Financial Incentives and Student Achievement: Evidence from Randomized Trials

Roland G. Fryer, Jr.*

Harvard University, EdLabs, and NBER

July 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, Kevin Murphy, 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, Theodora Chang, Vilsa E. Curto, Nancy Cyr, Will Dobbie, Katherine Ellis, Corinne Espinoza, 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 as effective. Relative to achievement-increasing 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 (OECD Indicators, 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 performance 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 for 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 students lack the structural resources or knowledge to convert effort to measurable achievement

[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 the United States.

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

or if the production function has important complementarities out of their control (e.g., effective teachers, engaged parents, or peer dynamics), 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 (figures include treatment and control 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]

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

[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 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).

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 (e.g., risk aversion, noisy output, or if students only have a vague idea of how to produce output), 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 between .093 (.057) and .123 (.074), but had no effect on achievement test scores in math or reading.

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 of 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 .142 (.090) and .166 (.103) standard deviations, and increases mathematics achievement between .103 (.104) and .121 (.120) 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, are economically significant, but only marginally statistically 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 [see 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 may have 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, may have 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 [.282 (.242) standard deviations in reading and .114 (.267) 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 when used in isolation.

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

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

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 incentives.

There are three pieces of evidence that support this theory. First, evidence from our qualitative team found consistent narratives suggesting that the typical student was elated by the incentive program 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 of 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 (e.g., effective parents) 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

[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.

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.

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

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

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 experimental group was comprised of 4,008 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 $1,500 to lower the cost of implementation.

Students were paid $2 per book read for up to 20 books per semester. Upon finishing a book, each student took an Accelerated Reader (AR) computer-based comprehension quiz, which provided evidence as to whether the student read the book. The student earned a $2 reward for scoring eighty percent or better on the book quiz. 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 8,355 total students. A participating

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

school received $2,500 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 $2,500 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 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, student were required to turn in signed parental consent forms; 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 student savings accounts 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 consent forms 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 Assessment of Educational Progress (NAEP). 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 92.4 percent black or Hispanic; 70 percent of students are eligible for free or reduced

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

lunch.

Washington, DC, is a relatively small school district, containing only thirty-five schools with middle school grades at the time of randomization. Thirty-four schools signed up to participate in the experiment and we randomly selected seventeen of them to be treated. The remaining seventeen 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, 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, ten 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 students opened up accounts 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 a stipend for participation in the program based on the number of students in the 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.

[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.

[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.

Chicago is the third largest school district in the U.S. 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 that wanted to participate and then randomly selected twenty to treat within this smaller set. Once a school was selected, students were required to return signed parental consent forms to participate. The set of students eligible to be treated consisted of 4,396 ninth graders. Eighty-nine 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, mathematics, 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 and 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; the data include 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

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

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 grade, or second grade. Spanish-speaking students in Dallas take a different exam called Logramos.[20] In New York City, the mathematics and English Language Arts tests, developed by McGraw-Hill, are administered each winter to students in grades three through eight. In Washington, DC, the District of Columbia 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 imbalance between treatment and control. The most important controls are reading and math achievement test scores from the previous two years, which we include in all regressions along with their squares. Previous years' test scores are available for most students who were in the district in previous years (see Appendix Tables 1A through 1F for exact percentages of experimental group students with valid test scores from previous years).[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 otherwise.

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

[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.

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

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: percent of student body that is black, percent Hispanic, and percent free lunch eligible. 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 attended. We assign anyone who moved into the school district at a later date to the school 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 attended first, and we assign everyone else to the school 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 from these surveys 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 in treatment schools and seventy-one percent of surveys in control schools were returned. 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 in treatment schools and thirty-nine percent of surveys in control schools were returned.[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 in control schools.

Appendix Tables 1A through 1F provide descriptive statistics for Dallas (A and B), New York City fourth graders (C), New York City seventh graders (D), Washington, DC (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 not to 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

[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.

[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.

and control schools on observables and unobservables. The standard method to check whether a school-based randomization was successful is to estimate models 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 treatment schools. The number of treatment and control schools ranges 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 the number of groups in Washington, DC, is smaller than usual.

Recall that we randomized among all schools that previously expressed interest in participating. Suppose there are X schools that are interested in participating and we aim to have a treatment group of size Y. Then, there are X choose Y potential treatment-control designations. 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 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 the most important school-level variables and controls from our 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, although treatment schools have slightly higher mean GPAs. In New York, we had a larger number of treatment and control schools than in other cities 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 [.037 (.015)]

[24]There is an active debate on which randomization procedures have the best properties. Karlan and Valdivia (2006) prefer a method similar to that 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.

and previous year's reading score [.004 (.013)], as well as mean number of behavioral incidents [.148 (.050)]. In seventh grade, the sample is similarly balanced. Treatment schools in Chicago have higher previous year's math scores [.110 (.043)] and lower reading scores [-.120 (.058)] than control schools. In Washington, DC, the smallest of the four school districts, only thirty-four (out of thirty-five possible) schools with grades six through eight signed up to participate. As a result, our independent variables are not as balanced across treatment and control as in the other cities. Schools in the treatment group have higher proportions of blacks, Hispanics, and students who report to be "other race," and are significantly less likely to be kindergarten through eighth grade 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 schools (Offenberg, 2001).[26]

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

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 for assignment to treatment, 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:

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

The ITT is an average of the causal effects for students in schools that were randomly selected for treatment at the beginning of the year and students in schools that signed up for treatment but

[26]A joint significance test for each city yields an F-statistic of 3.57 in Dallas, 7.06 in NYC fourth grade, .89 in NYC seventh grade, 2.37 in Chicago, and 4.66 in Washington, DC. These statistics cannot be compared across cities because the data available at the time of randomization differed across sites.

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, that 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, for participating students, of being in a school that was assigned to treatment. 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:

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

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

[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).

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 students 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 almost all specifications. For fourth graders, the ITT estimate of the effect of being offered an incentive program that 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. 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. Note: despite being statistically insignificant, we cannot reject the possibility of a modest effect in fourth grade math.

[28] If we allow entry and exit into treatment schools (some students enter and exit treatment schools multiple times a year), the estimates are thirty to fifty percent higher.

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

The results are similar for seventh graders. The ITT for mathematics is .008 (.048) standard deviations without controls and -.018 (.035) standard deviations with controls. The ITT for reading is .040 (.036) standard deviations without controls 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 reading 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 ITT for mathematics is -.031 (.031) standard deviations without controls and -.011 (.023) standard deviations with controls. TOT estimates for mathematics are very similar. The ITT for reading is -.028 (.045) standard deviations without controls and -.006 (.028) standard deviations with controls. Our TOT estimates for reading are also small: -.035 (.055) standard deviations without controls and -.007 (.034) 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 .123 (.074) [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 for second grade students in Dallas (separated according to whether they took the 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 (all estimates are 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 is 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 be .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 providing incentives for reading predominantly English-language books has 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 three pieces of evidence that, 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 on the Logramos test are entirely driven by the lowest performing students. 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 with 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 confusing the lowest performing Spanish-speaking students who were being provided with 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 .042 (.198) standard deviation decrease in reading and a .053 (.190) standard deviation decrease in mathematics. Adding our set of controls changes these estimates to .142 (.090) and .103 (.104) 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 fraction minority in the school and whether or not the school is kindergarten through eighth grades or a traditional middle school.

Our estimate of the effect of participating in the Washington, DC, experiment is -.054 (.251) standard deviations for reading and -.068 (.240) standard deviations for math (both estimates without controls). Adding controls changes the estimates to .166 (.103) and .121 (.120), respectively. Hence, once one accounts for our set of controls, paying middle school students for various inputs to the educational production function, such as attendance and behavior, has a positive, though only marginally significant, effect on reading scores and a slightly smaller (not significant) effect on math scores. 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 standard deviations in reading (Decker, Mayer, and Glazerman, 2004). The effect of Head Start is 0.147 (0.103) standard deviations in applied problems and 0.319 (0.147) standard deviations 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). Conversely, the average charter school across fifteen states shows no statistically significant advantages over traditional public schools (Gleason et al., 2010).[30]

Incentive programs based on inputs have effect sizes consistent with those of other achievement-increasing reforms in recent decades, but cost considerably less. 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 $4,497 per student per year (in 2009 dollars). Early childhood investments are typically even more expensive.

Tables 4A and 4B investigate treatment effects for subsamples that we deemed important before any analysis was conducted – gender, race/ethnicity, previous year's test score, previous year's

[30]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.

behavior, and an income proxy.[31] In these tables, as we have done throughout, we report standard errors that have been clustered at the school level. Appendix Tables 9A through 9E provide our analysis of subsamples where we adjust the standard errors for multiple hypothesis testing using a simple Bonferroni correction and the Free Step-Down Resampling Method detailed in Westfall and Young (1993) and Anderson (2008). In what follows, we refer to the unadjusted p-values but make a point to alert the reader if more conservative p-values do not confirm our results.

Gender was divided into two categories and race/ethnicity was divided into five categories: non-Hispanic white, non-Hispanic black, Hispanic, non-Hispanic Asian and non-Hispanic other race.[32] We only include a racial/ethnic category in our analysis if there are at least one hundred students from that racial/ethnic category in our experimental group. This restriction eliminates whites and Asians in Dallas and other race in other cities. Previous year's test scores are partitioned into four groups – evenly distributed terciles for students with valid pre-treatment test scores and a missing category for students without valid pre-treatment scores. Because of frequent mobility in and out of urban school districts, there are many students whose previous year's test scores are missing from our data. The proportion of students with missing previous year's test scores is between seventeen and nineteen percent for English-speaking students in Dallas, between nine and eleven percent for Spanish-speaking students in Dallas, fourteen percent in Chicago, eleven percent in Washington, DC, and between six and eight percent for fourth and seventh graders in NYC. Eligibility for free lunch is used as an income proxy.[33] The final distinction is between those students with at least one behavioral incident in the previous year that led to a suspension and those with none.[34]

Table 4A presents TOT and LATE estimates for gender and race/ethnicity subsamples.[35] The first column provides estimates on the full sample, restricted to only students who contain valid

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

[32] The eighty-four students in Dallas and sixty-nine students in NYC with missing gender information were not included in the gender subsample estimates. The eighty-four students in Dallas and seventy-two students in NYC with missing race/ethnicity information are not included in the race/ethnicity subsample estimates.

[33] 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.

[34] These are relatively major infractions. Five randomly chosen descriptions of behavioral incidents 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."

[35] Appendix Table 3 provides ITT estimates.

information for the particular subsample. There are no gender differences in the two output experiments. In New York, seventh grade boys gain .037 (.040) standard deviations in reading and -.011 (.070) standard deviations in math. Effect sizes for seventh grade girls are similar: .027 (.035) standard deviations in reading and -.046 (.064) standard deviations in math. The same pattern holds for fourth graders. In Chicago, the treatment effect on reading scores was .017 (.037) standard deviations for boys and -.034 (.039) standard deviations for girls. Estimates for math are -.027 (.034) and -.002 (.033) standard deviations, respectively.

In sites where incentives were relatively effective, boys seem to gain more from the experiment than girls on reading test score outcomes. 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 .237 (.129) standard deviations for boys and .089 (.081) standard deviations for girls. These results still hold for reading comprehension in Dallas, using family-wise error correction to adjust p-values, but gender differences in DC are no longer statistically significant after this adjustment. This finding is surprising, given results from previous demonstration projects that seem to favor girls (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 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 in language. Reading vocabulary scores show a

bigger increase for blacks than Hispanics, but estimates of these effects 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, blacks, Hispanics, and Asians all benefited from the experiment more than whites in math. Asians also benefited more in reading than each of the other racial/ethnic groups, with gains of .849 (.320) standard deviations. All racial/ethnic groups in New York show either small or inconsistent gains, although black fourth graders who participated in the incentives program did make larger gains in math than white or Hispanic fourth graders who participated. Whites in Chicago who received incentives fared somewhat worse in English than other racial/ethnic groups, but otherwise there were no differences among racial/ethnic groups in Chicago. [36]ifferences between racial/ethnic groups are not significant after one uses family-wise error correction to adjust p-values.

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 Spanish-language test in Dallas, the negative results at the mean are all driven by students in the bottom tercile.[37] The treatment effects for the bottom third of the students taking the Spanish-language exam are negative and very large in magnitude: -.370 (.156) in reading comprehension, -.559 (.137) in vocabulary, and -.294 (.199) in language.[38] The treatment effect for Spanish-speaking students in the middle and top terciles is statistically zero and substantively small. Treatment effects for English-speaking students are roughly equal across terciles. Splitting the sample by whether or not a student had a behavioral incident in the previous year shows no difference in treatment effects 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 [.282 (.242) standard deviations in reading and .114 (.267) 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

[36]D

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

[38]Treatment effects for the lowest tercile remain marginally significant for reading comprehension and significant for reading vocabulary even after adjusting p-values using family-wise error correction.

are .132 (.092) and .097 (.110) standard deviations.

Another interesting comparison is between students eligible for free lunch and their counterparts who are not eligible for free lunch.[39] For the input experiments, the point estimates for students eligible for free lunch are lower than for those students not eligible for free lunch, but one cannot reject the null hypothesis that they are the same.[40] In our output experiments, students not eligible for free lunch show modest gains from the experiment on several of the assessments. Seventh graders in NYC not eligible for free lunch had significantly larger gains from the experiment on both the reading and math exams, while treatment effects for fourth graders in NYC and ninth graders in Chicago did not differ significantly by free lunch eligibility. The point estimates for seventh graders who are not eligible for free lunch are .229 (.091) standard deviations in reading and .115 (.086) standard deviations in math. The corresponding treatment effects for seventh grade students who qualify for free lunch are .013 (.036) standard deviations in reading and -.052 (.061) standard deviations in math. Gains for seventh graders not eligible for free lunch in NYC are still significant in reading but are not significant in math after using family-wise error correction to adjust p-values.

This pattern was not seen in Chicago, where the treatment effects for students not eligible for free lunch are .024 (.076) standard deviations in English and -.073 (.051) standard deviations in math. The impact of incentives on achievement for students eligible for free lunch is -.006 (.035) standard deviations in reading and -.009 (.030) standard deviations in math.

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 in reading, while incentives for output may benefit those who do not qualify for free lunch.

[39] 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.

[40] This does not include the Spanish-speaking 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.[41] Every student in the district takes predictive assessments in October and June that are correlated with the state exam. Students were also given incentives for these predictive exams (but not the state assessments) as part of the ten total tests for which they were rewarded, which provides another outcome. 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 a mean of zero and a standard deviation of one among students in the same grade across the school district. In addition, Dallas has math scores from the Iowa Tests of Basic Skills; Washington, DC, has behavioral incidents and results from a set of interim assessments called the District of Columbia Benchmark Assessment System (DC-BAS); Chicago has total credits earned, total credits attempted, and PLAN composite, reading, and science scores; and New York City collects data on behavioral incidents as well as scores on interim 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 to read books on attendance rates.

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

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 in the incentive program gained .311 (.142) standard deviations. Math test scores also increased slightly but the treatment effect is not statistically significant.[42] There is no evidence that incentives had any impact on these alternative outcomes for students who took the Spanish-language exam.

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 [.169 (.235)] and fewer behavioral incidents [-.321 (.243)], but because of a lack of statistical power we cannot make confident conclusions. Report card grades did not increase [.047 (.149)]. Treatment students also had modest positive results on low stakes DC-BAS exams, but these effects are not significant. Conversely, our experiment in Chicago demonstrates that paying students for better course grades has a modest impact on their grades – an increase of .123 (.074) standard deviations – along with a larger increase in attendance of .200 (.107) standard deviations and an increase of 2.546 (1.479) credits earned. These estimates are marginally significant. The typical course in Chicago is worth four credits. Treatment students, therefore, passed approximately one-half of a course more on average than control students. The incentives program had no effect on PLAN composite, reading, and science scores. 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 that paying students for higher test scores changed their behavior in any quantifiable way.

Effort

[42]Second grade math tests have a significant reading component.

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 that 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 her achievement; (8) whether she pushes herself hard at school; and (9) how many hours per week she spends on homework.[43] See Appendix B for further details.

Our results, shown in Table 6, indicate that there are few differences on the dimensions of effort described above between those students who received treatment and those who did not. 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 are self-paced activities that made students feel more independent from their teachers or they simply did not need to ask as may questions. Working harder on schoolwork is consistent with the increases shown in their grade point averages. In Washington, DC, students reported that they were .317 (.051) standard deviations more likely to complete their homework and had .146 (.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 increase enthusiasm for school so much that students engage differently with their teachers. Tables 5 and 6 provide evidence that such effects are most likely not present, as the impacts of incentive programs seem relatively localized. Many qualitative observations confirm general excitement by students

[43] 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).

about earning rewards, but they seem to have focused their behavioral changes on precisely those elements that were incentivized.

Intrinsic Motivation

One of the major criticisms of the use of incentives to boost student achievement is that the incentives may destroy a student's "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 administered the Intrinsic Motivation Inventory, developed by Ryan (1982), to students in our experimental groups.[44] The instrument assesses participants' interest/enjoyment, perceived competence, effort, value/usefulness, pressure and tension, and perceived choice while performing a given activity. There is a subscale score 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 instrument 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 for these statements (reversing the sign on statements (3) and (4)). Only students with valid responses to all statements are included in our analysis of the overall score, as non-response may be confused with low intrinsic motivation. 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 on each survey statement independently. Coefficients reported are TOT

[44]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)].

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 intrinsic 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 1.274 (1.055) on a mean of 25.527. 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. This suggests that 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 students when the incentives are discontinued. Many believe that students will experience decreased motivation and thus (net) negative impacts on achievement when the incentives are removed following the experiment [see 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 outperforming students who were in the control group. More precisely, our estimate of the effect of incentives on achievement one year after the experiment has ended is between .088 (.080) [ITT] and .120 (.107) [TOT] standard deviations in reading and between .150 (.108) [ITT] and .206 (.145) [TOT] standard deviations in math [these results are using scores on the Texas Assessment of Knowledge and Skills (TAKS) from 2008-09 and are not shown in tabular form]. In other words, the point estimate a year after the experiment is roughly half as large as the point estimate on reading comprehension from the experimental year, and the effect is not statistically significant. This is similar to the fade-out effect that has been documented from other achievement-increasing interventions such as Head Start, having a high-quality teacher for one year, or a lower class size (Nye, Hedges, and Konstantopoulos, 1999; Puma, 2010). For students who took the Spanish-language exam in Dallas, the effects of the program a year after being discontinued are small and statistically insignificant [estimates are .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 treatment year did not

persist or the negative impacts can be explained by crowd-out.

6 Interpretation

Our field 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 leads to a marginally significant increase in reading achievement and has a similar but statistically insignificant impact on math achievement. Paying seventh grade students for test scores does not boost their achievement. Paying high school freshmen for their grades in core courses leads to modest increases in 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. 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. An argument against the former explanation is the fact that the fourth graders in New York demonstrated insignificant 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 there were not as important for achievement as reading books. This is our leading interpretation.[45]

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 functions – that is, the precise relationship between the vector of inputs and the corresponding output.[46] 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 function lies in our qualitative data. During the 2008-2009 school year, seven full-time qualitative researchers in New York observed twelve students and their families, as well as ten classrooms. From detailed interview notes, we gather that students were uniformly excited about the incentives and the prospect of earning money for school performance.[47] Despite showing that students were excited about the incentive programs, the qualitative data also demonstrate that students had little idea about how to translate their enthusiasm into tangible steps designed to increase their achievement. After each of the ten exams administered in New York, our qualitative team asked

[45]One might also worry that the marginal value of an increase in achievement is not similar across treatments and that might explain the results. (Many thanks to Kevin Murphy for making this excellent point.) To investigate this, we estimated the amount of money a student would earn across treatments for a .25 standard deviation increase in achievement. In Dallas, a .25 increase in achievement was associated with earning $13.81. In NYC, a .25 standard deviation increase in achievement would have resulted in a $13.66 increase in earnings for fourth graders and a $33.71 increase in earnings for seventh graders. In Chicago, the corresponding marginal increase in earnings is $31.84.

[46]Technically, students are only assumed to have more knowledge of their production function than a social planner.

[47]In a particularly illuminating example, one of the treatment schools asked their students to propose 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.

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 discussed test-taking strategies rather than salient inputs into the education production function or improving their general understanding of a subject area.[48] 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 had entered them correctly." Not a single student mentioned: reading the textbook, studying harder, completing homework, or asking teachers or other adults for help with 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 (subject to rounding) 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.[49] This finding is consistent with the empirical data in Table 6.

Yet when probed about why other 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 – one female student remarked "I never thought about it."[50]

The basic messages from students in Chicago centered on the excitement generated by the program at the beginning of the year. This excitement triggered more effort initially – coming to school, paying attention in class, and so on – but students indicated that they did not notice any change in 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

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

[49]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.

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

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 are 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 of 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, before the experiment started, one could collect information on the discount rates of all students in treatment and control schools and then test for heterogeneous treatment effects between those 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 rates. A potential limitation of this approach is that it critically depends on the metric for deciphering high and low discount 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 place 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 in order 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 would have been 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.[51]

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 data. This subset comprises approximately 47 percent of the full sample. The results from this subset of students are similar to those for the full sample, save that in fourth grade math we observe a sizable treatment effect. The next two columns divide students according to whether their teachers are above or below the median value-added for teachers in New York City. 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 quality teachers gain .223 (.111) standard deviations. The exact opposite pattern is observed for

[51] 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.

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 between these two groups were only statistically significant for seventh graders in NYC. 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 for her performance at school. Her 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.[52]

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 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

[52]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.

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 perhaps 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 (e.g., reading books). 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.[53]

***

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; these were designed to test the impact of incentives on student achievement.

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 those of so-called successful reforms in the past few decades, and obtained at lower cost. Yet incentives are by no means a silver bullet. They pass a

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

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.[54]

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 the context of teacher incentive pay as well (Duflo and Hanna, 2006; Muralidharan and Sundararaman, 2009). A complete understanding will require more experimentation and constant refinement 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 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.

[54] KIPP pays students on a point system that is 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.

8 Appendix A: Implementation Manual (Not For Publication)

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 not to 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 treatment school principal 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, DC, 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 his/her student to participate - in year 1, nine students out of 3,269 opted out. The experiments continued through the 2009-2010 school year.

C. PERFORMANCE METRICS AND INCENTIVE STRUCTURE

Members of the Capital Gains team conducted a meeting at each of the treatment schools during the first two weeks of school to explain the program to the school's staff 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 the participating schools at the start of the program. Representatives from the bank visited the schools and signed students up for accounts during their lunch and free periods. Each student was 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 savings 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 savings accounts) 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: A certificate was sent to each participating student displaying the amount of money earned based on his/her performance on each of the school's metrics. Certificates provided both a description of the student's behavior (e.g., "You were late to class 6 times this pay period") and 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 gauge 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.

Check-Cashing Letters: "Check-cashing letters" were provided, which had 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. A similar survey was administered at the end of year 2.

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 student's performance (i.e., scores on each metric, consent status, bank account status, and reward history). Coordinators also received a list of the top ten earners in each grade for a given pay period as well as a list of the ten students with the largest increases in rewards relative to 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 online survey as part of an effort to further understand the effects of the program. The survey results contain valuable insights and feedback on program implementation and impact. Project Managers also visited each of the schools

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.

9.3 SURVEY QUESTIONS

New York City/Washington, D.C.

Effort Questions

1. During the school year, how often have you been late for school?

   (a) Never

   (b) Once a month or less

   (c) Once every two weeks (2-3 times a month)

   (d) Once a week (4-5 times a month)

   (e) Several times a week (2-4 times a week)

   (f) Every day (5 times a week)

2. I would ask the teacher for help, if I needed it.

   (a) Never

(b) Usually not

(c) Sometimes

(d) Usually

(e) Always

3. About how much of your assigned homework do you usually complete, either during school hours or outside of school?

(a) All

(b) Three quarters

(c) Half

(d) One quarter

(e) Almost none

4. I work very hard on my schoolwork.

(a) Not at all true

(b) Not very true

(c) Sort of true

(d) Very true

5. I don't really care whether I arrive on time for class

(a) Totally untrue

(b) Mostly untrue

(c) Somewhat true

(d) Mostly true

(e) Totally true

6. My behavior is a problem for the teachers in my classes

(a) Totally untrue

(b) Mostly untrue

(c) Somewhat true

(d) Mostly true

(e) Totally true

7. I am satisfied with what I have achieved in my classes

(a) Totally untrue

(b) Mostly untrue

(c) Somewhat true

(d) Mostly true

(e) Totally true

8. I have pushed myself hard to completely understand my lessons in school

(a) Totally untrue

(b) Mostly untrue

(c) Somewhat true

(d) Mostly true

(e) Totally true

9. Which of these is closest to the amount of time you usually spend on homework outside of school each week?

(a) 1-4 hours

(b) 5-9 hours

(c) 10-14 hours

(d) 15-19 hours

(e) 20 or more hours

Intrinsic Motivation Questions

For each of the following statements, please indicate how true it is for you, using the following scale: 1 (totally untrue) to 7 (totally true)

1. I enjoy doing schoolwork very much.
2. Doing schoolwork is fun.
3. I think doing schoolwork is boring.
4. Doing schoolwork does not hold my attention at all.
5. I would describe doing schoolwork as very interesting.
6. I think doing schoolwork is quite enjoyable.
7. When I am doing schoolwork, I think about how much I enjoy it.

Dallas

Effort Questions

1. Do you work very hard on your schoolwork?
   (a) Never
   (b) Some of the time
   (c) Half of the time
   (d) Most of the time
   (e) All of the time
2. Do you ask the teacher for help when you need it?
   (a) Never
   (b) Some of the time

(c) Half of the time

(d) Most of the time

(e) All of the time

Intrinsic Motivation Questions

For each of the following statements, please indicate how true it is for you, using the following scale: 1 (totally untrue) to 7 (totally true)

1. I enjoy reading very much.
2. I think reading is boring.
3. When I am reading, I think about how much I enjoy it.
4. Reading is very interesting.
5. Reading is fun.
6. Reading is a good way to spend time.

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	-0.028	-0.035
	(0.034)	(0.052)	(0.036)	(0.063)	(0.045)	(0.055)
N	6594	6594	10252	10252	7616	7616
Reading: All Controls	-0.021	-0.036	0.018	0.033	-0.006	-0.007
	(0.033)	(0.051)	(0.018)	(0.032)	(0.028)	(0.034)
N	6594	6594	10252	10252	7616	7616
Math: Raw	0.052	0.081	0.008	0.015	-0.031	-0.038
	(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	-0.011	-0.013
	(0.046)	(0.070)	(0.035)	(0.063)	(0.023)	(0.028)
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.

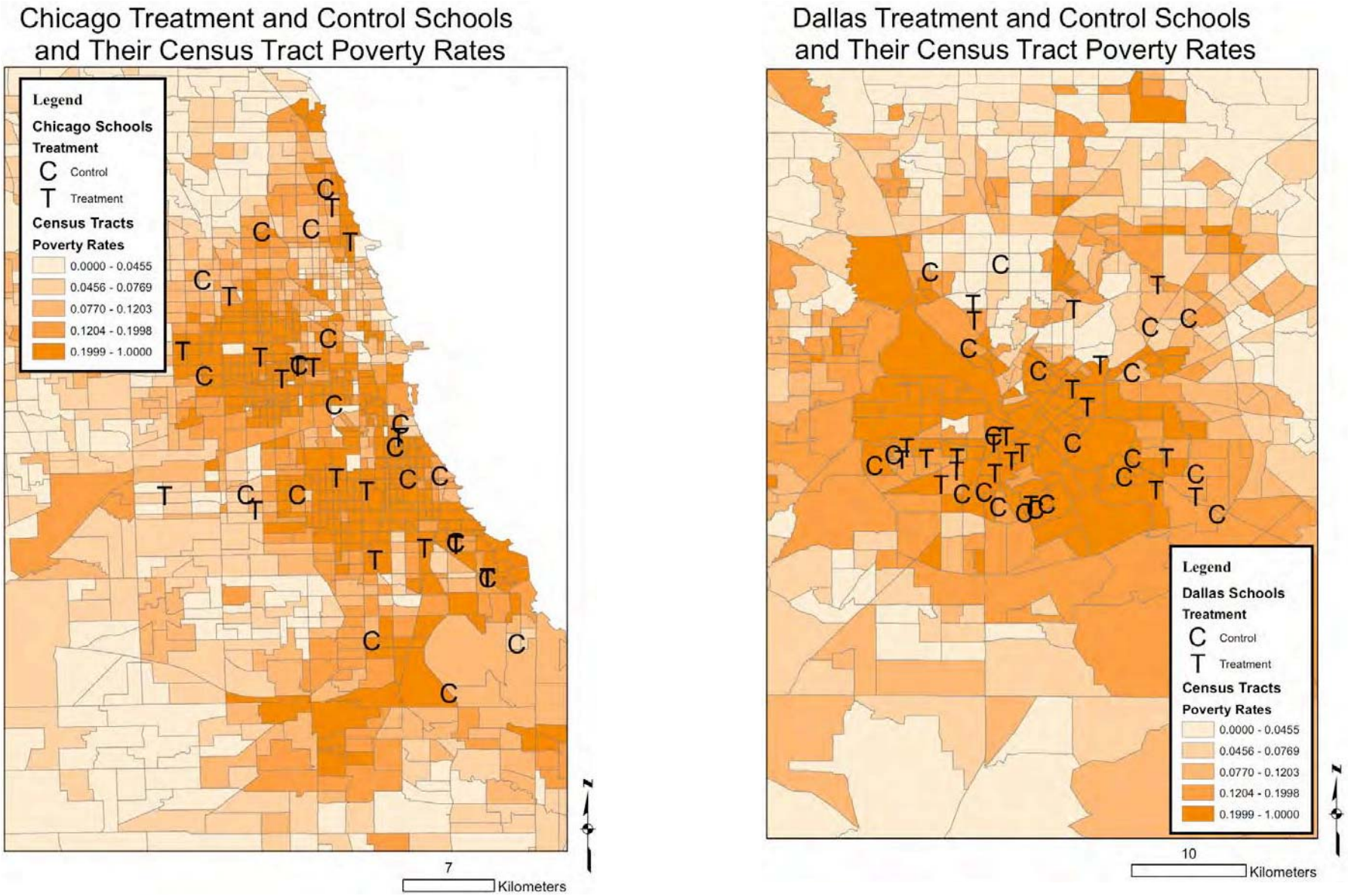
Table 3: The Effect of Financial Incentives on Student Achievement: Inputs

	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	-0.042	-0.054
	(0.071)	(0.097)	(0.096)	(0.116)	(0.198)	(0.251)
N	1900	1900	1756	1756	5844	5844
Rdg. Comp.: All Controls	0.180	0.249	-0.165	-0.200	0.142	0.166
	(0.075)	(0.103)	(0.090)	(0.108)	(0.090)	(0.103)
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	–	–	–	–	-0.053	-0.068
					(0.190)	(0.240)
N					5846	5846
Math: All Controls	–	–	–	–	0.103	0.121
					(0.104)	(0.120)
N					5846	5846

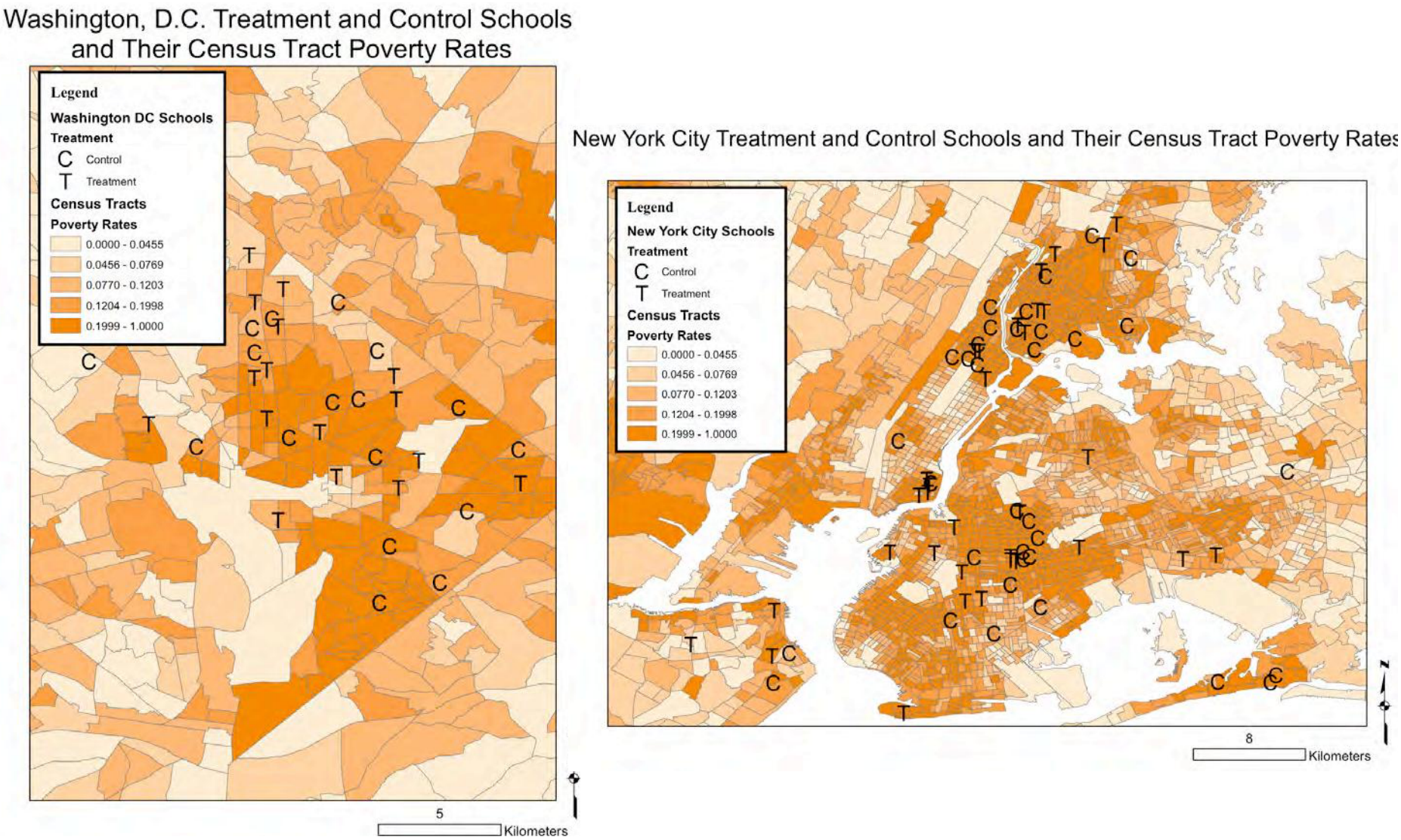
NOTES: The dependent variable is the state assessment taken in each respective city. There were no incentives provided for this test. In Dallas, there were two different types of exams: English-speaking students took the Iowa Tests of Basic Skills (ITBS), and Spanish-speaking students took Logramos tests. 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	Full Sample	White	Black	Hispanic	Asian
Dallas (Books)	2nd	Reading Comp.	0.249	0.319	0.178	0.252	–	0.169	0.314	–
			(0.103)	(0.110)	(0.106)	(0.102)		(0.123)	(0.122)	
		N	1900	995	905	1812		789	1023	
		Reading Vocab.	0.071	0.148	-0.012	0.062	–	0.107	0.035	–
			(0.093)	(0.106)	(0.095)	(0.092)		(0.129)	(0.093)	
		N	1954	1030	924	1863		818	1045	
		Language	0.186	0.241	0.127	0.181	–	0.179	0.197	–
			(0.107)	(0.101)	(0.144)	(0.107)		(0.102)	(0.135)	
		N	1944	1020	924	1854		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	0.166	0.237	0.089	0.166	0.044	0.149	0.253	0.849
			(0.103)	(0.129)	(0.081)	(0.103)	(0.280)	(0.107)	(0.127)	(0.320)
		N	5844	2903	2941	5842	233	4956	555	98
		Math	0.121	0.160	0.073	0.121	-0.770	0.103	0.163	0.551
			(0.120)	(0.132)	(0.114)	(0.120)	(0.152)	(0.124)	(0.135)	(0.416)
		N	5846	2905	2941	5844	233	4948	561	102



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 9B: The Effect of Financial Incentives on Student Achievement: Family-Wise Error Correction – Race

City	Grade Level	Subject	Sample	Effect Size	Naive p value	FWER p value	Bonf. p value	Cons. FWER p value	Cons. Bonf. p value	N
Dallas (Books)	2nd	Reading Comp.	Black	0.117	0.171	0.458	0.513	0.632	1.025	789
				(0.084)						
			Hispanic	0.236	0.020	0.090	0.059	0.180	0.118	1023
				(0.097)						
		Reading Vocab.	Black	0.074	0.404	0.531	1.212	0.733	2.424	818
				(0.088)						
			Hispanic	0.027	0.715	0.734	2.144	0.739	4.288	1045
				(0.072)						
		Language	Black	0.125	0.092	0.395	0.276	0.500	0.553	809
				(0.072)						
			Hispanic	0.148	0.169	0.341	0.508	0.632	1.016	1045
				(0.106)						
DC (Att./Behavior)	6th - 8th	Reading	White	0.031	0.890	0.919	1.780	0.917	7.120	233
				(0.215)						
			Black	0.127	0.179	0.324	0.358	0.792	1.434	4956
				(0.093)						
			Hispanic	0.224	0.069	0.311	0.139	0.689	0.554	555
				(0.119)						
			Asian	0.901	0.058	0.502	0.116	0.689	0.464	98
				(0.430)						
		Math	White	-0.533	0.004	0.189	0.008	0.393	0.032	233
				(0.133)						
			Black	0.088	0.418	0.494	0.836	0.838	3.345	4948
				(0.108)						
			Hispanic	0.145	0.252	0.466	0.504	0.838	2.018	561
				(0.124)						
			Asian	0.618	0.300	0.619	0.601	0.838	2.404	102
				(0.571)						
Chicago (Grades)	9th	English	White	-0.118	0.044	0.305	0.088	0.626	0.352	361
				(0.056)						
			Black	0.010	0.760	0.946	1.520	0.998	6.081	4171
				(0.032)						
			Hispanic	-0.009	0.806	0.976	1.612	0.998	6.450	2943
				(0.038)						
			Asian	0.176	0.259	0.843	0.518	0.979	2.072	136
				(0.149)						

Appendix Table 9E: The Effect of Financial Incentives on Student Achievement: Family-Wise Error Correction – Behavior

City	Grade Level	Subject	Sample	Effect Size	Naive p value	FWER p value	Bonf. p value	Cons. FWER p value	Cons. Bonf. p value	N
DC (Att./Behavior)	6th - 8th	Reading	No Behavioral Incidents	0.113	0.169	0.322	0.338	0.567	0.676	5129
				(0.081)						
			$\geq$ 1 Behavioral Incident	0.222	0.276	0.565	0.553	0.669	1.106	216
				(0.199)						
		Math	No Behavioral Incidents	0.083	0.394	0.477	0.789	0.681	1.577	5123
				(0.096)						
			$\geq$ 1 Behavioral Incident	0.090	0.695	0.755	1.391	0.765	2.781	214
				(0.226)						
NYC (Test Scores)	4th	ELA	No Behavioral Incidents	-0.025	0.452	0.482	0.905	0.723	1.809	6217
				(0.032)						
			$\geq$ 1 Behavioral Incident	0.109	0.273	0.519	0.545	0.658	1.091	191
				(0.098)						
		Math	No Behavioral Incidents	0.072	0.121	0.260	0.242	0.451	0.484	6197
				(0.046)						
			$\geq$ 1 Behavioral Incident	0.042	0.641	0.659	1.282	0.723	2.564	188
				(0.089)						
	7th	ELA	No Behavioral Incidents	0.023	0.230	0.438	0.459	0.691	0.919	9562
				(0.019)						
			$\geq$ 1 Behavioral Incident	-0.047	0.311	0.598	0.621	0.733	1.242	430
				(0.046)						
		Math	No Behavioral Incidents	-0.019	0.591	0.630	1.182	0.733	2.363	9576
				(0.035)						
			$\geq$ 1 Behavioral Incident	0.071	0.378	0.598	0.757	0.733	1.513	423
				(0.079)						

NOTES: The effect sizes reported are intent-to-treat estimates, and the table reports p values computed under various different assumptions. The sixth column reports naive p values, the seventh column reports p values computed using the free step-down resampling procedure (see Anderson (2008) and Westfall and Young (1993)), and the eighth column reports Bonferroni p values. The p values reported in the seventh and eighth columns correspond to defining subsample results for each outcome to be a separate family. The ninth and tenth columns report p values computed under the more conservative assumption that subsample results for multiple outcomes together constitute a family of hypotheses. The final column reports the number of observations for each subsample.