] Ryan, Richard M., Richard Koestner, and Edward L. Deci. 1991. "Ego-Involved Persistence: When Free-Choice Behavior is Not Intrinsically Motivated." Motivation and Emotion, 15(3): 185-205.

[80] Sanbonmatsu, Lisa, Jeffrey R. Kling, Greg J. Duncan, and Jeanne Brooks-Gunn. 2006. "Neighborhoods and Academic Achievement." Journal of Human Resources, 41(4): 649-691.

[81] Shapka, Jennifer D., and Daniel P. Keating. 2003. "Effects of a Girls-Only Curriculum during Adolescence: Performance, Persistence, and Engagement in Mathematics and Science." American Educational Research Journal, 40(4): 929-960.

[82] Swanson, Cristopher B. 2009. Cities in Crisis 2009: Closing the Graduation Gap. Bethesda, MD: Editorial Projects in Education, Inc. http://www.americaspromise.org/Our-

Work/Dropout-Prevention/Cities-in-Crisis.aspx

[83] Thernstrom, Abigail. 1992. "The Drive for Racially Inclusive Schools." Annals of the American Academy of Political and Social Science, 523: 131-143.

[84] Wong, Kenneth L., and Francis X. Shen. 2002. "Do School District Takeovers Work? Assessing the Effectiveness of City and State Takeovers as a School Reform Strategy." State Education Standard, 3(2): 19-23.

[85] Wong, Kenneth L., and Francis X. Shen. 2005. "When Mayors Lead Urban Schools: Assessing the Effects of Takeover." In Besieged: School Boards and the Future of Education Politics, ed. William G. Howell, 81-101. Washington, D.C.: Brookings Institution Press.

8 Appendix A: Implementation Manual

8.1 Capital Gains: An Experiment in DC Public Schools

A. BACKGROUND AND OVERVIEW

On August 8, 2008, DCPS Chancellor Michelle Rhee and Education Innovation Laboratory (EdLabs) director Roland Fryer conducted an introductory meeting with all principals of schools with students in sixth, seventh, or eighth grade. More than any other district leader involved with the incentive experiments, newly minted Chancellor Michelle Rhee made the Capital Gains program one of her signature initiatives. As such, schools and students were expected to participate unless they had a compelling reason to not do so.

B. RECRUITMENT AND SELECTION

Schools

After hearing the premise of the program, 28 principals asked for their schools to be included in the randomization process. Fourteen schools were selected as treatment schools, but one declined to participate. The remaining thirteen treatment schools that were selected were provided with school specific training to help set up the program. After the initial randomization, five more schools that had not originally attended the introductory meeting were also added to the pool. Three schools were selected for treatment: two of these schools chose to participate in the program (the other did not respond to EdLabs within the required twenty-four hours). In total, there were 34 schools: 17 selected into the treatment group (two of which did not participate) and 17 selected into the control group.

Each principal of a treatment school received sample student consent forms, brochures, and general overviews to share with their staffs. Each treatment school was asked to identify a school coordinator to manage the on-site operations of the program.

Students

In September 2008, students were given Capital Gains "parent packets" to take home with them. These packets included:

- A letter from the DCPS Chancellor with details about the program

- A letter from the Capital Gains team with details about the partnership between the program and SunTrust banks
- A parental consent/opt-out form
- A list of frequently asked questions about the program
- An overview of the school-specific metrics
- A program calendar with details about pay periods and payment dates

Washington, D.C. was the only school district that allowed passive consent. Once treatment schools were selected, each student in grades six through eight in those schools was assumed to be part of the program unless a parent consent form was returned indicating the parent did not want their student to participate - in year 1, nine students out of 3,269 were opted out by their parents. The experiments are still on-going.

C. PERFORMANCE METRICS AND INCENTIVE STRUCTURE

Members of the Capital Gains team conducted meetings at each of the treatment schools to explain the program to the school's staff during the first two weeks of school and to help them select the school-specific metrics that would be used to assess and reward their students.

Each school selected three metrics, along with attendance and behavior, which were used to evaluate students. The most popular metrics included homework completion, grades on tests, and wearing a proper uniform. Students could earn up to ten points for each of the five metrics, and each point was worth $2. Teachers kept track of the students' performance for a two week period and rewards were distributed in the week following the close of the previous period. There were a total of 15 two-week periods in the first year.

D. PAYMENT PROCESS

Preparation and Set-up

Student rewards were distributed via direct deposit into savings accounts or by check. Deposits were heavily promoted by schools as the safest distribution method and as a means of encouraging fiscal responsibility and increasing familiarity with banking. In order to set up and deposit funds,

a partnership was formed with SunTrust to create and manage student savings accounts that were interest-earning and child-owned (child is sole custodian).

SunTrust organized "Bank Days" at each of the participating schools at the start of the program. Representatives from the bank visited the schools and signed up students for accounts during their lunch and free periods. All students were required to have a social security number and picture ID before setting up an account. Social security numbers were verified by the Capital Gains project managers who also attended bank days. After establishing their accounts, students signed forms authorizing EdLabs to make direct deposits over the course of the year.

Students and families who could not (no social security number) or would not (unwilling to provide personal information) open saving accounts were paid by check. EdLabs contracted with Netchex, a check processing vendor, to process check payments.

Payment Logistics

Teachers were responsible for filling out hard copy spreadsheets every two weeks. The sheets allowed teachers to record individual student performance on each of the metrics for the two-week reward period. The spreadsheets were shipped to a scanning company which scanned the spreadsheets and sent the images to a data entry company. The data entry company entered all student performance data into electronic spreadsheets that EdLabs project managers accessed via a secure (File Transfer Protocol) site. Once the sheets were downloaded by EdLabs, payment amounts were calculated and audited for accuracy.

Once student payments were calculated and audited, a "pay list" was sent to a payroll vendor. The vendor then accessed a Harvard-owned bank account set up specifically for processing student payment transactions to initiate direct deposits (for those students who signed up for a savings account) and create checks for the remaining students. Those checks were delivered to DCPS project management staff for distribution to school coordinators, who then handed them out to students. In year 1, spreadsheets were collected from teachers on Friday at the end of a two week pay period and checks were delivered the following Thursday. In year 2, teachers were required to enter information into the database by Saturday evening and payments were delivered the following Wednesday.

E. PROGRAM SUPPORT

Throughout the program, targeted strategies were employed to increase participation and awareness and to ensure smooth implementation in all schools.

Student Support

Certificates: Certificates were sent to each participating student displaying the amount of money earned based on their performance on each of their schools' metrics. Certificates both described the student's behavior (e.g. "You were late to class 6 times this pay period") as well as reported the amount earned for each metric.

Assemblies: Schools held school assemblies and/or pep rallies to further introduce the program. School administrators and coordinators used these forums to generate excitement about the program, go over details about earning money and getting paid, and answer any questions students might have.

Knowledge Quizzes: To gage students' understanding of the basic elements of the Capital Gains program, a short quiz was administered to participating students in the fall and spring of year 1. In year 2, students were given quizzes during mandatory financial literacy sessions throughout the school year. A final quiz is planned for administration to students during the spring of 2010.

Check Cashing Letters: "Check-cashing letters" were provided with instructions on free check-cashing options.

Student Survey: At the end of year 1, students were surveyed about their attitude, effort, and motivation in school. The questions were not specific to the programmatic structure of Capital Gains but student responses were included in the analysis.

School Support

Parents' Nights: During the first year of Capital Gains, community forums (or "parents' nights") were held to inform parents of the details of the program, but turnout was low. In year 2, the program manager held information sessions during Back-to-School Night at selected schools.

Materials: Each school also hung posters throughout the building to promote the program and to explain the school-specific performance metrics.

School Communication: Capital Gains project managers contacted all coordinators regularly to confirm that rewards were being distributed in a timely manner, and contacted the principal via e-mail or phone to provide updates on program operations or to address potential concerns.

Coordinator Reports and Graphs: For each pay period, EdLabs sent the school coordinator an overall report that presented data on each of their students' performance (i.e. scores on each metric, consent status, bank account status, and reward history). Coordinators also received lists of the top ten earners in each grade for a given pay period as well as a list of the top ten students with the largest increase in rewards from the last period. Additionally in year 1, schools were provided with graphs that showed how each grade level scored on each of the metrics so they could compare performance across grade levels. Some schools requested that these graphs compare classrooms instead of grades. Halfway through the program and at the end of the year, schools were provided with graphs that showed their performance across periods on each metric so they could see how student performance was changing over time.

School Stipends: Each school received a stipend to help offset the additional work the program created for its staff. The stipend amounts were based on the number of students participating in the program, with small schools receiving around $1,000-$3,000 and the largest school receiving around $20,000. The principal decided whether the funds were to be given to the coordinator or split among the coordinator and other staff members.

Implementation Reviews: In January of year 1, Capital Gains project managers invited all coordinators, principals, and other staff members to complete an on-line survey as part of an effort to further understand the effects of the program. The survey results contain valuable insights and feedback from schools on program implementation and impact. Project Managers also visited each of the schools at the end of the first year to discuss possible improvements for the second year.

8.2 Earning By Learning: An Experiment in Dallas Public Schools

A. BACKGROUND AND OVERVIEW

During the 2005-2006 school year, EdLabs implemented a pilot incentive program in select elementary schools in Dallas Independent School District (DISD). The pilot was based on the Earning by Learning (EBL) of Dallas incentive program. Established in 1996, Earning by Learning provides small monetary rewards to encourage students to read books.

Two important lessons were learned from a small pilot, which informed the methodology and research design of the incentive experiment that was implemented in 2007-2008. First, in the pilot study the randomization of students into treatment and control groups occurred within classrooms;

there was significant evidence that teachers provided compensatory incentives for students in the control group. The decision was made to randomize across schools for the actual experiment.

Secondly, EBL students typically earned their rewards at the end of the semester. While we know of no theoretical model that credibly describes an optimal incentive schedule, related research indicates that children heavily discount the future. It was therefore decided to make every effort to increase the frequency of rewards given to treated students.

During implementation, EdLabs assigned a Boston-based project manger to oversee program implementation. The project manager was dispatched on several occasions for extended periods to increase EBL staff capacity during phases of the experiment which required higher than normal person-hours: school recruitment, technology set-up and testing, payment calculation, and reward distribution.

B. RECRUITMENT AND SELECTION

Schools

Implementation began in mid-July 2007 with a principals' orientation meeting. At this juncture, principals learned about EBL's prior success in encouraging students to read by offering monetary rewards. Forty-three principals attended the meeting or sent a representative in their place. Every principal agreed to participate in a lottery to determine which schools would be in the treatment group and the control group with the treatment group receiving EBL implementation in the 2007-2008 school year.

Principals were recommended to the EBL program by district supervisors and staff at Dallas Independent School District's (Dallas ISD) Instructional Technology department. Together, these administrators recommended schools that had no prior experience with EBL yet had the technological resources to participate in the program. Schools in the control group were told they would be given priority for expansion into the full-service EBL program in the 2008-2009 school year.

An agreement with schools was reached such that each participating school would run the program for their second grade classes, as well as one additional cohort of the principals' choosing from either the first, third, or fourth grades. Principals chose fourth grade. The early elementary grades were selected because EBL administrators, as well as a host of education researchers, have observed that early intervention in reading is most important to future achievement.

EdLabs then performed a computerized randomization, selecting 22 treatment schools and 21 control schools based on their standardized testing performance and demographic data. In schools that were selected into the treatment group, principals selected a member of their staff to serve as EBL coordinator; typically the Media Specialist, Reading Instructor or Technologist.

Students

EdLabs and EBL arranged for consent forms to be distributed and collected at the 22 campuses and delivered to EdLabs for data entry prior to the start of the experiment. Students who had not returned consent forms prior to the start of the experiment were allowed to turn in consent forms throughout the program, but could not be paid for their performance prior to a form being returned. Standard EBL program material and consent forms were sent to families, with slight modifications to make sure the research and changes made to support the experimental design were fully explained.

C. PERFORMANCE METRICS AND INCENTIVE STRUCTURE

The Accelerated Reader platform enabled tracking and assessment – in the form of brief on-line quizzes – of students' reading. Students were rewarded with $2 for each completed Accelerated Reader (AR) quiz with at least 80 percent of questions answered correctly.

The initial incentive structure also included a provision to pay fourth graders $4 for each AR quiz they passed. EBL administrators and veteran EBL coordinators advised EdLabs not to pursue this avenue because they believed a $2 incentive sufficiently motivated fourth graders without causing them to become fixated on the amount of the reward.

D. PAYMENT PROCESS

Preparation and Set-Up

Renaissance Learning (RL) received student information from Dallas ISD's Data and Accountability department to create Accelerated Reader accounts for every 2nd and 4th grader at each of the treatment schools. RL then created a website that allowed EBL coordinators to verify the accuracy of these student accounts. Each school's technologist installed the necessary software and shortcuts on the computers that students would use to be able to access Accelerated Reader.

Once the school's roster had been verified, RL sent students' login information to the EBL coordinator. Our goal was to finish these initial preparations by September 17, the first day of the EBL program. However, initial training for EBL coordinators and technologists did not begin until September 17. Some schools were able to make the necessary preparations before training by relying only on email and phone conversations with EBL, RL, and EdLabs staff. The majority of coordinators began EBL at their school in the two weeks following September 17. During this period, RL, EBL, and EdLabs assisted EBL coordinators and technologists when they had difficulty accessing Accelerated Reader or using the software.

Once students received their account information, they were able to login to the Renaissance Learning website to begin testing. The on-line version of Accelerated Reader has over 100,000 multiple choice quizzes available to students. Since there is only one reading practice quiz per book, to prevent cheating, students were not able to retake quizzes unless their EBL coordinator believed there were mitigating circumstances, such as a fire drill, power outage or medical situation.

Payment Logistics

The initial payment strategy was to reward students with checks every two weeks. The on-line version of Accelerated Reader allowed EdLabs and EBL to automate much of the check writing process, which greatly reduced the cost of processing. However, frequent distributions of rewards via check became a less tenable option for two reasons: high administrative costs incurred by school personnel for distributing thousands of checks several times a semester and concern that lower valued checks would be less likely to be cashed. Instead, checks were distributed three times during the 2007-2008 school year during the weeks of October 17, December 10, and March 17. One additional distribution would clearly be less intrusive on school coordinators than several throughout the term.

EdLabs contracted with Renaissance Learning to develop a macro that would update a student's earnings each time they passed a test using Accelerated Reader. While a student would not have the additional money in hand they would be aware that they had just earned an additional reward.

Payments were issued via check along with instructions as to how students could cash their checks for free in the Dallas area.

E. PROGRAM SUPPORT

Students Support

Celebrations: With the help of the EBL coordinators, EdLabs and EBL organized celebrations between October 17 and 22 to distribute checks to students in the treatment schools. These mid-session celebrations represented a change to the traditional EBL program model. Typically celebrations occurred only at the end of each semester's session. EBL and EdLabs chose the timing of these payments to engender trust among the students, allowing coordinators and reward recipients to encourage non-participating students to return parent consent forms or take AR quizzes before the end of the semester. EdLabs worked with EBL leadership to implement the traditional end of semester celebrations for the fall, beginning the week of December 10, and for the spring, beginning the week of March 17.

Check Cashing Letters: "Check-cashing letters" were provided with instructions on free check-cashing options.

Student Survey: At the end of year following the experiment, students were surveyed to gage the longer-term effects of treatment on their attitude, effort, and motivation.

Program Extension: There was a continued push to collect parent consent forms in order to reward students who were taking quizzes, but were not approved for payments. In an effort to collect more data for the spring session, EdLabs was able to negotiate on behalf of EBL for a two-week extension of the session end date.

School Support

School Communication and Visits: Coordinators in the bottom quartiles with respect to percent of consent forms returned, percent of students taking at least one quiz, and average quizzes taken per student were contacted. Personal visits were made to three schools that were not testing consistently by this point and were not responding to emails or phone calls. In total, this allowed EdLabs to learn about and address administrative or infrastructure challenges that may have been dampening participation. The three schools that required mid-term visits began testing shortly afterward.

School Summaries (Mid-Session): In an effort to improve participation numbers, brief summaries on student participation were sent to each principal and EBL coordinator midway through the fall session. At the same time schools were informed of their percentile ranking relative to all treatment schools. It was expected that this benchmark information would incite schools to

improve relative to their more successful peers. The mid-session summaries were implemented in the spring session as well.

School Summaries (End of Session): At the conclusion of the fall session, EdLabs sent a separate summary (one that was more detailed than the mid-session summaries) to each principal and EBL coordinator. Each individual school receiving a summary could compare itself against the top five schools ranked according to student participation and perhaps even contact them on best practices. These summaries also included customized suggestions to improve participation for each school. These summaries were also sent to select District personnel.

Accelerated Reader: EdLabs purchased school licenses for the web-based version of Accelerated Reader, which allowed schools to access a dynamic and wider range of book quizzes. The software also eliminated the need for school coordinators to send tally sheets of books read to EBL; that information was accessed by EBL and EdLabs directly from the web-based platform.

8.3 The Paper Project: An Experiment in Chicago Public Schools

A. BACKGROUND AND OVERVIEW

In the spring of 2008, Chicago Public Schools Superintendent, Arne Duncan, and EdLabs Faculty Director, Roland Fryer, met to discuss possible collaborations. Reducing drop-out rates was identified as a key component of CPS reform efforts. Preliminary analysis of CPS data revealed a connection between graduation and the number of credits accumulated during freshman year in high school. EdLabs developed an experiment to target 9th graders, entitled "Paper Project," in hope that providing students with incentives would improve course performance, increase the number of credits earned, improve scores on state assessments, and ultimately raise the graduation rate.

A Project Manager and an Assistant Project Manager were designated to serve as project managers for Paper Project.

B. RECRUITMENT AND SELECTION

Schools

EdLabs held a meeting for principals of the seventy high schools with the lowest graduation rates in Chicago Public Schools at the end of their annual back-to-school event in August 2008.

In the meeting we explained the incentive program and its potential benefits, and schools were given until the end of the week to decide whether they wanted to participate. The vast majority of schools signed up immediately, and eventually every school present at the meeting signed up to participate. To control costs, we selected 40 of the smallest schools out of the 70 who wanted to participate and then randomly selected 20 to form the treatment group.

Students

Once a school was selected, students were required to return a signed parent consent form to participate. Students were given a brochure to take home to parents, which described the program and explained the rules of the project as well as the long and short term incentives. A designated coordinator at each school was responsible for collecting consent forms, which were audited periodically by the project manager.

The project managers worked with Area Instructional Officers, principals, and school coordinators to hold a consent form drive to help drive participation rates at schools. Schools responded with creative methods for increasing consent rates. For example, flyers were sent home with students multiple times, robo-calls were made to all 9th graders, parent assemblies were held, and program information was provided during report card pick up times. Ultimately, 95 percent of eligible students signed up for Paper Project.

C. PERFORMANCE METRICS AND INCENTIVE STRUCTURE

When Paper Project was conceived, CPS had just adopted Gradebook, a system where teachers would enter grades daily. Every 5 weeks, grades were issued to students (progress report, quarter grade, progress report, semester grade, repeat). These grades served as the basis of students' earnings.

Students were provided incentives for their grades in five core courses: English, mathematics, science, social science, and gym. Each student was rewarded with $50 for each A, $35 for each B, $20 for each C, and $0 for each D. If a student made an F (failing) in a core course, he received $0 for that course and temporarily "lost" all other monies earned from other courses in the grading period. Once the student made up the failing grade in the next period, through credit recovery, or night school (we could not pay out for summer school credit recovery because the program ended earlier than planned) all the money "lost" was reimbursed. Students could earn $250 every five

weeks and $2,000 per year. Half of their rewards were paid out immediately after the five week grading periods ended, while the other half was held in an account and will be given in a lump sum conditional upon high school graduation. The average student earned $695.61; the highest achiever earned $1,875.

D. PAYMENT PROCESS

Preparation and Set-Up

Student rewards were distributed via deposit into savings accounts or in the form of a personal check. Deposits were heavily promoted by schools as the safest distribution method and as a means of encouraging fiscal responsibility and increasing familiarity with banking. In order to set up and deposit funds into student savings accounts, a partnership was formed with JP Morgan Chase Bank. Through this partnership, students were able to open interest-earning child-owned (child is sole custodian) savings accounts.

In order to open an account, bank officials visited each school, where students completed a bank form with mailing address, social security number, and birth date. An account was opened immediately after processing.

Students and families who could not (did not have social security number) or would not (unwilling to provide personal information) open savings accounts were rewarded via personal checks. EdLabs contracted with Netchex, a check processing vendor, to process check payments.

A corporate bank account was set up in the name of the American Inequality Lab at Bank of America. Payment funds were requested, and then wired, from the grant holding agent (National Bureau of Economic Research) to the bank account, from which payments were drawn.

Payment Logistics

Grades were uploaded from Gradebook to IMPACT (CPS' system of record). CPS's IT group pulled the grades from IMPACT and uploaded them to an FTP site from which they could be retrieved by EdLabs. After processing the grades, EdLabs calculated award amounts, uploaded payment information to Netchex (check processing vendor), and generated student certificates. The project manager had certificates printed at CPS' internal print shop, and sent out certificates to schools via CPS' internal mail system.

EdLabs then uploaded grades to the UBOOST system. If students came to coordinators with questions or issues regarding their grades or payments, coordinators would verify student claims and determine whether adjustments needed to be made. Coordinators (or project manager) would have about two weeks to make any grade corrections in the UBOOST system. UBOOST fed changes directly back to Edlabs for incorporation into the next round of payments to be issued. EdLabs verified all payment amounts through extensive internal audits.

Coordinators received checks via FedEx and certificates via CPS mail. Most coordinators organized checks by division (homeroom) and had homeroom teachers pass out certificates and checks to students during division period. Some coordinators passed out checks after school hours on Fridays to minimize chances for issues on school property. Student certificates contained information on the amount earned (x out of a possible $250), grade for each class, a reminder that they would only receive half the money now (the remainder would be paid out upon graduation), notes regarding makeup grades or missing grades, as well as a reminder that checks could be cashed for a small fee ($6) at any Bank of America.

E. PROGRAM SUPPORT

Student Support

Certificates: Certificates were sent to each student, including non-earners, displaying the amount of money earned based on test scores. Students, who had not yet returned consent forms, were informed that all earned monies would be distributed once a valid consent form was submitted.

Student Assemblies: Many coordinators arranged celebrations/assemblies throughout the year to reinforce the positive messages behind the program.

Knowledge Quizzes: To gage students' awareness of the basic elements of Capital Gains and the incentive structure, a short quiz was administered to participating students.

Check Cashing letters: "Check-cashing letters" were provided with instructions on check-cashing options.

Student Survey: At the end of the year, students were surveyed about their attitude, effort, and motivation in school. The questions were not specific to the programmatic structure of the Paper Project.

Student and Family Communication: Students and families received ongoing updates and support from school-based coordinators. Many coordinators sat down with students to discuss what they could do to begin earning rewards or to earn more than they were already. Questions from parents were responded to on a daily basis.

School Support

Materials: Creative bulletin boards and other celebratory messaging encouraged and reminded students about the program.

School Communication: The project managers managed communication and coordination with schools on a daily basis, made several school visits each month, handled student-specific inquiries, and developed a newsletter to disseminate program information as well as studying tips, etc.

School Stipend: Participating schools received up to $1,500 to provide a bonus for the school liaison that served as the main contact for our implementation team.

8.4 Spark: An Experiment in New York City Public Schools

A. BACKGROUND AND OVERVIEW

During the 2003-2004 school year, Roland Fryer, Richard Freeman, and Alexander Gelber implemented a pilot incentive program at PS 70, an elementary school in the Bronx, NY. The pilot informed the design of future incentive projects related to: type of reward (monetary vs. social), frequency and timeliness of rewards, level of randomization (within vs. across schools), and teacher and administrative expectations.

During the 2007-2008 and 2008-2009 school years, EdLabs implemented NYC Spark to measure the impact of incentives on student achievement. In year 1, the program was administered to 4th and 7th grade students in a sample of New York City public schools. In year 2, eligibility expanded to 8th graders in about half of the participating middle schools to test two-year treatment effects.

During the first year, EdLabs assigned two Boston-based project mangers to oversee program implementation, and one to oversee the payment process. In the second year, EdLabs' project managers were stationed at the New York City Department of Education in order to be more

accessible to schools via in person visits, local phone and fax numbers and a familiar DOE email address.

B. RECRUITMENT AND SELECTION

Schools

Implementation of the Spark program began on June 1, 2007 when EdLabs sent a letter inviting all schools with 4th and 7th graders in the New York City Public School District to participate. Roland Fryer then visited the 143 schools that expressed initial interest in response to our letter. All 143 schools signed up for the experiment. Twelve schools were removed due to their participation in other Opportunity NYC programs. Sixty-three schools were randomized into treatment. The schools that were not selected served as a control group and were told that they would be given first priority if the NYC Department of Education decided to expand the program.

The principals of the participating schools were invited to an orientation meeting in August 2007 and again in August 2008. Each principal was asked to select a Spark coordinator from their staff (typically the Assistant Principal or the Assessments Administrator) to serve as the point of contact between students and the Spark Program.

Students

Principals were given information packets with program details, consent forms, and applications for bank accounts to be distributed to students at the start of each school year. Schools were provided with FedEx account numbers to facilitate the return of consent forms to EdLabs. During year 2, EdLabs contracted with marketing firm Droga5 to produce T-shirts, pencils, and various other items with the program logo in order to generate excitement within the student population about the program.

Students were allowed to turn in consent forms until just before the last testing period of each school year. In year 1, there were over 5,800 participating students (roughly 70 percent of the more than 8,300 eligible students). In year 2, over 8,000 students (80 percent of about 10,000 eligible students) participated in the program.

C. PERFORMANCE METRICS AND INCENTIVE STRUCTURE

Students in Spark schools completed ten assessments, five in reading and five in math, over the course of each year. Of these tests, six were computer-adaptive assessments and four were predictive assessments. All of the assessments were administered through the Department of Education's Periodic Assessment Team, allowing schools to fulfill their city assessment requirements.

The computer-adaptive assessments respond to each question a student answers by generating an appropriate next question based on how the student responds. In effect, these assessments "zero-in" on a student's performance level and set an empirical benchmark from which the student's progress could be tracked. The computer-adaptive assessments indicate the instructional level at which the student performs, regardless of their current grade level placement in reading, language arts, and math.[50] Student results include a scaled score, ranging from 1300-3700, an associated grade level estimate, a comparison to NY state standards, and a national percentile rank. In addition, the results also offer a Standard Item Pool score (SIP), which expresses the expected percent of questions within the grade level item pool that the student could answer correctly.

The predictive assessments mirror the state exams, provide an indication of student progress against state standards, and provide diagnostic information to inform instruction. The first assessment, administered 6-8 weeks before the New York State exam, assesses what short-term tailored interventions students need before the state exam. The second assessment, administered in late May or early June, measures the student's growth within and across academic years and serves to collect information for instructional planning for the following year. Reports provide detailed diagnostic information by standard, performance indicator, sub-skill, level of difficulty of the item, and each student's correct and incorrect answers. These analyses help target instruction and intervention to accelerate student learning. A scaled score is provided to us along with a prediction of how the student may perform on the state exam.

In year 1, each 4th and 7th grader earned \$5 and \$10, respectively, for simply taking each assessment. Based on feedback from schools, this base payment was instituted to encourage poorly

[50]There are four additional reasons that schools used Scantron computerized tests. First, we were concerned that in high poverty schools administrators might cheat to help students earn extra money. Second, computerized tests were the only assessment option that allowed us to grade and record the exams in a timely manner. Third, ITAs are created by staff at each school based on how quickly students are progressing through the curriculum. We were worried about making the level of the test endogenous. Fourth, the NYC DOE was concerned that their efforts to ensure that the interim assessments were low stakes test would face criticism if monetary rewards were tied to them.

performing students. For each assessment, a 4th grader could earn a maximum of $25 per assessment for a total of $250 over the course of the year, and a 7th grader could earn a maximum of $50 per assessment for a total of $500 over the course of the year. In year 2, 8th graders could earn the same amount as 7th graders, a maximum of $50 per assessment for a total of $500 over the course of the year. The overall reward scale between the lower and upper bounds was linear.

For Computer-Adaptive assessments, each 4th grade student received $5 as a base payment plus $0.025 for each additional point gained from 1900-2700. Each 7th and 8th grade student received $10 plus $0.05 for each additional point gained from 2200-3000.

For Predictive assessments, each 4th grade student received $5 as a base payment and an additional $0.20 for answering an additional 1% of the questions correctly. Each 7th and 8th grader earned a base amount of $10 plus $0.40 for correctly answering an additional 1%.

D. PAYMENT PROCESS

Preparation and Set-up

Student rewards were distributed via deposit into savings accounts or by check. Deposits were heavily promoted by schools as the safest distribution method and as a means of encouraging fiscal responsibility and increasing familiarity with banking. In order to set up and deposit funds into student savings accounts, a partnership was formed with Washington Mutual (now part of J.P. Morgan Chase) to operate a derivative of their School Savings account program. A School Savings account is an interest-earning child-owned (child is sole custodian) savings account.

In order to open an account, students completed a bank form with mailing address, social security number, and birth date. Those forms were sent to EdLabs to screen for completeness and then forwarded to Washington Mutual. An account was opened approximately two weeks after Washington Mutual received a valid completed application.

Students and families who could not (no social security number) or would not (unwilling to provide personal information) open saving accounts were rewarded with individual checks. EdLabs contracted with Netchex, a check processing vendor, to process check payments.

A corporate bank account was set up in the name of the American Inequality Lab at Bank of America. Payment funds were requested, and then wired, from the grant holding agent (National Bureau of Economics) to the bank account, from which payments were drawn.

Payment Logistics

After each assessment was taken, EdLabs requested and received scores from New York City DOE. The test scores were merged with the existing New York City student enrollment database, and internal records on consent and method of payment – savings or check. EdLabs then generated a pay list based on the payment algorithm described in section IV.

In year 1, EdLabs manually entered deposit amounts into the School Savings interface. By year 2, EdLabs created a script allowing this process to be automated. EdLabs then sent a check via FedEx to Washington Mutual to cover the deposits. The deposits were available to the students on the next business day. Netchex mailed all personal checks to the Spark coordinator at each school for distribution.

Due to the test vendor's processing time, EdLabs typically received test scores within four to five weeks of the completion of student testing. Students typically received rewards within one to two weeks of EdLabs receiving assessment data.

In year 1, total payments exceeded $1.1 million. In year 2, payments were over $1.6 million.

E. PROGRAM SUPPORT

Student Support

Throughout the program, targeted strategies were employed to increase participation and awareness and to ensure smooth implementation in all schools.

Certificates: Certificates were sent to each student, including non-earners, displaying the amount of money earned based on test scores. Students who had not yet returned consent forms, were informed that all earned monies would be distributed once a valid consent form was submitted.

Knowledge Quizzes: To gage students' awareness of the basic elements of the Spark program, a short quiz was administered to participating students in year 2.

Check Cashing letters: "Check-cashing letters" were provided with instructions on free check-cashing options.

Student Survey: At the end of year 2, students were surveyed about their attitude, effort, and motivation in school. The questions were not specific to the programmatic structure of Spark but student responses were included in the analysis.

School Support

School Visits: In year 1, each school received at least two visits from EdLabs' project managers. In year 2, EdLabs' project managers targeted schools with low participation to be visited. All schools, regardless of participation, were visited upon request. School visits consisted of hosting assemblies to provide program overviews, distributing certificates, speaking with students in classrooms or during lunch, presenting to parents during parent-teacher conferences, and meeting with teachers and school leadership team members

School Communication: EdLabs' project managers contacted all coordinators regularly to confirm that rewards were being distributed in a timely manner, and contacted the principal or network leader via e-mail or phone to provide updates on program operations or to address potential concerns.

Coordinator Reports: EdLabs sent each coordinator an overall report that presented data on each of their students (i.e. test scores, consent status, bank account status, and reward history).

Tracking Database: In year 2, EdLabs worked with Uboost, an external vendor, to develop an internet-accessible database which provided real time information similar to "Coordinator Reports": consent status, bank account status and earned rewards.

School Stipends: Each school was given the opportunity to receive up to $5,000 for each year of the two-year program to help offset the added work the program created for its staff. Each year there were two allocations of up to $2,500 per school, one at the middle of the year and the other after the end of year.

In the first year, for the first $2,500, schools had to have turned in at least one student's consent form to show that they intended to participate in the program. In order to receive the full second installment of $2,500, EdLabs required that 60 percent of a school's 4th or 7th grade students received rewards for the third computer-adaptive assessment. If a school was below this 60 percent benchmark, the school received a pro-rated amount.

In the second year, the first $2,500 was contingent upon a school achieving an 80 percent participation rate and having tested at least 70 percent of students on each Spark-eligible assessment. The second $2,500 was contingent on having tested at least 70 percent of students on the remaining Spark-eligible assessments.

Implementation Reviews: On three separate occasions, twice in September 2007, and once in January 2008, EdLabs' project managers hosted an on-line chat room for principals and coordi-

nators to ask and get answers to program implementation questions.

In late March of the first year, EdLabs' project managers conducted 40-minute on-site interviews at each school to review program implementation. A list of best practices was compiled from these interviews and shared with all schools before the start of the second year of the program.

In February of the second year, EdLabs' project managers invited all coordinators, principals and other staff members to complete an on-line survey as part of an effort to further understand the effects of the program. The survey results contain valuable insights and feedback from schools on program implementation and impact (Appendix B).

9 Appendix B: Data Description and Construction of Variables

9.1 VARIABLE CONSTRUCTION

Chicago

Assignment to treatment/control at the beginning of the year

Chicago Public Schools gave us a file containing the IDs of the students who were assigned to the treatment group and a separate file for students in the control group. We combined these files and used them to assign students to treatment or control.

Instrument in TOT regressions

Students who were originally assigned to treatment/control were given their original assignments. Any other student who ever attended a control or treatment school according to the Chicago Public Schools 2009 attendance file, and was in 9th grade according to the Chicago Public Schools 2009 enrollment file, was given a value of zero for the TOT instrument.

Amount of time treated in TOT regressions

This was constructed by adding up all of the days that a student was present in a treatment school if the student was in 9th grade, and gave consent (according to a Chicago Public Schools file listing all students who gave consent). This number was divided by 180 (approximately the number of school days in the year; 180 is used to be consistent with other cities). Unlike other districts, the CPS attendance file did not contain the grade level of the student for each observation, so we could not figure out which students may have skipped into the 9th grade or been demoted into the 9th grade. For this reason, we could only consider students that were assigned to the 9th grade in

the enrollment file to have been treated.

Demographic variables

Demographic variables that should not vary from year to year (race, gender) were pulled from several different Chicago Public Schools enrollment files, with precedence given in the following order: 2010 enrollment file, 2009 enrollment file, 2009 district file, 2009 bilingual students file. Race consisted of the following categories: Black, Hispanic, White, Asian, and Other. These categories were considered mutually exclusive.The "Other" category consisted of students who were coded as Multiracial or Native American. Gender was coded as male, non-male, or missing.

Demographic variables that may vary from year to year (free lunch, English language learner) were only pulled from Chicago Public Schools 2009 files, with precedence given in the following order: 2009 enrollment file, 2009 district file, 2009 bilingual students file. A student was considered free lunch if any of the files had a "F"or "R"for the student's lunch status or was entered as "Eligible For Free Meals"or "Eligible for Reduced Price Meals."All other entries("D","FT", "N/A", "Denied - Income Over Allowable Limit", and "Not Applying Or Refusal To Cooperate With") were considered non free lunch and blanks were coded as missing. The English Language Learner dummy variable was coded as one if the students had "Yes"entered under the bilingual variable. All other values, including blanks, were coded as zero.

Illinois Standards Achievement Test (ISAT) Scores

ISAT scores for math, reading, science and writing were pulled from a file listing scores for all students in Chicago Public Schools. Eighth graders do not take the science portion of the test, and we decided to use only math and reading scores to keep the analysis consistent across districts. Scores were standardized to have mean of zero and standard deviation of one in the experimental group.

PLAN Test Scores

PLAN test scores were pulled from the Chicago test score file for 2009-10. Scores were standardized to have mean of and standard deviation of one in the experimental group (the experimental group mean was subtracted from each student's score, and the result was divided by the experimental group standard deviation).

Grades

Grades were pulled from files containing the transcripts for all students in Chicago Public

Schools. Observations were at the student-course-semester level in the raw data. Letter grades were converted to a 4.0 scale. Each student's grades were averaged within a semester to yield a Fall, Spring, and Summer (where available) GPA. Each student's GPA for the 2008-09 school year was calculated by taking the mean of the Fall, Spring, and Summer GPAs. As with ISAT test scores, GPAs were standardized to have mean of zero and standard deviation of one in the experimental group.

School-level variables

School-level variables were constructed for each school based on the population of students assigned to that school, using the following method to assign students and the Chicago Public Schools 2009 attendance data file: If a student attended only one school, then he was assigned to this school. If a student attended more than one school, then he was assigned to the school he attended for the most number of days. If a student attended multiple schools for the same maximum number of days, then he was assigned to the school with the lower school ID. We would have ideally liked to assign students to the schools they attended at the beginning of the year, but the format of the attendance data in Chicago did not provide enough information to do this. The data file only had the total number of days enrolled and number of days present in each school for every student; it did not contain any information on which months the student was present.

Attendance Rate

Each student's attendance rate was calculated as the total number of days present in any school divided by the total number of days enrolled in any school, according to the CPS 2008-09 attendance file.

Credits attempted, credits earned, and absences

These were pulled directly from files containing transcripts for all students in CPS from each semester.

Dallas

Assignment to treatment/control at the beginning of the year

The Dallas 2007-08 attendance file was used to determine the first school that each student attended. A student was assigned to the control group if the student's first school was a control school, if the student was in 2nd grade in that school, and if the student was present in that school

before October 1, 2008. The treatment group was defined similarly.

Instrument in TOT regressions

Students who were originally assigned to treatment/control were given the same assignments. Any other student who ever attended a control or treatment school according to the Dallas 2007-08 attendance file and was in 2nd grade while attending that school was assigned a value of zero for the TOT instrument.

Amount of time treated in TOT regressions

This was constructed by adding up all of the days that a student was present in a treatment school, was in 2nd grade in that school, and gave consent (according to a Dallas file listing all students who gave consent). This number was divided by 179 (the most school days attended by any student).

Demographic variables

Demographic variables that should not vary from year to year (race, gender) were pulled from Dallas enrollment files from 2002-03 through 2008-09, with precedence given to the most recent files first. Race consisted of the following categories: Black, Hispanic, White, Asian, and Other. These categories were considered mutually exclusive. The "Other" category consisted of students who were coded as "I", which was assumed to stand for American Indian (or Native American). Gender was coded as male, non-male, or missing.

Demographic variables that may vary from year to year (free lunch, English language learner, special education) were pulled only from the Dallas enrollment file for 2007-08, the experimental year. A student was considered free lunch if he had a 1 for the lunch variable in the enrollment file for 200708. A student was considered not free lunch if he was in the 2007-08 enrollment file, but had a blank value for the lunch variable(lunch was coded exclusively as 1 or blank in the raw data). If a student was not in the 2007-08 enrollment file, then he was coded as having a missing free lunch value. English Language Learner status was coded similarly, as the raw data has a variable for limited English that was coded as "Y"or blank. Special education had values of zero and one in the raw data, so a one was coded as one for the dummy variable and zero was coded as zero. Blanks were considered missing values for that variable.

Assignment to ITBS vs. Logramos sample

Students were assigned to the ITBS sample if they had at least one non-missing score from the

Reading Comprehension, Reading Vocabulary, or Language Total ITBS 2007-08 tests. Similarly, students were assigned to Logramos if they had at least one non-missing score from the Reading Comprehension, Reading Vocabulary, or Language Total Logramos 2007-08 tests.

In cases where a student had non-missing scores from both ITBS and Logramos tests, the student was assigned to the Logramos category and the ITBS scores were dropped. In cases where students switched between Logramos and ITBS from 2006-07 to 2007-08, the 2006-07 scores of those students who switched were dropped and the student was considered a part of the sample for the test he took in 2007-08.

Missing 2007-08 test scores

If a student was missing Reading Comprehension, Reading Vocabulary, and Language Total scores from 2007-08, then he is dropped from both samples since he is missing outcomes.

Grades

Grades were pulled from files containing the transcripts for all students in Dallas from 2007-08. Student-course-semester observations that had values for grades that were non-numeric and could not obviously be converted to numeric values were dropped. Each student's grades from each semester were averaged to yield a GPA for the semester. These semester GPAs were averaged to yield a GPA for the year. The year-long GPAs were standardized to have mean of zero and standard deviation of one in the experimental group.

School-level variables

School-level variables were constructed for each school based on the population of students assigned to that school, using the following method to assign students: If a student attended only one school, then he was assigned to that school. If a student attended multiple schools and attended at least one of these schools at the beginning of the school year, defined as before October 1, then we assigned him to the first school he attended. This first school is determined by the looking at the earliest start date for each student. If a student had the same start date at multiple schools, then the school with the earliest end date was assumed to be his first school. If we couldn't figure out which school the student attended first, we assigned him to the school with the lower school ID. If the student attended multiple schools but did not attend any of these schools at the beginning of the year, he was assigned to the school he attended for the greatest number of days. If there was a tie for the greatest number of days, then he was assigned to the school with the lower school

ID. These school-level variables included the percents of the student population that were Black, Hispanic or classified as free lunch.

Construction of effort index

Effort variables were constructed from a file containing survey responses that were entered into a computer(Dallas was the only district in which students took surveys on the computer). If a student took the survey multiple times and had more than one response to a particular question and these responses conflicted, those responses were dropped. If the responses were consistent or only filled in for one observation, then the values were kept. All question responses were converted to a numerical scale so that higher numbers indicated more effort. There were two effort variables constructed from two questions on the survey. A total effort index equal to the sum of these effort variables (when they were both non-missing) was constructed. This effort index was standardized to have mean of zero and standard deviation of one in the experimental group.

Construction of intrinsic motivation index

Intrinsic motivation variables were constructed from a file containing survey responses that were entered into a computer. If a student had more than one response to a particular question and these responses conflicted, the responses were dropped. All question responses were converted to a numerical scale so that higher numbers indicated more intrinsic motivation. There were six intrinsic motivation variables in total. A total intrinsic motivation index equal to the sum of these intrinsic motivation variables (when they were all non-missing) was constructed. This intrinsic motivation index was standardized to have mean of zero and standard deviation of one in the experimental group, but regressions were also run on the non-standardized index because the index itself has meaning.

Washington, D.C.

Assignment to treatment/control at the beginning of the year

The DC 2008-09 attendance file was used to determine the first school that each student attended. A student was assigned to the control group if the student's first school was a control school, if the student was in 6th, 7th, or 8th grade, and if the student was present in that school before October 1, 2008. The treatment group was defined similarly.

Instrument in TOT regressions

Students who were originally assigned to treatment/control were given the same assignments. Any other student who ever attended a control or treatment school and was in 6th, 7th, or 8th grade at that school according to the DC 2009 attendance file was assigned a value of zero for the TOT instrument.

Amount of time treated in TOT regressions

The amount of time treated was constructed by adding up all of the days that a student was present in a treatment school if the student was in 6th, 7th, or 8th grade (DC had passive consent rates of around 98 percent, so whether or not each student had consent was not determined). This number was divided by 180 (the number of school days in the year).

Demographic variables

Demographic variables that should not vary from year to year (race, gender) were pulled from the following files (in order of precedence): 2008-09 DCPS enrollment file, 2008-09 DCCAS file, 2007-08 DCPS enrollment file, 2006-07 DCPS enrollment file. Race consisted of the following categories: Black, Hispanic, White, Asian, and Other. These categories were considered mutually exclusive.The "Other" category consisted of students that were coded as "American Indian." Gender was coded as male, non-male, or missing.

Demographic variables that may vary from year to year (free lunch, English language learner, and special education) were only pulled from the 2008-09 DCPS enrollment file and the 2008-09 DCCAS file. A student was considered free lunch if he was coded as "Free"or "Reduced" in the DC enrollment file. A student was considered not free lunch if he was coded as "Pay All" in the lunch status variable. All blanks were coded as missing. For English Language Learner status, a student was given a value of one if he had a status of "ELL," "ELL Level 1"-"ELL Level 4," or "ELLm (Return to ESL)" in the enrollment file or a "Y" in the ELL variable in the 2008-09 DCCAS file. All blanks in the enrollment file were coded as non-ELL and had a value of zero as were "N"in the DCCAS file. Special education was coded similarly with a one for "Special Education" or "Referred" in the 2008-09 enrollment file or "Y" for the special education variable in the 2008-09 DCCAS file. A student was considered to be missing free lunch, ELL, or special education data if he was not in the DC enrollment file and did not have values for these variables in the 2008-09 DCCAS file.

DCCAS Test Scores

These were pulled from the DCCAS 2008-09 and 2007-08 files. Scores were standardized by grade level to have mean of zero and standard deviation of one in the experimental group (the experimental group mean was subtracted from each student's score, and the result was divided by the experimental group standard deviation). Proficiency levels were also taken directly from these files.

Grades

Grades were pulled from files containing the transcripts for all students in DCPS from 2008-09 and 2007-08. Letter grades were converted to a 4.0 scale. Each student's grades from each semester were averaged to yield a GPA for the year. As with the DCCAS test scores, GPAs were standardized to have mean of zero and standard deviation of one in the experimental group.

School-level variables

School-level variables were constructed for each school based on the population of students assigned to that school, using the same method of assignment described in the School-level variables in Dallas section of the Appendix. The first school was determined by looking at the pattern of attendance across schools. These school-level variables included the percents of the student population that were black, Hispanic or classified as free lunch. Additionally, a dummy variable was constructed to capture the type of school. Schools in the experiment were either traditional middle schools with grades six through eight or elementary schools with grades from pre-kindergarten to eighth grade. The dummy variable was set to zero if the school was a traditional middle school or one if it was a PreK-8 school.

Behavior

The number of behavioral incidents for each student was pulled from a DCPS file listing all behavioral incidents from 2008-09. Students not listed in this file were assumed to have zero behavioral incidents. The total number of behavioral incidents for each student was summed and then standardized to have mean of zero and standard deviation of one in the experimental group. The number of behavioral incidents for 2007-08 was constructed similarly using a DCPS file listing all behavioral incidents from 2007-08. School-level behavior variables were created by summing behavioral incidents by school rather than by student. A school was assumed to have zero incidents if there were no students who attended that school in the behavioral incident file.

A dummy variable was constructed for bad behavior in 2007-08, with the dummy equal to zero

if a student had had zero behavioral incidents in 2007-08, and equal to one if the student had had any behavioral incidents in 2007-08.

Attendance Rate

Each student's attendance rate was calculated as the total number of days present in any school divided by the total number of days enrolled in any school, according to the DCPS 2008-09 attendance file. Each student's attendance rate from 2007-08 was calculated similarly using the DCPS 2007-08 attendance file.

Construction of effort index

Effort variables were constructed from a file containing paper survey responses that were manually entered into a computer. If a student had more than one response to a particular question and these responses conflicted, those responses were dropped. All question responses were converted to a numerical scale so that higher numbers indicated more effort. There were nine effort variables in total. A total effort index equal to the sum of these effort variables (when they were all non-missing) was constructed. This effort index was standardized to have mean of zero and standard deviation of one in the experimental group.

Construction of intrinsic motivation index

Intrinsic motivation variables were constructed from a file containing paper survey responses that were manually entered into a computer. If a student had more than one response to a particular question and these responses conflicted, those responses were dropped. All question responses were converted to a numerical scale so that higher numbers indicated more intrinsic motivation. There were seven intrinsic motivation variables in total. A total intrinsic motivation index equal to the sum of these intrinsic motivation variables (when they were all non-missing) was constructed. This intrinsic motivation index was standardized to have mean of zero and standard deviation of one in the experimental group, but regressions were also run on the non-standardized index because the index itself has meaning.

New York City

Assignment to treatment/control at the beginning of the year

The New York 2008-09 attendance file was used to determine the first school that each student attended. A student was assigned to the control group if the student's first school was a control

school, if the student was in 4th or 7th grade, and if the student was present in that school before October 1, 2008. The treatment group was defined similarly.

Instrument in LATE regressions

Students who were originally assigned to treatment/control were given the same assignments. Any other student who ever attended a control or treatment school and was in 4th or 7th grade according to the NYC 2009 attendance file was assigned a value of of zero for the LATE instrument.

Amount of time treated in LATE regressions

The amount of time treated was calculated by adding up the number of the days that a student was present in a school that received treatment (as opposed to a school that was assigned treatment), was in the relevant grade and had consent to participate in the program. This number was divided by 182 (the number of school days in the year). Estimating the number of days in this way accounts for both schools that were assigned treatment and did not participate and schools that were assigned to the control group but ended up receiving treatment.

Demographic variables

Demographic variables that should not vary from year to year (race, gender) were pulled from New York City enrollment files from 2003-04 through 2008-09, with precedence given to the most recent files first. Race consisted of the following categories: Black, Hispanic, White, Asian, and Other. These categories were considered mutually exclusive.The "Other" category consisted of students that were coded as "American Indian". Gender was coded as male, non-male, or missing.

Demographic variables that may vary from year to year (free lunch, English language learner, and special education) were only pulled from the 2008-09 NYC enrollment file. A student was considered free lunch if he was coded as "A"or "1"in the raw data, which corresponds to free lunch or "2"which corresponds to reduced-price lunch. A student was considered non free lunch if the student was coded as a "3"in the NYC enrollment file which corresponds to Full Price. All other values, including blanks, were coded as missing. For English Language Learner status, a student was given a value of one if he was coded as "Y"in the limited English proficiency variable. All other students in the NYC 2008-09 enrollment file were coded as zero for English Language Learner status. Special education was coded similarly.

New York State Test Scores

NYC state test scores were pulled from the NYC test score files for 2008-09 and 2007-08.

Scores were standardized by grade level to have mean of zero and standard deviation of one in the experimental group (the experimental group mean was subtracted from each student's score, and the result was divided by the experimental group standard deviation). Proficiency levels were also taken directly from these files.

Grades

Grades were pulled from files containing the transcripts for all students in NYC from 2008-09 and 2007-08. Grades were averaged to create a yearly GPA for each student. If a student had multiple grades for the same course in the same semester (usually because he switched schools) both grades were used. As with test scores, GPAs were standardized to have mean of zero and standard deviation of one in the experimental group.

School-level variables

School-level variables were constructed for each school based on the population of students assigned to that school, using the same method of assignment described in the School-level variables in Dallas section of the Appendix. The method of determining the first school was slightly more complicated than in other cities because of the format of the attendance data for 2008-09. In previous years' data, observations were student-school level and had the number of presents and days enrolled in each month. In 2008-09 the data was still student-school level, but each month kept track of the cumulative number of days a student had spent in NYC public schools over the year. We assumed that a student attended first the school with the lowest number of cumulative days present. These school-level variables included the percents of the student population that were black, Hispanic or classified as free lunch.

Behavior

The number of behavioral incidents for each student was pulled from a NYC file listing all behavioral incidents from 2008-09. Students not listed in this file were assumed to have zero behavioral incidents. The total number of behavioral incidents for each student was summed and then standardized to have mean of zero and standard deviation of one in the experimental group. The number of behavioral incidents for 2007-08 was constructed similarly using a NYC file listing all behavioral incidents from 2007-08. School-level behavior variables were constructed similarly, but behavioral incidents were summed within a school rather than at the individual level.

A dummy was constructed for bad behavior in 2007-08, with the dummy equal to zero if a

student had had zero behavioral incidents in 2007-08, and equal to one if the student had had any behavioral incidents in 2007-08.

Attendance Rate

Each student's attendance rate was calculated as the total number of days present in any school in NYC divided by the total number of days enrolled in any school, according to the NYC 2008-09 attendance file. Each student's attendance rate from 2007-08 was calculated similarly using the NYC 2007-08 attendance file.

Construction of effort index

Effort variables were constructed from a file containing paper survey responses that were manually entered into a computer. If a student had more than one response to a particular question and these responses conflicted, those responses were dropped. All question responses were converted to a numerical scale so that higher numbers indicated more effort. There were nine effort variables. A total effort index equal to the sum of these effort variables (when they were all non-missing) was constructed. This effort index was standardized to have mean of zero and standard deviation of one in the experimental group.

Construction of intrinsic motivation index

Intrinsic motivation variables were constructed from a file containing paper survey responses that were manually entered into a computer. If a student had more than one response to a particular question and these responses conflicted, those responses were dropped. All question responses were converted to a numerical scale so that higher numbers indicated more intrinsic motivation. There were seven intrinsic motivation variables in total. A total intrinsic motivation index equal to the sum of these intrinsic motivation variables (when they were all non-missing) was constructed. This intrinsic motivation index was standardized to have mean of zero and standard deviation of one in the experimental group but regressions were also run on the non-standardized index because the index itself has meaning.

9.2 STATE ASSESSMENTS

Dallas

In May of every school year, students in Dallas ISD elementary schools take either the Iowa Tests of Basic Skills (ITBS) or Logramos if they are in kindergarten, first or second grade. Logramos is

a Spanish academic achievement test that takes the place of the ITBS for all students with limited English proficiency. Both tests have similar formats. The tests have four subject areas: reading comprehension, reading vocabulary, language arts, and mathematics. Students receive scores in each of these four areas. ITBS/Logramos scores are an estimate of the grade level of the student taking the test. For example, a score of 2.5 is what the average second grade student would score if she took the test in the middle of the school year. Students in Dallas take the Texas Assessment of Knowledge and Skills beginning in 3rd grade. This test has reading and math sections and is administered in both English and Spanish.

New York City

The state mathematics and English Language Arts tests, developed by McGraw-Hill, are exams conducted in the winters of third through eighth grade.[51] Students in third, fifth, and seventh grades must score proficient or above on both tests to advance to the next grade. 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.[52]

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 who are Limited English Proficient 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.

Washington, D.C.

The DC-CAS is the D.C. Comprehensive Assessment System and is administered each April to students in grades 3 through 8 and 10. It measures knowledge and skills in reading and math. Students in grades 4, 7, and 10 also take a composition test; students in grades 5 and 8 also take

[51] Sample tests can be found at http://www.emsc.nysed.gov/osa/testsample.html.

[52] Content breakdown by grade and additional exam information is available at http://www.emsc.nysed.gov/osa/pub/reports.shtml.

a science test; and students in grades 9-12 who take biology also take a biology test.[53]

In 2008-09, 49 percent of students were proficient in reading and 49 percent were proficient in math at the elementary level. At the secondary level, 41 percent were proficient in reading and 40 percent were proficient in math.

Chicago

Chicago high school students take the PLAN assessment created by ACT in late September/early October of their sophomore year. The test is comprised of four academic achievement tests in Reading, English, Math and Science Reasoning. Each of the tests consists of 25 to 50 multiple choice questions that are curriculum based. PLAN also contains four components designed to help students prepare for the future, the "Needs Assessment," which collects information about students' perceived needs for help; the High School Course and Grade Information, which gathers lists of completed courses; the UNIACT Interest Inventory, which helps students explore possible career options; and the Education Opportunity Service, which links students with relevant colleges and scholarship programs. The test is used to "provide baseline information at 10th grade about student readiness for college and to assist in educational and career planning." PLAN is a "pre-ACT" test and a good predictor of student performance on the ACT portion of the Prairie State Achievement Examination(PSAE) in 11th grade. The PSAE is required for graduation in Chicago Public Schools.

[53]There was also an alternative assessment to the DC-CAS for students with severe cognitive disabilities who are unable to participate in the general assessment even with accommodations and/or modifications.

Table 1: Incentive Treatments by School District

	Dallas	NYC	DC	Chicago
Schools	43 schools opted in to participate, 22 schools randomly chosen for treatment	143 schools opted in to participate, 63 schools randomly chosen for treatment	17 schools randomly chosen to participate from the set of all DC middle schools	70 schools opted in to participate, 20 schools randomly chosen for treatment from a pre-determined set of 40
Students	4,008 2nd grade students: 23% black, 74% Hispanic, 58% free lunch eligible	17,744 4th and 7th grade students: 43% black, 42% Hispanic, 90% free lunch eligible	6,039 6th-8th grade students: 85% black, 9% Hispanic, 72% free lunch eligible	10,628 9th grade students: 55% black, 38% Hispanic, 93% free lunch eligible
Reward Structure	Students paid $2 per book to read books and pass a short test to ensure they read it. The average student earned $13.81 ($80 max).	4th graders could earn up to $25 per test and $250 per year. 7th graders could earn up to $50 per test and $500 per year. The average 4th grader earned $139.43 ($244 max). The average 7th grader earned $231.55 ($495 max).	Students could earn up to $100 every two weeks, $1500 per year. The average student earned $532.85 ($1322 max).	Students could earn up to $250 per report card and $2,000 per year. A=$50, B=$35, C=$20, D=$0, F=$0 (and resulted in $0 for all classes). Half of the rewards were given immediately, the other half at graduation. The average student earned $695.61 ($1875 max).
Frequency of Rewards	3 times per year	5 times per year	Every 2 weeks	Every 5 weeks / report card
Outcomes of Interest	ITBS and Logramos reading scores	New York state assessment ELA and math scores	DC-CAS reading and math scores	PLAN English and math scores
Operations	$360,000 total cost, 80% consent rate. One dedicated project manager.	$6,000,000 distributed. 66% opened bank accounts. 82% consent rate. 90% of students understood the basic structure of the incentive program. Three dedicated project managers.	$2,300,000 distributed. 99.9% consent rate. 86% of students understood the basic structure of the incentive program. Two dedicated project managers.	$3,000,000 distributed. 88.97% consent rate. 91% of students understood the basic structure of the incentive program. Two dedicated project managers.

NOTES: Each column represents a different city. Entries are descriptions of the schools, students, reward structure, frequency of rewards, outcomes of interest, and basic operations of our incentive treatments. See Appendix A for more details. The number of students given is the number in our ITT samples; that is, students who were in the treatment or control schools and grades at the beginning of the treatment school year (2007-08 for Dallas and 2008-09 for NYC, DC, and Chicago).

Table 2: The Effect of Financial Incentives on Student Achievement: Outputs

	NYC (Test Scores)				Chicago (Grades)	
	4th Grade		7th Grade		9th Grade	
	ITT	LATE	ITT	LATE	ITT	TOT
Reading: Raw	-0.023	-0.036	0.040	0.072	-0.027	-0.035
	(0.034)	(0.052)	(0.036)	(0.063)	(0.044)	(0.056)
N	6594	6594	10252	10252	7616	7616
Reading: All Controls	-0.021	-0.036	0.018	0.033	-0.006	-0.008
	(0.033)	(0.051)	(0.018)	(0.032)	(0.027)	(0.035)
N	6594	6594	10252	10252	7616	7616
Math: Raw	0.052	0.081	0.008	0.015	-0.030	-0.039
	(0.046)	(0.072)	(0.048)	(0.084)	(0.031)	(0.039)
N	6617	6617	10338	10338	7599	7599
Math: All Controls	0.067	0.092	-0.018	-0.030	-0.010	-0.013
	(0.046)	(0.070)	(0.035)	(0.063)	(0.023)	(0.029)
N	6617	6617	10338	10338	7599	7599

NOTES: The dependent variable is the state assessment taken in each respective city. There were no incentives provided for this test. All tests have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district with valid test scores. Thus, coefficients are in standard deviation units. The ITT is the difference between mean achievement of students in schools randomly chosen to participate and mean achievement of students in schools that were not chosen. TOT and LATE estimates are obtained by instrumenting for the number of days in a treatment school with original treatment and control assignment. See Section 3 in the text for formal definitions of ITT, TOT, and LATE. All standard errors, located in parentheses, are clustered at the school level.

Table 3: The Effect of Financial Incentives on Student Achievement: Inputs

	Dallas (Books)				DC (Att./Behavior)	
	2nd Grade		2nd Grade Spanish		6th - 8th Grade	
	ITT	TOT	ITT	TOT	ITT	TOT
Rdg. Comp.: Raw	0.182	0.253	-0.199	-0.239	-0.041	-0.053
	(0.071)	(0.097)	(0.096)	(0.116)	(0.223)	(0.282)
N	1900	1900	1756	1756	5844	5844
Rdg. Comp.: All Controls	0.180	0.249	-0.165	-0.200	0.152	0.179
	(0.075)	(0.103)	(0.090)	(0.108)	(0.092)	(0.106)
N	1900	1900	1756	1756	5844	5844
Rdg. Vocab.: Raw	0.045	0.062	-0.256	-0.307	–	–
	(0.068)	(0.093)	(0.101)	(0.122)		
N	1954	1954	1759	1759		
Rdg. Vocab.: All Controls	0.051	0.071	-0.232	-0.281	–	–
	(0.068)	(0.093)	(0.099)	(0.119)		
N	1954	1954	1759	1759		
Lang. Total: Raw	0.150	0.207	-0.054	-0.064	–	–
	(0.079)	(0.105)	(0.114)	(0.135)		
N	1944	1944	1742	1742		
Lang. Total: All Controls	0.136	0.186	-0.061	-0.073	–	–
	(0.080)	(0.107)	(0.125)	(0.149)		
N	1944	1944	1742	1742		
Math: Raw					-0.055	-0.071
					(0.217)	(0.274)
N	–	–	–	–	5846	5846
Math: All Controls					0.114	0.134
					(0.106)	(0.122)
N	–	–	–	–	5846	5846

NOTES: The dependent variable is the state assessment taken in each respective city. There were no incentives provided for this test. In Dallas, there were two different types of exams: English-speaking students took the Iowa Tests of Basic Skills (ITBS), and Spanish-speaking students took Logramos tests. All tests have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district with valid test scores. Thus, coefficients are in standard deviation units. The ITT is the difference between mean achievement of students in schools randomly chosen to participate and mean achievement of students in schools that were not chosen. TOT and LATE estimates are obtained by instrumenting for the number of days in a treatment school with original treatment and control assignment. See Section 3 in the text for formal definitions of ITT, TOT, and LATE. All standard errors, located in parentheses, are clustered at the school level.

Table 4A: The Effect of Financial Incentives on Student Achievement: Gender and Race

City	Grade Level	Subject	Full Sample	Male	Female	White	Black	Hispanic	Asian
Dallas (Books)	2nd	Reading Comp.	0.249	0.319	0.178	–	0.169	0.314	–
			(0.103)	(0.110)	(0.106)		(0.123)	(0.122)	
		N	1900	995	905		789	1023	
		Reading Vocab.	0.071	0.148	-0.012	–	0.107	0.035	–
			(0.093)	(0.106)	(0.095)		(0.129)	(0.093)	
		N	1954	1030	924		818	1045	
		Language	0.186	0.241	0.127	–	0.179	0.197	–
			(0.107)	(0.101)	(0.144)		(0.102)	(0.135)	
		N	1944	1020	924		809	1045	
	2nd Spanish	Reading Comp.	-0.200	-0.190	-0.198	–	–	–	–
			(0.108)	(0.139)	(0.090)				
		N	1756	888	868				
		Reading Vocab.	-0.281	-0.274	-0.280	–	–	–	–
			(0.119)	(0.143)	(0.106)				
		N	1759	890	869				
		Language	-0.073	-0.035	-0.090	–	–	–	–
			(0.149)	(0.191)	(0.116)				
		N	1742	878	864				
DC (Att./Behavior)	6th - 8th	Reading	0.179	0.267	0.091	-0.244	0.160	0.302	-0.077
			(0.106)	(0.132)	(0.081)	(0.171)	(0.109)	(0.116)	(0.237)
		N	5844	2903	2941	233	4956	555	98
		Math	0.134	0.188	0.076	-1.036	0.116	0.168	0.330
			(0.122)	(0.136)	(0.114)	(0.096)	(0.125)	(0.132)	(0.266)
		N	5846	2905	2941	233	4948	561	102

Table 4A (Continued)

City	Grade Level	Subject	Full Sample	Male	Female	White	Black	Hispanic	Asian
Chicago (Grades)	9th	English	-0.008	0.018	-0.034	-0.153	0.013	-0.011	0.218
			(0.035)	(0.038)	(0.040)	(0.070)	(0.041)	(0.045)	(0.172)
		N	7616	3629	3987	361	4171	2943	136
		Math	-0.013	-0.028	-0.002	-0.094	0.008	-0.009	-0.176
			(0.029)	(0.035)	(0.034)	(0.123)	(0.033)	(0.042)	(0.152)
		N	7599	3629	3970	361	4155	2942	136
NYC (Test Scores)	4th	ELA	-0.036	-0.013	-0.064	-0.188	-0.009	-0.020	-0.122
			(0.051)	(0.054)	(0.060)	(0.158)	(0.066)	(0.058)	(0.077)
		N	6594	3365	3222	233	2940	2857	507
		Math	0.092	0.100	0.076	-0.425	0.177	0.016	0.135
			(0.070)	(0.074)	(0.074)	(0.220)	(0.091)	(0.075)	(0.084)
		N	6617	3388	3225	237	2939	2871	516
	7th	ELA	0.033	0.037	0.027	0.225	-0.008	0.036	0.138
			(0.032)	(0.040)	(0.035)	(0.208)	(0.047)	(0.035)	(0.241)
		N	10252	5178	5061	709	4253	4247	992
		Math	-0.030	-0.011	-0.046	0.154	-0.079	-0.074	0.590
			(0.063)	(0.070)	(0.064)	(0.196)	(0.061)	(0.067)	(0.309)
		N	10338	5226	5105	709	4241	4317	1026

NOTES: The dependent variable is the state assessment taken in each respective city. There were no incentives provided for this test. All tests have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district with valid test scores. Thus, coefficients are in standard deviation units. All entries are TOT (Dallas, DC, and Chicago) or LATE (NYC) estimates with our set of controls. See Section 3 in the text for formal definitions of ITT, TOT, and LATE. An estimate for a racial group was only included if there were more than 100 individuals in that racial group in the experimental group. All standard errors, located in parentheses, are clustered at the school level.

Table 4B: The Effect of Financial Incentives on Student Achievement: Previous Year Test Scores, Free Lunch, and Behavior

City	Grade Level	Subject	Full Sample	Lowest Tercile	Middle Tercile	Highest Tercile	Missing Scores	Free Lunch	No Free Lunch	No Beh. Incidents	≥ 1 Beh. Incident
Dallas (Books)	2nd	Reading Comp.	0.249	0.257	0.173	0.383	0.111	0.231	0.296	–	–
			(0.103)	(0.140)	(0.131)	(0.102)	(0.172)	(0.093)	(0.121)		
		N	1900	539	513	499	349	835	1062		
		Reading Vocab.	0.071	0.151	0.089	0.109	-0.159	0.064	0.099	–	–
			(0.093)	(0.093)	(0.107)	(0.130)	(0.173)	(0.094)	(0.109)		
		N	1954	567	562	486	339	863	1088		
		Language	0.186	0.205	0.324	0.129	0.106	0.110	0.260	–	–
			(0.107)	(0.119)	(0.137)	(0.136)	(0.203)	(0.116)	(0.119)		
		N	1944	642	433	517	352	860	1081		
	2nd Spanish	Reading Comp.	-0.200	-0.370	-0.046	-0.050	-0.464	-0.219	-0.149	–	–
			(0.108)	(0.156)	(0.111)	(0.078)	(0.408)	(0.111)	(0.136)		
		N	1756	596	517	456	187	1276	479		
		Reading Vocab.	-0.281	-0.559	-0.089	-0.129	-0.490	-0.262	-0.346	–	–
			(0.119)	(0.137)	(0.130)	(0.081)	(0.364)	(0.124)	(0.124)		
		N	1759	589	499	511	160	1280	478		
		Language	-0.073	-0.294	0.047	-0.004	0.079	-0.067	-0.114	–	–
			(0.149)	(0.199)	(0.161)	(0.106)	(0.355)	(0.152)	(0.156)		
		N	1742	541	518	517	166	1271	470		
DC (Att./Behavior)	6th - 8th	Reading	0.179	0.071	0.077	0.073	0.169	0.156	0.214	0.175	0.400
			(0.106)	(0.106)	(0.094)	(0.073)	(0.115)	(0.097)	(0.133)	(0.097)	(0.235)
		N	5844	1517	1462	1374	1491	4163	1636	5129	216
		Math	0.134	0.046	0.036	0.020	0.128	0.124	0.150	0.139	0.164
			(0.122)	(0.149)	(0.105)	(0.106)	(0.129)	(0.125)	(0.107)	(0.115)	(0.274)
		N	5846	1442	1512	1392	1500	4161	1641	5123	214

Table 4B (Continued)

City	Grade Level	Subject	Full Sample	Lowest Tercile	Middle Tercile	Highest Tercile	Missing Scores	Free Lunch	No Free Lunch	No Beh. Incidents	≥ 1 Beh. Incident
Chicago (Grades)	9th	English	-0.008	0.015	-0.008	-0.016	0.052	-0.006	0.025	–	–
			(0.035)	(0.037)	(0.038)	(0.048)	(0.118)	(0.036)	(0.077)		
		N	7616	2066	2453	2398	699	6975	569		
		Math	-0.013	0.007	-0.011	0.006	-0.140	-0.009	-0.075	–	–
			(0.029)	(0.044)	(0.033)	(0.046)	(0.099)	(0.030)	(0.052)		
		N	7599	2098	2319	2521	661	6961	567		
NYC (Test Scores)	4th	ELA	-0.036	0.022	0.012	-0.112	-0.030	-0.006	-0.141	-0.041	0.060
			(0.051)	(0.088)	(0.052)	(0.063)	(0.147)	(0.057)	(0.112)	(0.050)	(0.158)
		N	6594	2263	2002	1959	370	4688	506	6217	191
		Math	0.092	0.147	0.171	0.048	-0.164	0.056	0.179	0.099	0.057
			(0.070)	(0.089)	(0.072)	(0.090)	(0.182)	(0.072)	(0.143)	(0.070)	(0.141)
		N	6617	2133	2160	2016	308	4725	505	6197	188
	7th	ELA	0.033	-0.047	-0.020	0.138	0.204	0.013	0.229	0.043	-0.078
			(0.032)	(0.041)	(0.036)	(0.065)	(0.118)	(0.036)	(0.091)	(0.033)	(0.090)
		N	10252	3485	2907	3257	603	7329	879	9562	430
		Math	-0.030	-0.067	-0.074	0.047	-0.131	-0.052	0.116	-0.031	0.153
			(0.063)	(0.078)	(0.058)	(0.100)	(0.139)	(0.061)	(0.086)	(0.064)	(0.162)
		N	10338	3178	3397	3107	656	7410	884	9576	423

NOTES: The dependent variable is the state assessment taken in each respective city. There were no incentives provided for this test. All tests have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district with valid test scores. Thus, coefficients are in standard deviation units. All entries are TOT (Dallas, DC, and Chicago) or LATE (NYC) estimates with our set of controls. See Section 3 in the text for formal definitions of ITT, TOT, and LATE. Behavioral incident data were not available for Dallas and Chicago. All standard errors, located in parentheses, are clustered at the school level.

Table 5: Alternative Outcomes

A. Dallas (Books)

Grade Level	Attendance Rates	Report Card Grades	Math Test Scores
2nd	-0.055	0.311	0.081
	(0.106)	(0.142)	(0.133)
N	1957	1935	1953
2nd Spanish	-0.025	-0.058	0.062
	(0.043)	(0.154)	(0.139)
N	1765	1760	1759

B. DC (Attendance/Behavior)

Grade Level	Attendance Rates	Report Card Grades	Behavioral Incidents
6th - 8th	0.171	0.049	-0.323
	(0.235)	(0.148)	(0.245)
N	6039	5802	6039

C. Chicago (Grades)

Grade Level	Attendance Rates	Report Card Grades	Total Credits Earned
9th	0.214	0.131	2.697
	(0.113)	(0.078)	(1.556)
N	10628	10613	10221

D. NYC (Test Scores)

Grade Level	Attendance Rates	Report Card Grades	Behavioral Incidents	Predictive ELA 1	Predictive ELA 2	Predictive Math 1	Predictive Math 2
4th	0.042	-0.050	-0.062	-0.115	-0.106	-0.076	-0.091
	(0.059)	(0.153)	(0.079)	(0.059)	(0.065)	(0.060)	(0.077)
N	6898	2162	6898	6032	6000	5791	5878
7th	-0.146	-0.082	0.290	-0.047	-0.078	0.017	-0.177
	(0.075)	(0.133)	(0.188)	(0.056)	(0.077)	(0.059)	(0.069)
N	10846	8252	10846	7990	8144	8315	8090

NOTES: Dependent variables vary from city to city and column to column. All dependent variables, except credits earned, have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district. Thus, all coefficients are in standard deviation units (except for the coefficient on credits earned). All entries are TOT (Dallas, DC, and Chicago) or LATE estimates with our set of controls. Predictive ELA 1 is taken in October. Predictive ELA 2 is taken in May of the same school year. Predictive math scores are similar. All standard errors, located in parentheses, are clustered at the school level.

Table 6: The Effect of Financial Incentives on Student Effort

City	Grade Level	Not Late for School	Ask Teacher for Help	Complete Hmwk.	Work Hard in School	Arrive on Time	Beh. Not Problem for Teachers	Satisfied with Achvmnt.	Push Myself Hard in School	Time Spent on Hmwk.
Dallas (Books)	2nd	–	-0.293	–	0.216	–	–	–	–	
			(0.130)		(0.106)					
	N		884		888					
	2nd Spanish	–	-0.150	–	-0.075	–	–	–	–	
			(0.068)		(0.139)					
	N		1022		1026					
DC (Att./Behavior)	6th - 8th	-0.034	-0.002	0.318	0.011	0.066	0.144	-0.028	0.010	0.016
		(0.070)	(0.069)	(0.050)	(0.056)	(0.042)	(0.052)	(0.051)	(0.034)	(0.049)
	N	3444	3454	3441	3361	3350	3331	3337	3338	3265
NYC (Test Scores)	7th	0.063	0.097	-0.106	-0.044	0.097	-0.056	-0.010	0.044	-0.016
		(0.100)	(0.100)	(0.114)	(0.071)	(0.076)	(0.076)	(0.107)	(0.083)	(0.075)
	N	3347	3374	3350	3340	3307	3294	3291	3290	3286

NOTES: Dependent variables vary by column and are gleaned from surveys administered in every district as part of the experiment. All dependent variables have been normalized to have a mean of zero and a standard deviation of one in the experimental group. All entries are TOT (Dallas and DC) or LATE (NYC) estimates with our set of controls. See Appendix B for more details regarding survey questions. All standard errors, located in parentheses, are clustered at the school level.

Table 7: The Effect of Financial Incentives on Student Attitudes Toward Schoolwork

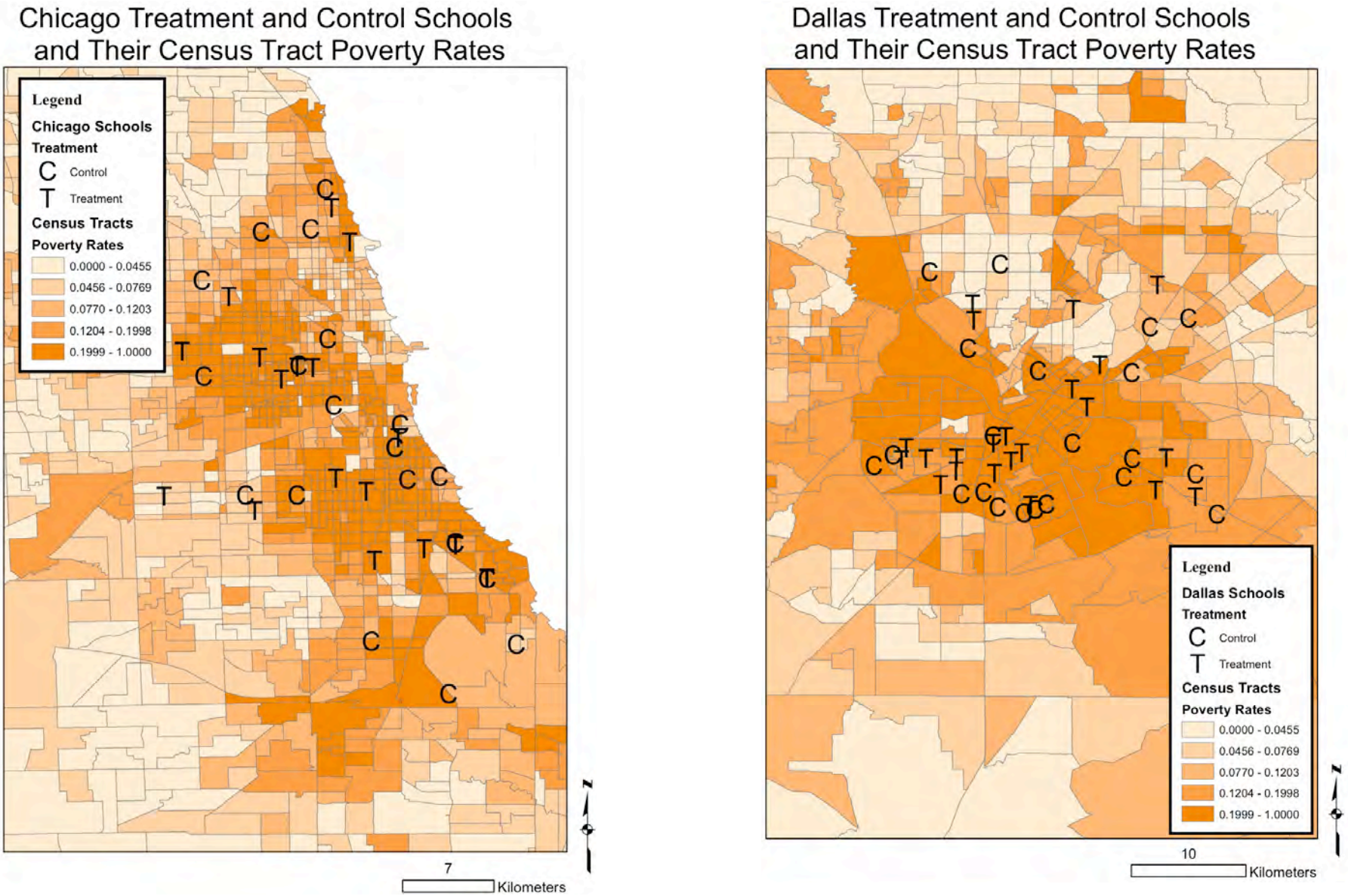
City	Grade Level	Mean/SD	Intrinsic Motivation Inventory	Enjoy Schlwk.	Schlwk. Is Fun	Schlwk. Is Not Boring	Schlwk. Holds Attention	Schlwk. Interesting	Schlwk. Enjoyable	Think About Schlwk. Enjoyment
Dallas (Books)	2nd	23.517	0.611	0.123	0.072	0.103	–	0.109	-0.021	0.077
		(6.266)	(0.816)	(0.147)	(0.144)	(0.122)		(0.171)	(0.181)	(0.151)
	N	797	797	887	851	872		874	876	867
	2nd Spanish	24.223	-0.902	-0.130	-0.098	-0.080	–	-0.152	-0.069	-0.232
		(5.760)	(0.643)	(0.127)	(0.168)	(0.118)		(0.113)	(0.129)	(0.103)
	N	936	936	1021	987	1016		1013	1020	1002
DC (Att./Behavior)	6th - 8th	27.314	0.732	0.196	0.205	-0.058	0.053	0.160	0.139	0.023
		(9.710)	(0.597)	(0.122)	(0.107)	(0.102)	(0.084)	(0.085)	(0.082)	(0.126)
	N	2766	2766	3094	3064	3048	3006	3027	3027	3071
NYC (Test Scores)	7th	25.520	-1.277	-0.229	-0.260	-0.445	0.002	-0.270	-0.317	-0.185
		(9.721)	(1.064)	(0.163)	(0.183)	(0.232)	(0.185)	(0.175)	(0.182)	(0.176)
	N	2829	2829	3161	3133	3092	3073	3121	3100	3162

NOTES: Dependent variables vary by column and are taken directly from the Intrinsic Motivation Inventory developed in Ryan (1982) and administered in every district as part of the experiment. All dependent variables for NYC and DC are measured on a seven-point Likert scale, whereas variables for Dallas are on a five-point scale. Schoolwork is abbreviated as in the column headings. All entries are TOT (Dallas and DC) or LATE (NYC) estimates with our set of controls. See Appendix B for more details regarding survey questions. All standard errors, located in parentheses, are clustered at the school level.

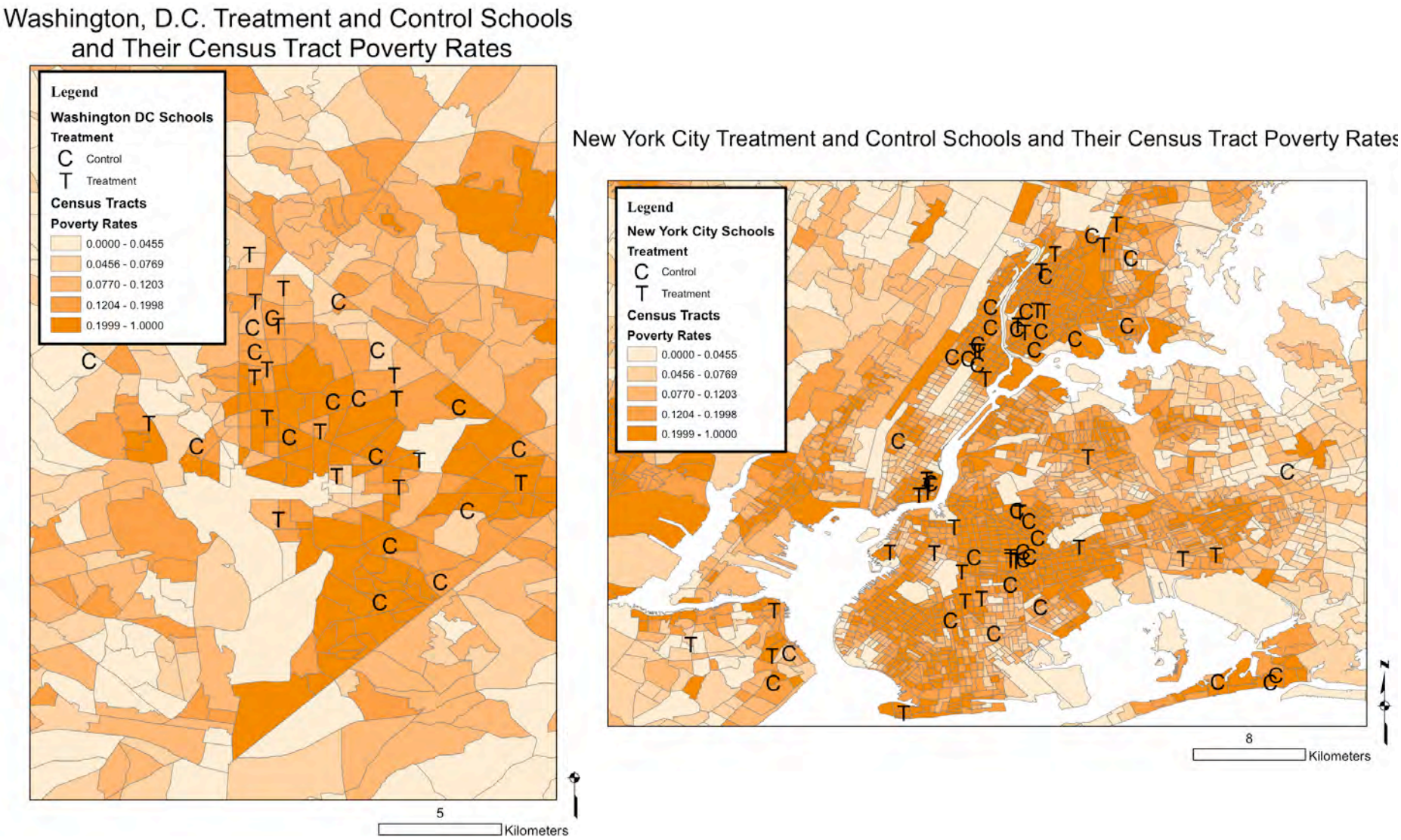
Table 8: The Impact of Incentives on Achievement:
by Teacher Value Added

Grade Level	Subject	With Valid Teacher Data	Below Median Teacher Value Added	Above Median Teacher Value Added
4th	ELA	-0.016	0.017	-0.066
		(0.068)	(0.079)	(0.126)
	N	3399	1712	1687
	Math	0.163	0.103	0.223
		(0.099)	(0.130)	(0.111)
	N	3277	1646	1631
7th	ELA	-0.043	-0.031	-0.086
		(0.056)	(0.058)	(0.082)
	N	3765	1970	1795
	Math	-0.097	-0.051	-0.276
		(0.113)	(0.069)	(0.265)
	N	4826	2432	2394

NOTES: The dependent variable is the state assessment taken in New York. There were no incentives provided for this test. All tests have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district. Thus, coefficients are in standard deviation units. All entries are LATE estimates with our set of controls. Teacher Value was calculated for New York by the Battelle Institute (http://www.battelleforkids.org/) for all teachers in math and ELA in grades four through eight. All standard errors, located in parentheses, are clustered at the school level.



Appendix Figure 1: Geographic Distribution of Treatment and Control Schools, Dallas and Chicago



Appendix Figure 2: Geographic Distribution of Treatment and Control Schools, Washington, DC and New York City

Appendix Table 1A: Dallas Summary Statistics, English

	Experimental Group			Treatment			Control		
	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N
White	0.029	0.168	1963	0.016	0.124	955	0.042	0.200	1008
Black	0.420	0.494	1963	0.409	0.492	955	0.430	0.495	1008
Hispanic	0.534	0.499	1963	0.568	0.496	955	0.502	0.500	1008
Asian	0.014	0.119	1963	0.005	0.072	955	0.023	0.149	1008
Other race	0.003	0.055	1963	0.002	0.046	955	0.004	0.063	1008
Male	0.527	0.499	1963	0.510	0.500	955	0.543	0.498	1008
Free lunch	0.441	0.497	1960	0.454	0.498	954	0.429	0.495	1006
Special education	0.063	0.243	1960	0.066	0.248	954	0.060	0.237	1006
English Language Learner (ELL)	0.129	0.335	1960	0.130	0.336	954	0.128	0.335	1006
Percent black	0.297	0.283	1963	0.279	0.251	955	0.315	0.309	1008
Percent Hispanic	0.678	0.276	1963	0.704	0.248	955	0.653	0.299	1008
Percent free lunch	0.563	0.114	1963	0.580	0.095	955	0.546	0.128	1008
Std. ITBS reading comprehension 2007-08	-0.058	0.971	1900	-0.042	0.954	917	-0.073	0.988	983
Std. ITBS reading vocabulary 2007-08	-0.151	0.848	1954	-0.189	0.812	951	-0.115	0.879	1003
Std. ITBS language total 2007-08	-0.017	0.987	1944	-0.029	0.987	949	-0.006	0.988	995
Std. ITBS math total 2007-08	0.007	0.989	1953	-0.045	0.953	951	0.056	1.020	1002
ITBS reading comprehension 2006-07	1.554	0.988	1963	1.460	0.974	955	1.643	0.994	1008
ITBS reading vocabulary 2006-07	1.247	0.954	1963	1.148	0.927	955	1.340	0.971	1008
ITBS language total 2006-07	1.424	0.922	1963	1.301	0.872	955	1.540	0.953	1008
ITBS math total 2006-07	1.406	0.938	1963	1.263	0.880	955	1.541	0.971	1008
ITBS reading comprehension 2005-06	0.120	0.404	1963	0.124	0.413	955	0.117	0.396	1008
ITBS reading vocabulary 2005-06	0.530	0.721	1963	0.476	0.690	955	0.580	0.746	1008
ITBS language total 2005-06	0.687	0.610	1963	0.639	0.587	955	0.733	0.628	1008
ITBS math total 2005-06	0.686	0.593	1963	0.650	0.578	955	0.719	0.605	1008
Std. attendance rate 2007-08	-0.179	1.110	1957	-0.214	1.215	953	-0.146	0.999	1004
Std. GPA 2007-08	0.022	1.032	1935	0.050	1.000	943	-0.004	1.061	992
Missing free lunch status	0.002	0.039	1963	0.001	0.032	955	0.002	0.045	1008
Missing special education status	0.002	0.039	1963	0.001	0.032	955	0.002	0.045	1008
Missing ELL status	0.002	0.039	1963	0.001	0.032	955	0.002	0.045	1008
Missing reading comprehension 2006-07	0.192	0.394	1963	0.215	0.411	955	0.171	0.376	1008
Missing reading vocabulary 2006-07	0.173	0.378	1963	0.191	0.393	955	0.156	0.363	1008
Missing language total 2006-07	0.180	0.385	1963	0.192	0.394	955	0.170	0.376	1008
Missing math total 2006-07	0.175	0.380	1963	0.190	0.392	955	0.161	0.367	1008
Missing reading comprehension 2005-06	0.911	0.285	1963	0.909	0.288	955	0.913	0.282	1008
Missing reading vocabulary 2005-06	0.263	0.440	1963	0.286	0.452	955	0.241	0.428	1008
Missing language total 2005-06	0.268	0.443	1963	0.288	0.453	955	0.249	0.433	1008
Missing math total 2005-06	0.260	0.439	1963	0.279	0.449	955	0.243	0.429	1008

Appendix Table 1B: Dallas Summary Statistics, Spanish

	Experimental Group			Treatment			Control		
	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N
White	0.001	0.024	1767	0.001	0.035	825	0.000	0.000	942
Black	0.001	0.024	1767	0.001	0.035	825	0.000	0.000	942
Hispanic	0.999	0.034	1767	0.998	0.049	825	1.000	0.000	942
Asian	0.000	0.000	1767	0.000	0.000	825	0.000	0.000	942
Other race	0.000	0.000	1767	0.000	0.000	825	0.000	0.000	942
Male	0.506	0.500	1767	0.507	0.500	825	0.505	0.500	942
Free lunch	0.728	0.445	1766	0.738	0.440	825	0.719	0.450	941
Special education	0.024	0.152	1766	0.028	0.165	825	0.020	0.141	941
English Language Learner (ELL)	0.973	0.163	1766	0.967	0.178	825	0.978	0.148	941
Percent black	0.156	0.150	1767	0.174	0.174	825	0.141	0.123	942
Percent Hispanic	0.817	0.154	1767	0.808	0.175	825	0.824	0.132	942
Percent free lunch	0.618	0.066	1767	0.616	0.068	825	0.619	0.063	942
Std. Logramos reading comprehension 2007-08	-0.036	1.019	1756	-0.113	1.056	819	0.032	0.980	937
Std. Logramos reading vocabulary 2007-08	-0.029	1.000	1759	-0.140	1.019	822	0.069	0.974	937
Std. Logramos language total 2007-08	-0.017	1.015	1742	-0.016	1.013	822	-0.018	1.017	920
Std. Logramos math total 2007-08	-0.055	0.995	1759	-0.022	1.000	822	-0.084	0.990	937
Logramos reading comprehension 2006-07	3.840	2.450	1767	3.823	2.481	825	3.854	2.424	942
Logramos reading vocabulary 2006-07	3.511	2.345	1767	3.566	2.344	825	3.463	2.345	942
Logramos language total 2006-07	3.258	2.441	1767	3.399	2.484	825	3.134	2.397	942
Logramos math total 2006-07	0.787	0.841	1767	0.812	0.848	825	0.765	0.836	942
Logramos reading comprehension 2005-06	1.394	1.896	1767	1.342	1.853	825	1.440	1.933	942
Logramos reading vocabulary 2005-06	0.557	1.315	1767	0.622	1.438	825	0.500	1.195	942
Logramos language total 2005-06	2.008	2.214	1767	1.997	2.244	825	2.017	2.188	942
Logramos math total 2005-06	0.000	0.005	1767	0.000	0.000	825	0.000	0.007	942
Std. attendance rate 2007-08	0.274	0.654	1765	0.267	0.655	825	0.280	0.654	940
Std. GPA 2007-08	0.006	0.883	1760	0.006	0.873	824	0.006	0.893	936
Missing free lunch status	0.001	0.024	1767	0.000	0.000	825	0.001	0.033	942
Missing special education status	0.001	0.024	1767	0.000	0.000	825	0.001	0.033	942
Missing ELL status	0.001	0.024	1767	0.000	0.000	825	0.001	0.033	942
Missing reading comprehension 2006-07	0.110	0.313	1767	0.127	0.333	825	0.094	0.293	942
Missing reading vocabulary 2006-07	0.092	0.289	1767	0.096	0.294	825	0.089	0.285	942
Missing language total 2006-07	0.096	0.294	1767	0.101	0.301	825	0.091	0.288	942
Missing math total 2006-07	0.434	0.496	1767	0.428	0.495	825	0.438	0.496	942
Missing reading comprehension 2005-06	0.291	0.455	1767	0.305	0.461	825	0.279	0.449	942
Missing reading vocabulary 2005-06	0.701	0.458	1767	0.698	0.459	825	0.703	0.457	942
Missing language total 2005-06	0.263	0.440	1767	0.278	0.448	825	0.251	0.434	942
Missing math total 2005-06	0.999	0.024	1767	1.000	0.000	825	0.999	0.033	942

Appendix Table 1C: NYC Summary Statistics, 4th Grade

	Experimental Group			Treatment			Control		
	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N
White	0.036	0.186	6872	0.043	0.202	3440	0.029	0.167	3432
Black	0.445	0.497	6872	0.441	0.497	3440	0.450	0.498	3432
Hispanic	0.434	0.496	6872	0.439	0.496	3440	0.429	0.495	3432
Asian	0.078	0.268	6872	0.072	0.258	3440	0.084	0.277	3432
Other race	0.007	0.084	6872	0.006	0.076	3440	0.008	0.092	3432
Male	0.515	0.500	6874	0.511	0.500	3441	0.520	0.500	3433
Free lunch	0.904	0.295	5342	0.891	0.311	2805	0.917	0.276	2537
Special education	0.119	0.324	6676	0.126	0.332	3343	0.112	0.315	3333
English Language Learner (ELL)	0.166	0.372	6676	0.171	0.376	3343	0.161	0.368	3333
Individual-level behavior 2007-08	0.044	0.316	6633	0.037	0.316	3315	0.051	0.317	3318
School-level behavior 2007-08	31.135	39.183	6898	21.478	21.984	3455	40.825	49.029	3443
Percent black	0.430	0.282	6898	0.424	0.290	3455	0.436	0.274	3443
Percent Hispanic	0.438	0.259	6898	0.444	0.255	3455	0.432	0.263	3443
Percent free lunch	0.901	0.095	6898	0.889	0.128	3455	0.912	0.038	3443
Std. ELA 2008-09	-0.159	0.923	6594	-0.151	0.916	3296	-0.167	0.931	3298
Std. math 2008-09	-0.156	0.967	6617	-0.115	0.998	3315	-0.198	0.934	3302
ELA 2007-08	653.030	36.128	6366	654.140	36.792	3173	651.927	35.427	3193
Math 2007-08	677.909	33.278	6506	678.865	34.433	3243	676.958	32.065	3263
ELA 2006-07	600.707	33.671	328	602.699	28.894	156	598.901	37.475	172
Math 2006-07	632.976	31.332	338	632.248	34.348	157	633.608	28.541	181
Std. attendance rate 2008-09	-0.107	1.056	6898	-0.095	1.018	3455	-0.119	1.094	3443
Std. GPA 2008-09	-0.110	0.983	2162	-0.113	1.034	821	-0.108	0.952	1341
Std. individual behavior 2008-09	0.045	1.245	6898	0.009	1.267	3455	0.081	1.221	3443
Std. predictive ELA 1 2008-09	-0.165	0.955	6032	-0.182	0.979	2939	-0.149	0.930	3093
Std. predictive ELA 2 2008-09	-0.183	1.011	6000	-0.202	1.050	3042	-0.162	0.969	2958
Std. predictive math 1 2008-09	-0.231	0.953	5791	-0.241	0.964	2932	-0.221	0.942	2859
Std. predictive math 2 2008-09	-0.218	0.947	5878	-0.240	0.982	3016	-0.195	0.908	2862
Missing race	0.004	0.061	6898	0.004	0.066	3455	0.003	0.056	3443
Missing sex	0.003	0.059	6898	0.004	0.064	3455	0.003	0.054	3443
Missing free lunch status	0.226	0.418	6898	0.188	0.391	3455	0.263	0.440	3443
Missing special education status	0.032	0.176	6898	0.032	0.177	3455	0.032	0.176	3443
Missing ELL status	0.032	0.176	6898	0.032	0.177	3455	0.032	0.176	3443
Missing individual-level behavior 2007-08	0.038	0.192	6898	0.041	0.197	3455	0.036	0.187	3443
Missing school-level behavior 2007-08	0.000	0.000	6898	0.000	0.000	3455	0.000	0.000	3443
Missing ELA 2007-08	0.077	0.267	6898	0.082	0.274	3455	0.073	0.260	3443
Missing math 2007-08	0.057	0.232	6898	0.061	0.240	3455	0.052	0.223	3443
Missing ELA 2006-07	0.952	0.213	6898	0.955	0.208	3455	0.950	0.218	3443
Missing math 2006-07	0.951	0.216	6898	0.955	0.208	3455	0.947	0.223	3443

Appendix Table 1D: NYC Summary Statistics, 7th Grade

	Experimental Group			Treatment			Control		
	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N
White	0.068	0.251	10800	0.077	0.266	4880	0.060	0.238	5920
Black	0.414	0.493	10800	0.367	0.482	4880	0.453	0.498	5920
Hispanic	0.417	0.493	10800	0.431	0.495	4880	0.405	0.491	5920
Asian	0.098	0.297	10800	0.123	0.328	4880	0.077	0.266	5920
Other race	0.004	0.060	10800	0.002	0.047	4880	0.005	0.069	5920
Male	0.508	0.500	10801	0.516	0.500	4880	0.502	0.500	5921
Free lunch	0.895	0.306	8543	0.871	0.335	4040	0.917	0.276	4503
Special education	0.122	0.327	10537	0.132	0.338	4768	0.114	0.317	5769
English Language Learner (ELL)	0.145	0.352	10537	0.144	0.351	4768	0.145	0.352	5769
Individual-level behavior 2007-08	0.063	0.347	10448	0.087	0.414	4743	0.044	0.277	5705
School-level behavior 2007-08	53.690	52.878	10846	61.908	57.688	4900	46.917	47.505	5946
Percent black	0.409	0.294	10846	0.366	0.292	4900	0.445	0.290	5946
Percent Hispanic	0.419	0.285	10846	0.429	0.284	4900	0.411	0.285	5946
Percent free lunch	0.897	0.104	10846	0.880	0.140	4900	0.911	0.057	5946
Std. ELA 2008-09	-0.151	0.938	10252	-0.103	1.043	4657	-0.192	0.839	5595
Std. math 2008-09	-0.152	0.975	10338	-0.104	1.060	4685	-0.191	0.896	5653
ELA 2007-08	647.199	28.252	9941	648.031	29.521	4513	646.506	27.135	5428
Math 2007-08	660.091	40.741	10024	661.707	42.896	4616	658.712	38.757	5408
ELA 2006-07	648.918	37.566	9541	650.158	39.846	4312	647.895	35.547	5229
Math 2006-07	661.948	40.566	9750	663.784	42.888	4410	660.431	38.482	5340
Std. attendance rate 2008-09	-0.110	1.044	10846	-0.136	1.089	4900	-0.088	1.005	5946
Std. GPA 2008-09	-0.246	0.998	8252	-0.251	1.018	3617	-0.242	0.983	4635
Std. individual behavior 2008-09	0.084	1.419	10846	0.217	1.933	4900	-0.025	0.755	5946
Std. predictive ELA 1 2008-09	-0.152	1.036	7990	-0.145	1.079	3750	-0.159	0.996	4240
Std. predictive ELA 2 2008-09	-0.145	1.008	8144	-0.108	1.051	3717	-0.177	0.970	4427
Std. predictive math 1 2008-09	-0.192	1.002	8315	-0.131	1.072	3718	-0.243	0.939	4597
Std. predictive math 2 2008-09	-0.158	0.999	8090	-0.138	1.070	3654	-0.175	0.937	4436
Missing race	0.004	0.065	10846	0.004	0.064	4900	0.004	0.066	5946
Missing sex	0.004	0.064	10846	0.004	0.064	4900	0.004	0.065	5946
Missing free lunch status	0.212	0.409	10846	0.176	0.380	4900	0.243	0.429	5946
Missing special education status	0.028	0.166	10846	0.027	0.162	4900	0.030	0.170	5946
Missing ELL status	0.028	0.166	10846	0.027	0.162	4900	0.030	0.170	5946
Missing individual-level behavior 2007-08	0.037	0.188	10846	0.032	0.176	4900	0.041	0.197	5946
Missing school-level behavior 2007-08	0.000	0.000	10846	0.000	0.000	4900	0.000	0.000	5946
Missing ELA 2007-08	0.083	0.277	10846	0.079	0.270	4900	0.087	0.282	5946
Missing math 2007-08	0.076	0.265	10846	0.058	0.234	4900	0.090	0.287	5946
Missing ELA 2006-07	0.120	0.325	10846	0.120	0.325	4900	0.121	0.326	5946
Missing math 2006-07	0.101	0.301	10846	0.100	0.300	4900	0.102	0.303	5946

Appendix Table 1E: DC Summary Statistics

	Experimental Group			Treatment			Control		
	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N
White	0.039	0.193	6039	0.020	0.141	3495	0.064	0.245	2544
Black	0.849	0.358	6039	0.856	0.352	3495	0.840	0.367	2544
Hispanic	0.095	0.293	6039	0.110	0.312	3495	0.075	0.263	2544
Asian	0.017	0.129	6039	0.014	0.119	3495	0.021	0.143	2544
Other race	0.000	0.018	6039	0.000	0.017	3495	0.000	0.020	2544
Male	0.502	0.500	6039	0.496	0.500	3495	0.509	0.500	2544
Free lunch	0.719	0.450	5988	0.736	0.441	3476	0.695	0.460	2512
Special education	0.177	0.382	6035	0.181	0.385	3492	0.172	0.378	2543
English Language Learner (ELL)	0.059	0.236	6035	0.070	0.254	3492	0.045	0.208	2543
Individual-level behavior 2007-08	0.085	0.643	5514	0.088	0.590	3201	0.079	0.710	2313
School-level behavior 2007-08	16.231	14.684	6039	16.903	15.878	3495	15.309	12.808	2544
Percent black	0.836	0.218	6039	0.837	0.208	3495	0.835	0.232	2544
Percent Hispanic	0.106	0.179	6039	0.125	0.203	3495	0.080	0.135	2544
Percent free lunch	0.705	0.190	6039	0.723	0.177	3495	0.680	0.204	2544
Elementary school	0.218	0.413	6039	0.120	0.326	3495	0.351	0.478	2544
Grade 6	0.272	0.445	6039	0.272	0.445	3495	0.272	0.445	2544
Grade 7	0.354	0.478	6039	0.341	0.474	3495	0.372	0.483	2544
Grade 8	0.374	0.484	6039	0.387	0.487	3495	0.357	0.479	2544
Std. DC-CAS reading 2008-09	0.035	0.946	5844	0.014	0.906	3368	0.063	0.997	2476
Std. DC-CAS math 2008-09	0.040	0.953	5846	0.013	0.925	3372	0.077	0.989	2474
DC-CAS reading 2007-08	667.592	80.387	4425	668.185	79.957	2557	666.781	80.986	1868
DC-CAS math 2007-08	663.822	79.095	4425	664.419	78.827	2557	663.004	79.473	1868
DC-CAS reading 2006-07	568.728	83.087	4464	569.427	82.861	2571	567.779	83.405	1893
DC-CAS math 2006-07	565.400	80.152	4464	565.921	79.715	2571	564.693	80.756	1893
Std. attendance rate 2008-09	0.038	0.950	6039	0.090	0.955	3495	-0.034	0.939	2544
Std. GPA 2008-09	-0.083	0.985	5802	-0.125	0.997	3370	-0.025	0.964	2432
Std. individual-level behavior 2008-09	0.011	1.041	6039	-0.081	0.833	3495	0.137	1.261	2544
Missing race	0.000	0.000	6039	0.000	0.000	3495	0.000	0.000	2544
Missing sex	0.000	0.000	6039	0.000	0.000	3495	0.000	0.000	2544
Missing free lunch status	0.008	0.092	6039	0.005	0.074	3495	0.013	0.111	2544
Missing special education status	0.001	0.026	6039	0.001	0.029	3495	0.000	0.020	2544
Missing ELL status	0.001	0.026	6039	0.001	0.029	3495	0.000	0.020	2544
Missing individual-level behavior 2007-08	0.087	0.282	6039	0.084	0.278	3495	0.091	0.287	2544
Missing school-level behavior 2007-08	0.000	0.000	6039	0.000	0.000	3495	0.000	0.000	2544
Missing DC-CAS reading 2007-08	0.267	0.443	6039	0.268	0.443	3495	0.266	0.442	2544
Missing DC-CAS math 2007-08	0.267	0.443	6039	0.268	0.443	3495	0.266	0.442	2544
Missing DC-CAS reading 2006-07	0.261	0.439	6039	0.264	0.441	3495	0.256	0.436	2544
Missing DC-CAS math 2006-07	0.261	0.439	6039	0.264	0.441	3495	0.256	0.436	2544

Appendix Table 1F: Chicago Summary Statistics

	Experimental Group			Treatment			Control		
	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N
White	0.048	0.213	10628	0.051	0.219	4396	0.045	0.208	6232
Black	0.557	0.497	10628	0.581	0.493	4396	0.539	0.499	6232
Hispanic	0.379	0.485	10628	0.356	0.479	4396	0.395	0.489	6232
Asian	0.016	0.126	10628	0.011	0.103	4396	0.020	0.140	6232
Other race	0.001	0.031	10628	0.001	0.034	4396	0.001	0.028	6232
Male	0.506	0.500	10628	0.519	0.500	4396	0.497	0.500	6232
Free lunch	0.932	0.251	10484	0.937	0.243	4333	0.929	0.257	6151
English Language Learner (ELL)	0.009	0.094	10382	0.008	0.091	4304	0.009	0.096	6078
Percent black	0.558	0.381	10628	0.569	0.389	4396	0.550	0.376	6232
Percent Hispanic	0.367	0.340	10628	0.356	0.338	4396	0.376	0.342	6232
Percent free lunch	0.926	0.073	10628	0.934	0.055	4396	0.921	0.083	6232
Std. PLAN English 2009-10	-0.236	0.796	7616	-0.258	0.776	3266	-0.220	0.810	4350
Std. PLAN math 2009-10	-0.219	0.810	7599	-0.239	0.776	3260	-0.204	0.834	4339
ISAT reading 2007-08	238.769	19.150	9123	238.365	19.245	3726	239.048	19.081	5397
ISAT math 2007-08	255.739	21.842	9176	255.850	21.707	3746	255.663	21.937	5430
ISAT reading 2006-07	227.977	22.387	9041	227.764	22.670	3722	228.125	22.188	5319
ISAT math 2006-07	242.521	22.638	9022	242.647	22.511	3713	242.433	22.729	5309
Std. attendance rate 2008-09	0.036	0.853	10628	0.087	0.835	4396	-0.000	0.864	6232
Std. GPA 2008-09	-0.051	0.936	10613	-0.015	0.928	4387	-0.077	0.940	6226
Number of credits earned 2008-09	46.175	15.513	10221	46.999	15.223	4218	45.597	15.688	6003
Missing race	0.000	0.000	10628	0.000	0.000	4396	0.000	0.000	6232
Missing sex	0.000	0.000	10628	0.000	0.000	4396	0.000	0.000	6232
Missing free lunch status	0.014	0.116	10628	0.014	0.119	4396	0.013	0.113	6232
Missing ELL status	0.023	0.150	10628	0.021	0.143	4396	0.025	0.155	6232
Missing ISAT reading 2007-08	0.142	0.349	10628	0.152	0.359	4396	0.134	0.341	6232
Missing ISAT math 2007-08	0.137	0.343	10628	0.148	0.355	4396	0.129	0.335	6232
Missing ISAT reading 2006-07	0.149	0.356	10628	0.153	0.360	4396	0.147	0.354	6232
Missing ISAT math 2006-07	0.151	0.358	10628	0.155	0.362	4396	0.148	0.355	6232

Appendix Table 2: Randomization Check

	Dallas	DC	Chicago	NYC	
	2nd	6th - 8th	9th	4th	7th
Percent white	0.046		0.031	-0.007	
	(0.072)		(0.030)	(0.009)	
Percent black	0.076	0.066	0.025	0.004	0.007
	(0.069)	(0.025)	(0.024)	(0.008)	(0.007)
Percent Hispanic	0.073	0.085	0.019	0.005	0.008
	(0.066)	(0.045)	(0.023)	(0.007)	(0.008)
Percent Asian		0.020			0.016
		(0.079)			(0.008)
Percent other race	0.268	1.015	0.647	-0.097	-0.091
	(0.354)	(0.578)	(0.291)	(0.086)	(0.145)
Percent male	-0.054	-0.015		-0.010	-0.002
	(0.053)	(0.036)		(0.017)	(0.013)
Percent free lunch	0.008	-0.017	0.029	-0.010	-0.005
	(0.024)	(0.015)	(0.019)	(0.008)	(0.009)
Percent ELL	-0.006	-0.019	-0.044	0.006	-0.003
	(0.016)	(0.046)	(0.082)	(0.011)	(0.010)
Percent special education	0.030	0.002		0.030	0.007
	(0.049)	(0.027)		(0.014)	(0.015)
Prev. year reading	-0.102	-0.025	-0.115	0.021	0.007
	(0.694)	(0.050)	(0.053)	(0.010)	(0.017)
Prev. year math	0.173	0.019	0.100	-0.005	-0.004
	(1.046)	(0.052)	(0.041)	(0.010)	(0.012)
Percent took test prev. year					
Prev. year reading vocab.	0.435				
	(0.754)				
Prev. year language	-0.916				
	(1.053)				
K-8 school		-0.848			
		(0.323)			
Number of students		0.001			
		(0.001)			
Mean school behavior 2007-08		1.029			
		(1.371)			
Mean GPA 2007-08		0.604	0.050		
		(0.449)	(0.566)		
Prev. year mean credits attempted			-0.051		
			(0.210)		
Prev. year mean credits earned			0.127		
			(0.254)		
N	42	34	42	68	72
R^2	0.366	0.439	0.305	0.151	0.055

NOTES: The dependent variable is treatment status assigned by randomization. Columns contain data that was collected before, during, and after the experiment. Equations estimated are linear regressions identical to Equation (1). All independent variables are pre-treatment, taken from the administrative files in the first month of school. Robust standard errors are in parentheses.

Appendix Table 3: The Effect of Financial Incentives on Student Achievement: Gender and Race – Intent-to-Treat

City	Grade Level	Subject	Full Sample	Male	Female	White	Black	Hispanic	Asian
Dallas (Books)	2nd	Reading Comp.	0.180	0.219	0.134		0.117	0.236	
			(0.075)	(0.077)	(0.083)		(0.084)	(0.097)	
			1900	995	905		789	1023	
		Reading Vocab.	0.051	0.103	-0.009		0.074	0.027	
			(0.068)	(0.074)	(0.074)		(0.088)	(0.072)	
			1954	1030	924		818	1045	
		Language	0.136	0.168	0.096		0.125	0.148	
			(0.080)	(0.072)	(0.113)		(0.072)	(0.106)	
			1944	1020	924		809	1045	
	2nd Spanish	Reading Comp.	-0.165	-0.150	-0.171				
			(0.090)	(0.111)	(0.079)				
			1756	888	868				
		Reading Vocab.	-0.232	-0.217	-0.242				
			(0.099)	(0.114)	(0.093)				
			1759	890	869				
		Language	-0.061	-0.027	-0.078				
			(0.125)	(0.155)	(0.103)				
			1742	878	864				
DC (Att./Behavior)	6th - 8th	Reading	0.152	0.228	0.077	-0.175	0.136	0.267	-0.087
			(0.092)	(0.115)	(0.070)	(0.164)	(0.094)	(0.109)	(0.323)
			5844	2903	2941	233	4956	555	98
		Math	0.114	0.161	0.064	-0.744	0.099	0.149	0.378
			(0.106)	(0.120)	(0.098)	(0.145)	(0.109)	(0.122)	(0.364)
			5846	2905	2941	233	4948	561	102

Appendix Table 3 (Continued)

City	Grade Level	Subject	Full Sample	Male	Female	White	Black	Hispanic	Asian
Chicago (Grades)	9th	English	-0.006	0.014	-0.027	-0.116	0.010	-0.009	0.173
			(0.027)	(0.029)	(0.032)	(0.055)	(0.032)	(0.037)	(0.147)
			7616	3629	3987	361	4171	2943	136
		Math	-0.010	-0.021	-0.001	-0.071	0.006	-0.007	-0.137
			(0.023)	(0.027)	(0.027)	(0.098)	(0.025)	(0.035)	(0.145)
			7599	3629	3970	361	4155	2942	136
NYC (Test Scores)	4th	ELA	-0.021	-0.007	-0.040	-0.129	-0.003	-0.012	-0.060
			(0.033)	(0.035)	(0.038)	(0.120)	(0.046)	(0.034)	(0.061)
			6594	3365	3222	233	2940	2857	507
		Math	0.067	0.071	0.059	-0.259	0.125	0.011	0.206
			(0.046)	(0.048)	(0.049)	(0.161)	(0.063)	(0.044)	(0.069)
			6617	3388	3225	237	2939	2871	516
	7th	ELA	0.018	0.019	0.015	0.081	-0.005	0.021	0.034
			(0.018)	(0.021)	(0.021)	(0.057)	(0.028)	(0.020)	(0.071)
			10252	5178	5061	709	4253	4247	992
		Math	-0.018	-0.008	-0.028	0.037	-0.048	-0.044	0.153
			(0.035)	(0.037)	(0.037)	(0.067)	(0.035)	(0.041)	(0.076)
			10338	5226	5105	709	4241	4317	1026

NOTES: The dependent variable is the state assessment taken in each respective city. There were no incentives provided for this test. All tests have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district with valid test scores. Thus, coefficients are in standard deviation units. All entries are ITT estimates with our set of controls. See Section 3 in the text for a formal definition of ITT. An estimate for a racial group was only included if there were more than 100 individuals in that racial group in the experimental group. All standard errors, located in parentheses, are clustered at the school level.

Appendix Table 4: The Effect of Financial Incentives on Student Achievement: Previous Year Test Scores, Free Lunch, and Behavior – Intent-to-Treat

City	Grade Level	Subject	Full Sample	Lowest Tercile	Middle Tercile	Highest Tercile	Missing Scores	Free Lunch	No Free Lunch	No Beh. Incidents	$\geq$ 1 Beh. Incident
Dallas (Books)	2nd	Reading Comp.	0.180	0.171	0.129	0.303	0.073	0.164	0.217		
			(0.075)	(0.097)	(0.103)	(0.079)	(0.121)	(0.070)	(0.088)		
		N	1900	539	513	499	349	835	1062		
		Reading Vocab.	0.051	0.107	0.064	0.085	-0.106	0.046	0.073		
			(0.068)	(0.068)	(0.079)	(0.105)	(0.122)	(0.069)	(0.081)		
		N	1954	567	562	486	339	863	1088		
		Language	0.136	0.137	0.237	0.106	0.072	0.079	0.191		
			(0.080)	(0.086)	(0.107)	(0.118)	(0.147)	(0.087)	(0.091)		
		N	1944	642	433	517	352	860	1081		
	2nd Spanish	Reading Comp.	-0.165	-0.299	-0.039	-0.045	-0.319	-0.180	-0.123		
			(0.090)	(0.128)	(0.099)	(0.073)	(0.303)	(0.092)	(0.116)		
		N	1756	596	517	456	187	1276	479		
		Reading Vocab.	-0.232	-0.436	-0.077	-0.116	-0.341	-0.216	-0.285		
			(0.099)	(0.106)	(0.117)	(0.076)	(0.273)	(0.104)	(0.106)		
		N	1759	589	499	511	160	1280	478		
		Language	-0.061	-0.229	0.040	-0.004	0.055	-0.055	-0.093		
			(0.125)	(0.159)	(0.144)	(0.100)	(0.273)	(0.129)	(0.133)		
		N	1742	541	518	517	166	1271	470		
DC (Att./Behavior)	6th - 8th	Reading	0.152	0.059	0.066	0.063	0.144	0.131	0.204	0.150	0.316
			(0.092)	(0.090)	(0.082)	(0.066)	(0.103)	(0.083)	(0.129)	(0.084)	(0.192)
		N	5844	1517	1462	1374	1491	4163	1636	5129	216
		Math	0.114	0.038	0.030	0.018	0.110	0.104	0.143	0.119	0.129
			(0.106)	(0.126)	(0.091)	(0.096)	(0.114)	(0.107)	(0.107)	(0.100)	(0.230)
		N	5846	1442	1512	1392	1500	4161	1641	5123	214

Appendix Table 4 (Continued)

City	Grade Level	Subject	Full Sample	Lowest Tercile	Middle Tercile	Highest Tercile	Missing Scores	Free Lunch	No Free Lunch	No Beh. Incidents	$\geq$ 1 Beh. Incident
Chicago (Grades)	9th	English	-0.006	0.011	-0.007	-0.010	0.035	-0.005	0.020		
			(0.027)	(0.027)	(0.031)	(0.039)	(0.080)	(0.028)	(0.065)		
		N	7616	2066	2453	2398	699	6975	569		
		Math	-0.010	0.005	-0.010	0.006	-0.086	-0.007	-0.061		
			(0.023)	(0.033)	(0.027)	(0.039)	(0.066)	(0.024)	(0.044)		
		N	7599	2098	2319	2521	661	6961	567		
NYC (Test Scores)	4th	ELA	-0.021	0.016	0.008	-0.077	-0.014	-0.002	-0.111	-0.025	0.109
			(0.033)	(0.053)	(0.036)	(0.046)	(0.098)	(0.035)	(0.077)	(0.032)	(0.098)
		N	6594	2263	2002	1959	370	4688	506	6217	191
		Math	0.067	0.089	0.112	0.038	-0.078	0.037	0.124	0.072	0.042
			(0.046)	(0.054)	(0.046)	(0.064)	(0.124)	(0.045)	(0.102)	(0.046)	(0.089)
		N	6617	2133	2160	2016	308	4725	505	6197	188
	7th	ELA	0.018	-0.025	-0.013	0.075	0.110	0.007	0.135	0.023	-0.047
			(0.018)	(0.023)	(0.019)	(0.034)	(0.075)	(0.022)	(0.048)	(0.019)	(0.046)
		N	10252	3485	2907	3257	603	7329	879	9562	430
		Math	-0.018	-0.040	-0.043	0.023	-0.078	-0.033	0.076	-0.019	0.071
			(0.035)	(0.042)	(0.034)	(0.055)	(0.082)	(0.037)	(0.053)	(0.035)	(0.079)
		N	10338	3178	3397	3107	656	7410	884	9576	423

NOTES: The dependent variable is the state assessment taken in each respective city. There were no incentives provided for this test. All tests have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district with valid test scores. Thus, coefficients are in standard deviation units. All entries are ITT estimates with our set of controls. See Section 3 in the text for a formal definition of ITT. Behavioral incident data were not available for Dallas and Chicago. All standard errors, located in parentheses, are clustered at the school level.

Appendix Table 5: Alternative Outcomes – Intent-to-Treat

A. Dallas (Books)

Grade Level	Attendance Rates	Report Card Grades	Math Test Scores
2nd	-0.040	0.228	0.059
	(0.078)	(0.106)	(0.098)
	1957	1935	1953
2nd Spanish	-0.021	-0.048	0.051
	(0.036)	(0.130)	(0.118)
	1765	1760	1759

B. DC (Attendance/Behavior)

Grade Level	Attendance Rates	Report Card Grades	Behavioral Incidents
6th - 8th	0.146	0.042	-0.274
	(0.204)	(0.128)	(0.219)
	6039	5802	6039

C. Chicago (Grades)

Grade Level	Attendance Rates	Report Card Grades	Total Credits Earned
9th	0.152	0.093	1.968
	(0.083)	(0.057)	(1.153)
	10628	10613	10221

D. NYC (Test Scores)

Grade Level	Attendance Rates	Report Card Grades	Behavioral Incidents	Predictive ELA 1	Predictive ELA 2	Predictive Math 1	Predictive Math 2
4th	0.027	-0.038	-0.040	-0.073	-0.067	-0.044	-0.056
	(0.039)	(0.092)	(0.052)	(0.038)	(0.044)	(0.040)	(0.052)
	6898	2162	6898	6032	6000	5791	5878
7th	-0.079	-0.044	0.157	-0.025	-0.045	0.009	-0.107
	(0.041)	(0.070)	(0.099)	(0.028)	(0.045)	(0.034)	(0.043)
	10846	8252	10846	7990	8144	8315	8090

NOTES: Dependent variables vary from city to city and column to column. All dependent variables, except credits earned, have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district. Thus, all coefficients are in standard deviation units (except for the coefficient on credits earned). All entries are ITT estimates with our set of controls. Predictive ELA 1 is taken in October. Predictive ELA 2 is taken in May of the same school year. Predictive math scores are similar. All standard errors, located in parentheses, are clustered at the school level.

Appendix Table 6: The Effect of Financial Incentives on Effort – Intent-to-Treat

City	Grade Level	Not Late for School	Ask Teacher for Help	Complete Hmwk.	Work Hard in School	Arrive on Time	Beh. Not Problem for Teachers	Satisfied with Achvmnt.	Push Myself Hard in School	Time Spent on Hmwk.
Dallas (Books)	2nd		-0.234		0.174					
			(0.107)		(0.088)					
	N		884		888					
	2nd Spanish		-0.128		-0.064					
			(0.061)		(0.123)					
	N		1022		1026					
DC (Att./Behavior)	6th - 8th	-0.030	-0.002	0.282	0.010	0.059	0.128	-0.025	0.008	0.015
		(0.063)	(0.063)	(0.048)	(0.051)	(0.039)	(0.049)	(0.046)	(0.031)	(0.045)
	N	3444	3454	3441	3361	3350	3331	3337	3338	3265
NYC (Test Scores)	7th	0.031	0.050	-0.055	-0.024	0.051	-0.028	-0.005	0.021	-0.012
		(0.054)	(0.051)	(0.062)	(0.037)	(0.039)	(0.040)	(0.056)	(0.043)	(0.035)
	N	3347	3374	3350	3340	3307	3294	3291	3290	3286

NOTES: Dependent variables vary by column and are gleaned from surveys administered in every district as part of the experiment. All dependent variables have been normalized to have a mean zero and a standard deviation of one in the experimental group. All entries are ITT estimates with our set of controls. See Appendix B for more details regarding survey questions. All standard errors, located in parentheses, are clustered at the school level.

Appendix Table 7: The Effect of Financial Incentives on Attitudes Toward Schoolwork – Intent-to-Treat

City	Grade Level	Mean/SD	Intrinsic Motivation Inventory	Enjoy Schlwk.	Schlwk. Is Fun	Schlwk. Is Not Boring	Schlwk. Holds Attention	Schlwk. Interesting	Schlwk. Enjoyable	Think About Schlwk. Enjoyment
Dallas (Books)	2nd	23.517	0.488	0.099	0.058	0.082		0.087	-0.017	0.062
		(6.266)	(0.679)	(0.123)	(0.119)	(0.101)		(0.142)	(0.150)	(0.126)
	N	797	797	887	851	872		874	876	867
	2nd Spanish	24.223	-0.773	-0.111	-0.084	-0.069		-0.130	-0.059	-0.199
		(5.760)	(0.567)	(0.112)	(0.148)	(0.104)		(0.100)	(0.114)	(0.088)
	N	936	936	1021	987	1016		1013	1020	1002
DC (Att./Behavior)	6th - 8th	27.314	0.649	0.173	0.182	-0.052	0.047	0.142	0.123	0.020
		(9.710)	(0.543)	(0.111)	(0.098)	(0.092)	(0.076)	(0.077)	(0.074)	(0.114)
	N	2766	2766	3094	3064	3048	3006	3027	3027	3071
NYC (Test Scores)	7th	25.520	-0.675	-0.124	-0.139	-0.230	0.002	-0.144	-0.169	-0.100
		(9.721)	(0.493)	(0.077)	(0.081)	(0.100)	(0.097)	(0.080)	(0.080)	(0.086)
	N	2829	2829	3161	3133	3092	3073	3121	3100	3162

NOTES: Dependent variables vary by column and are taken directly from the Intrinsic Motivation Inventory developed in Ryan (1982) and administered in every district as part of the experiment. All dependent variables for NYC and DC are measured on a seven-point Likert scale, whereas variables for Dallas are on a five-point scale. All entries are ITT estimates with our set of controls. See Appendix B for more details regarding survey questions. All standard errors, located in parentheses, are clustered at the school level.

Appendix Table 8: The Impact of Incentives on Achievement:
by Teacher Value Added – Intent-to-Treat

Grade Level	Subject	With Valid Teacher Data	Below Median Teacher Value Added	Above Median Teacher Value Added
4th	ELA	-0.008	0.015	-0.035
		(0.042)	(0.055)	(0.067)
	N	3399	1712	1687
	Math	0.100	0.050	0.136
		(0.063)	(0.074)	(0.069)
	N	3277	1646	1631
7th	ELA	-0.015	-0.015	-0.048
		(0.025)	(0.030)	(0.034)
	N	3765	1970	1795
	Math	-0.043	-0.021	-0.091
		(0.052)	(0.045)	(0.068)
	N	4826	2432	2394

NOTES: The dependent variable is the state assessment taken in New York. There were no incentives provided for this test. All tests have been normalized to have a mean of zero and a standard deviation of one within each grade across the entire sample of students in the school district. Thus, coefficients are in standard deviation units. All entries are ITT estimates with our set of controls. Teacher Value was calculated for New York by the Battelle Institute (http://www.battelleforkids.org/) for all teachers in math and ELA in grades four through eight. All standard errors, located in parentheses, are clustered at the school level.