FINANCIAL INCENTIVES AND STUDENT ACHIEVEMENT: EVIDENCE FROM RANDOMIZED TRIALS*

ROLAND G. FRYER JR.

This article describes a series of school-based field experiments in over 200 urban schools across three cities designed to better understand the impact of financial incentives on student achievement. In Dallas, students were paid to read books. In New York, students were rewarded for performance on interim assessments. In Chicago, students were paid for classroom grades. I estimate that the impact of financial incentives on student achievement is statistically 0, in each city. Due to a lack of power, however, I cannot rule out the possibility of effect sizes that would have positive returns on investment. The only statistically significant effect is on English-speaking students in Dallas. The article concludes with a speculative discussion of what might account for intercity differences in estimated treatment effects. *JEL* Codes: I20, I21, I24, J15.

I. INTRODUCTION

The United States is the richest country in the world, but American ninth graders rank 33rd in math, 23rd in science, and 16th in reading achievement.[1] Seventy-seven percent of American students graduate from high school, which ranks the United States in the bottom third of OECD countries (OECD 2010). In large urban areas with high concentrations of blacks

*I am grateful to 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, Nancy Zimmerman, six anonymous referees and the editor, along with seminar participants at Brown, CIFAR, Harvard (Economics and Applied Statistics), Oxford, and University of Pennsylvania for helpful comments. Brad Allan, Austin Blackmon, Charles Campbell, Melody Casagrande, Theodora Chang, Vilsa E. Curto, Nancy Cyr, Will Dobbie, Katherine Ellis, Corinne Espinoza, Peter Evangelakis, 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 exceptional 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. The usual caveat applies.

1. Author's calculations based on data from the 2009 Program for International Student Assessment, which contains data on 65 countries including all OECD countries.

The Quarterly Journal of Economics (2011) 126, 1755–1798. doi:10.1093/qje/qjr045.
Advance Access publication on November 2, 2011.

and Hispanics, educational attainment and achievement are even more bleak, with graduation rates as low as 31% 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. One potentially cost-effective strategy, not yet tested in American urban public schools, is providing short-term financial incentives for students to achieve or exhibit certain behaviors correlated with student achievement. 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.[2] If students lack the structural resources or knowledge to convert effort to measurable achievement or if the production function has important complementarities out of their control (e.g., effective teachers, engaged parents, or social interactions), then incentives will have 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.[3] 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, and New York City—three prototypically low-performing urban school districts—distributing a total of $9.4 million to roughly 27,000

2. Economists estimate that the return to an additional year of schooling is roughly 10% and, if anything, is higher for black students relative to whites (Neal and Johnson 1996; Card 1999; Neal 2006). 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.

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

students in 203 schools (figures include treatment and control).[4] All treatments were school-based randomized trials, which 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. 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. There was a coordinated implementation effort among 20 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.

The incentive schemes were designed to be both simple and politically feasible. In Dallas, we paid second graders $2 per book to read and pass a short quiz to confirm they read it. In NYC, we paid fourth- and seventh-grade students for performance on a series of 10 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. It is important to note that these incentive schemes do not scratch the surface of what is possible. We urge the reader to interpret any results as specific to these incentive schemes and refrain from drawing more general conclusions.

An important potential limitation in our set of field experiments is that they were constructed to detect effects of 0.15 standard deviations or more with 80% power. Thus, we are underpowered to estimate effect sizes below this cutoff, many of which could have a positive return on investment.

The results from our incentive experiments are surprising. The impact of financial incentives on student achievement is statistically 0 in each city. Throughout the text we report intent-to-treat (ITT) estimates, which have been transformed to standard deviation units (hereafter σ). Paying students to read books yields a treatment effect of 0.012σ (0.069) in reading and 0.079σ (0.086) in math. Paying students for performance on standardized tests yielded treatment effects of 0.004σ (0.017) in mathematics

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

and -0.031σ (0.037) in reading in seventh grade and similar results for fourth graders. Rewarding ninth graders for their grades had no effect on achievement test scores in math or reading. Overall, these estimates suggests that incentives are not a panacea—but we cannot rule out small to modest effects (e.g., 0.10σ) which, given the relatively low cost of incentives, have a positive return on investment.

Perhaps even more surprising, financial incentives had little or no effect on the outcomes for which students received direct incentives, self-reported effort, or intrinsic motivation. In NYC, the effect of student incentives on the interim assessments is, if anything, negative. In Chicago, where we rewarded students for grades in five core subjects, the grade point average in these subjects increased 0.093σ (0.057) and treatment students earned 1.979 (1.169) more credits (half a class) than control students. Both of these impacts are marginally significant. We were unable to collect data on the number of books read for students in control schools in Dallas.

Treatment effects on our index of "effort," which aggregates responses to survey questions such as how often students complete their homework or ask their teacher for help, are small and statistically insignificant across all cities, though there may have been substitution between tasks. Finally, using the Intrinsic Motivation Inventory developed in Ryan (1982), we find little evidence that incentives decrease intrinsic motivation. Again, we urge the reader to interpret these results with the important caveat that there may be small effects that we cannot detect.

We conclude our statistical analysis by estimating heterogeneous treatment effects across a variety of subsamples. The key result from this analysis emerges when one partitions students in Dallas into two groups based on whether they took the exam administered to students in bilingual classes (Logramos) or the exam administered to students in regular classes (Iowa Test of Basic Skills). Splitting the data in this way reveals that there is a 0.173σ (0.069) *increase* in reading achievement among English-speaking students and a 0.118σ (0.104) *decrease* in reading achievement among students in bilingual classes. When we aggregate the results in our main analysis this heterogeneity canceled itself out. Similarly, the treatment effect for students who are not English language learners is 0.221σ (0.068) and -0.164 (0.095) for students who are English language learners. This pattern is

not repeated in other cities. Among all other subgroups in Chicago and New York there are no statistically significant differences.

The article is structured as follows. Section II gives a brief review of the emerging experimental literature on the effects of financial incentives on student achievement. Section III provides some details of our experiments and their implementation in each city. Section IV describes our data, research design, and econometric framework. Section V presents estimates of the impact of financial incentives on achievement tests in each city, outcomes that were directly incentivized, self-reported measures of effort, and intrinsic motivation. Section VI provides some discussion and speculation about potential theories that might reconcile the intercity differences in estimates treatment effects. There are two online appendixes. Online Appendix A is an implementation supplement that provides details on the timing of our experimental roll-out and critical milestones reached. Online Appendix B is a data appendix that provides details on how we construct our covariates and our samples from the school district administrative files and survey data used in our analysis.

II. A Brief Literature Review on Incentives for Student Achievement

There is a nascent but growing body of scholarship on the role of incentives in primary, secondary, and postsecondary education around the globe (Angrist et al. 2002; Angrist and Lavy 2009; Angrist, Bettinger, and Kremer 2006; Angrist, Lang, and Oreopoulos 2009; Behrman, Sengupta, and Todd 2005; Bettinger 2010; Barrera-Osorio et al. 2011; Hahn, Leavitt, and Aaron 1994; Jackson 2010; Kremer, Miguel, and Thornton 2009). In this section, we provide a brief overview of the literature on the effect of financial incentives on student achievement, limiting ourselves to analysis from field experiments.[5]

II.A. Incentives in Primary Schools

Psychologists argue that children understand the concept of money as a medium of exchange at a very young age (Marshall and MacGruder 1960), but the use of financial incentives to motivate

5. There are several papers that use nonexperimental methods, including Bettinger (2004), Dynarski (2008), Scott-Clayton (2008), and Jackson (2010).

primary school students is exceedingly rare. Bettinger (2010), who evaluates a pay-for-performance program for students in grades 3 through 6 in Coshocton, Ohio, is a notable exception. Coshocton is 94% white and 55% free/reduced lunch. Students in grades 3 through 6 took achievement tests in five different subjects: math, reading, writing, science, and social studies. Eligible students received $15 for each test on which they scored proficient or better. Students received $20 for scoring "Advanced" or "Accelerated." Bettinger (2010) reports a 0.13σ increase in math scores and no significant effects on reading, social science, or science. Pooling subjects produces an insignificant effect.

The use of nonfinancial incentives—gold stars, aromatic stickers, certificates, and so on—is a more common form of incentive for young children. Perhaps the most famous national incentive program is the Pizza Hut Book It! Program which provides one-topping personal pan pizzas for student readers. This program has been in existence for 25 years, but never credibly evaluated. The concept of the Book It! program, providing incentives for reading books, is very similar to our reading incentive experiment in Dallas.

II.B. Incentives in Secondary Schools

Experiments on financial incentives in secondary school have been concentrated outside the United States. Kremer, Miguel, and Thornton (2009) conduct a randomized evaluation of a merit scholarship program in Kenya for girls. Girls in grade 6 from program schools who scored in the top 15% in the district received an award over the next 2 years: in each year, a winner would receive a grant of $6.40 to cover school fees, paid to the winner's school; a grant of $12.80 for school supplies, paid to the winner's family; and public recognition at a school awards assembly. Scholarships were awarded on the basis of performance on district-wide exams in five subjects. Kremer, Miguel, and Thornton (2009) find that the program raises test scores by 0.19σ for girls and 0.08σ for boys, though boys were ineligible for any rewards.

In December 2000, the Israeli Ministry of Education selected 40 schools with low Bagrut passage rates to participate in an incentives program called the Achievement Awards program. Bagrut is a high school matriculation certificate. Angrist and Lavy (2009) evaluate results for high school seniors, who were offered approximately $1,500 for receiving the Bagrut. The results are

positive but insignificant in the full sample. When the sample is divided by gender, however, they find significantly positive effects on Bagrut receipt rates among girls.

II.C. Incentives in Postsecondary School

There are many programs to incentivize college students for various behaviors ranging from giving plasma to achieving a certain GPA as a condition to keep their financial aid (Cornwell, Mustard, and Sridhar, 2006). Angrist, Lang, and Oreopoulos (2009) present results from an evaluation of a program called the Student Achievement and Retention (STAR) Demonstration Project at a large Canadian university. Students who were below the top quartile in their incoming high school GPAs were randomly assigned to one of three treatment arms or to a control group. In the first treatment arm, students were offered access to a peer-advising service as well as supplemental instruction in the form of facilitated study groups. In the second treatment arm, students were offered fellowships of up to $5,000 cash (equivalent to a year's tuition) for maintaining at least a B average, and $1,000 for maintaining at least a C+ average. To be eligible for the program, students had to take at least four courses per term (regular course load is five courses per term) and register for the second year. In the third treatment arm, students were eligible both for the study services and for the fellowship. There were 250 students in each of the first two treatment arms and 150 students in the third treatment arm. The control group was comprised of 1,006 students who were not offered any of the treatments. Angrist, Lang, and Oreopoulos (2009) report that students in the services-only treatment arm did not earn higher GPAs and were not more like to be retained than students in the control group. The fellowship had no effect on GPAs. The combined services/fellowship treatment arm had positive effects on fall semester GPA, but the results were not sustained.

Leuven, Oosterbeek, and van der Klaauw (2010) examine the impacts of a randomized experiment on first-year students at the University of Amsterdam. Students were randomly assigned to one of three groups: a large reward group that could earn a bonus of 681 euros by completing all the first-year requirements by the start of the next academic year; a small reward group that could earn 227 euros for completing these requirements; and a control group that could not earn an award. For the randomization,

students were stratified on high school math score (8 intervals) and parents' education (3 intervals), so that there were 24 strata. Overall, 83 students were assigned to the large reward group, 84 students to the small reward group, and 82 students to the control group. They find that the large reward has a small and insignificant positive effect.

III. Program Details

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

In total, experiments were conducted in 203 schools across three cities, distributing $9.4 million to approximately 27,000 students.[6] All experiments had a similar implementation plan. 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. After principals were given information about the experiment, there was a brief sign-up period, typically 5–10 days. 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 not chosen. Students received their first payments the second week of October and their last payment was disseminated over the summer. All experiments lasted one academic year.

III.A. Dallas

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

Forty-two schools signed up to participate in the Dallas experiment, and we randomly chose 21 of those schools to be

6. Roughly half the students and half the schools were assigned to treatment and the other half to control.

TABLE I

SUMMARY OF INCENTIVES EXPERIMENTS

	Dallas	NYC	Chicago
Schools	42 schools opted in to participate, 21 schools randomly chosen for treatment	121 schools opted in to participate, 63 schools randomly chosen for treatment	70 schools opted in to participate, 20 schools randomly chosen for treatment from a predetermined set of 40
Number of students in treatment sample	1,777 2nd-grade students: 22% black, 77% Hispanic, 59% free lunch eligible	3,348 4th-grade students: 44% black, 44% Hispanic, 89% free lunch eligible; 4,605 7th-grade students: 37% black, 42% Hispanic, 87% free lunch eligible	3,275 9th-grade students: 57% black, 37% Hispanic, 93% free lunch eligible
Number of students in control sample	1,941 2nd-grade students: 22% black, 74% Hispanic, 57% free lunch eligible	3,234 4th-grade students: 45% black, 42% Hispanic, 92% free lunch eligible; 4,696 7th-grade students: 45% black, 38% Hispanic, 91% free lunch eligible	4,380 9th-grade students: 53% black, 40% Hispanic, 92% 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 $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 ($1,875 max).

TABLE I
(CONTINUED)

	Dallas	NYC	Chicago
Frequency of rewards	3 times per year	5 times per year	Every 5 weeks / report card
Outcomes of interest	ITBS/Logramos reading and math scores	New York state assessment ELA and math scores	PLAN English and math scores
Testing dates	May 12–23, 2008	January 13–15, 2009 (4th ELA), January 21–22 (7th ELA), March 3–5 (4th math), March 10–11 (7th math)	September 28–October 7, 2009
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.	$3,000,000 distributed. 88.97% consent rate. 91% of students understood the basic structure of the incentive program. Two dedicated project managers.

Notes. Entries are descriptions of the schools, students, reward structure, frequency of rewards, outcomes of interest, testing dates, and basic operations of the incentive treatments. See Appendix A for more details. In Dallas, 43 schools originally opted in to participate, however, one of these schools only had students in grades 4 and 5 and is therefore excluded from the second-grade analysis. In New York City, 143 schools originally opted in to participate. However, special education schools, schools with many students participating in the Opportunity NYC Family Rewards conditional cash transfer program, and schools that closed between 2007–2008 and 2008–2009 are dropped from the analysis. In Chicago, the experimental sample was limited to eligible schools with the lowest numbers of enrolled ninth graders due to budgetary constraints. The numbers of treatment and control students given are for those students who have non-missing reading or math test scores. See Online Appendix Table I for a full accounting of sample sizes.

treated (more on our randomization procedure shortly).[7] The experimental group was comprised of 3,718 second-grade students.[8] To participate, students were required to have a parental consent form signed; 83% of students in the treatment sample signed up to participate. Participating schools received $1,500 to lower the cost of implementation.

Table II explores differences, on a set of covariates, between schools that signed up to participate relative to those that did not sign up to participate. The first three columns compare the 42 experimental schools in Dallas with the other 109 schools in the DISD that contain a second grade. The experimental schools are more likely to have students on free lunch and have lower average reading scores. The racial distribution, percent English language learner, percent special education, and total enrollment are all similar between schools that opted to participate in the experiment and those that did not.

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 80% 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). To reduce the possibility of cheating, quizzes were taken in the library on a computer and students were only allowed one chance to take a quiz.

An important caveat of the Dallas experiment is that we combine AR (a known software program) with the use of incentives. If the AR program has an independent (positive) effect on student achievement, we will assign more weight to incentives than is warranted. The only two experimental analyses of the impact of AR on students who are learning to read that meet the standards of the "What Works Clearinghouse," Bullock (2005) and Ross, Nunnery, and Goldfeder (2004), report no discernible

7. Forty-three schools originally signed up, but one had students only in grades 4 and 5, so we exclude it from the discussion of the second-grade analysis.

8. This is the number of second-grade students in the experimental group with nonmissing reading or math achievement outcomes at the end of the school year.

TABLE II

BASELINE CHARACTERISTICS OF NONEXPERIMENTAL AND EXPERIMENTAL SCHOOLS

	Dallas			NYC			Chicago		
Variable	Non-experimental schools	Experimental schools	*p*-value	Non-experimental schools	Experimental schools	*p*-value	Non-experimental schools	Experimental schools	*p*-value
Student characteristics									
Percent black	0.296	0.269	0.606	0.327	0.453	0.000	0.582	0.669	0.211
Percent Hispanic	0.641	0.707	0.210	0.395	0.436	0.101	0.323	0.282	0.498
Percent Asian	0.010	0.007	0.325	0.124	0.057	0.000	0.026	0.015	0.235
Percent male	0.506	0.508	0.682	0.505	0.504	0.869	0.500	0.487	0.556
Percent free lunch	0.527	0.572	0.053	0.849	0.930	0.000	0.881	0.935	0.012
Percent English language learner	0.409	0.472	0.102	0.141	0.146	0.715	0.006	0.008	0.335
Percent special education	0.070	0.067	0.445	0.088	0.100	0.016	—	—	—
Total enrollment	570.954	575.476	0.900	636.661	632.405	0.894	856.489	933.450	0.591
Mean 2nd grade ITBS reading 2006–07	7.719	7.365	0.077	—	—	—	—	—	—
Mean 2nd grade ITBS math 2006–07	2.749	2.736	0.862	—	—	—	—	—	—
Mean 4th grade NYS ELA 2006–07	—	—	—	655.500	646.255	0.000	—	—	—
Mean 4th grade NYS math 2006–07	—	—	—	675.886	664.398	0.000	—	—	—
Mean 7th grade NYS ELA 2006–07	—	—	—	645.493	637.604	0.002	—	—	—
Mean 7th grade NYS math 2006–07	—	—	—	654.535	646.271	0.003	—	—	—
Mean 10th grade PLAN English 2008–09	—	—	—	—	—	—	14.542	13.349	0.011
Mean 10th grade PLAN math 2008–09	—	—	—	—	—	—	14.939	13.972	0.012

TABLE II
(CONTINUED)

Variable	Dallas			NYC			Chicago		
	Non-experimental schools	Experimental schools	*p*-value	Non-experimental schools	Experimental schools	*p*-value	Non-experimental schools	Experimental schools	*p*-value
Teacher characteristics									
Percent black	—	—	—	0.217	0.340	0.000	—	—	—
Percent Hispanic	—	—	—	0.148	0.176	0.044	—	—	—
Percent Asian	—	—	—	0.045	0.040	0.405	—	—	—
Percent male	—	—	—	0.177	0.215	0.001	—	—	—
Average years teaching experience	—	—	—	8.272	7.796	0.055	—	—	—
Average teacher salary/1,000	—	—	—	67.329	66.415	0.039	—	—	—
Progress report results (2006–07)									
School environment score	—	—	—	0.511	0.465	0.010	—	—	—
Student performance score	—	—	—	0.551	0.493	0.001	—	—	—
Student progress score	—	—	—	0.501	0.493	0.649	—	—	—
Overall score	—	—	—	54.023	51.327	0.047	—	—	—
Progress report grade	—	—	—	2.696	2.403	0.005	—	—	—
Number of schools	109	42		924	121		94	40	

Notes. The first, fourth, and seventh columns present nonexperimental school means of the variable indicated in each row. The second, fifth, and eighth columns present experimental school means. The third, sixth, and ninth columns present *p*-values for the differences in means between the previous two columns.

effect or mixed effects of the AR program. Bullock (2005) finds no significant effect of AR on third graders when measured using the Oral Reading Fluency subtest of the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). The positive effects on reading comprehension found by Ross, Nunnery, and Goldfeder (2004) are not statistically significant.

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.

III.B. 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 70% of NYC students are black or Hispanic, 15% are English language learners, and over 70% are eligible for free lunch.

One hundred twenty-one schools signed up to participate in the New York City experiment, and we randomly chose 63 schools (33 fourth grades and 31 seventh grades) to be treated.[9] The experimental sample consists of 15,883 students. A participating school received $2,500 if 80% 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 80% of students were signed up and if the school had administered all six assessments.

Columns four through six in Table II compare the schools that opted to participate in the incentive program with all other schools in NYC that contain either a fourth or seventh grade. Experimental schools are significantly more likely to have students who are nonwhite, on free lunch, in special education, and have lower test scores. In other words, the schools that opted

9. Grades and schools do not add up because there is one treatment school that contained both fourth and seventh grades and both grades participated. One hundred forty-three schools originally signed up to participate. However, special education schools, schools with many students participating in the Opportunity NYC Family Rewards conditional cash transfer program, and schools that closed between 2007–2008 and 2008–2009 are dropped from the analysis.

to participate in NYC were predominantly minority and poor performing.

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. 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, students were required to turn in signed parental consent forms; 73% 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 66% 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; 90% of students understood the basic structure of the incentive program. See Online Appendix A for more details.

III.C. Chicago

The Chicago experiment took place in 20 low-performing Chicago public high schools. Chicago is the third largest school district in the United States with over 400,000 students, 88.3% of whom are black or Hispanic. Seventy-five percent of students in Chicago are eligible for free or reduced lunch, and 13.3% are English language learners.

Seventy schools signed up to participate in the Chicago experiment. To control costs, we selected 40 of the smallest schools out of the 70 that wanted to participate and then randomly selected 20 to treat within this smaller set. Once a school was selected, students were required to return signed parental consent forms to participate. The experimental sample consisted of 7,655 ninth

graders, of whom 3,275 were in the treatment group.[10] Ninety-four percent of the treatment 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.

Columns seven through nine in Table II compare the schools that opted to participate in the incentive program with all other schools in Chicago that enrolled ninth graders. Experimental schools had higher percentages of students who were eligible for free lunch and had lower scores on the PLAN English and math tests. As in other cities, experimental schools tended to serve low-income students and perform poorly on achievement tests.

Students in Chicago were given incentives for their grades in five core courses: English, mathematics, science, social science, and gym.[11] 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 was held in an account and given in a lump sum conditional on high school graduation. The average student earned $695.61 and the highest achiever earned $1,875.

IV. Data, Research Design, and Econometrics

We collected both administrative and survey data. The richness of the administrative data varies by school district. For all cities, the data include information on each student's first and last name, birth date, address, race, gender, free lunch eligibility, attendance, matriculation with course grades, special education status, and English language learner (ELL) status. In Dallas and New York, we are able to link students to their classroom teachers. New York City administrative files contain teacher value-added

10. This is the sample of students with nonmissing PLAN reading or math achievement outcomes. See Online Appendix Table I for a full accounting of sample sizes.

11. Gym may seem like an odd core course in which to provide incentives for achievement, but roughly 22% of ninth-grade students failed their gym courses in the year prior to our experiment.

data for teachers in grades 4–8, as well as data on student suspensions and behavioral incidents.

Our main outcome variable is an achievement test unique to each city. We did not provide incentives of any form for these assessments. All Chicago 10th graders take the PLAN assessment, an ACT college-readiness exam, in October. In May of every school year, students in regular classes in Dallas elementary schools take the Iowa Tests of Basic Skills (ITBS) if they are in kindergarten, first grade, or second grade. Students in bilingual classes in Dallas take a different exam, called Logramos.[12] In New York City, the mathematics and English Language Arts tests, developed by McGraw-Hill, are administered each winter to students in grades 3–8. There are two important drawbacks of the PLAN assessment in Chicago. First, PLAN is a pre-ACT test and is only loosely related to the everyday teaching and learning in the classroom. Thus, there may be a low causal effect of GPA on PLAN. Second, the PLAN is administered in the fall, approximately four months after our experiment ended. See Online 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 2 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 Table III, Panels A through C for exact percentages of experimental group students with valid test scores from previous years). We also include an indicator variable that takes on the value of 1 if a student is missing a test score from a previous year and 0 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 ELL. A student is income-eligible for free lunch if her family income is below 130% 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

12. The Spanish test is not a simple translation of ITBS. Throughout the text, we present estimates of these tests together. In Table IV, our analysis of subsamples, we estimate the effect of incentives on each test separately.

TABLE III
STUDENT BASELINE CHARACTERISTICS

Variable	Control mean	Treatment mean	p-value: C v. T
A. Dallas			
ITBS reading 2006–07	8.608	8.392	0.683
ITBS math 2006–07	1.655	1.503	0.089
ITBS reading 2005–06	5.428	5.715	0.655
ITBS math 2005–06	0.947	0.893	0.192
Took English ITBS test in 2006–07	0.497	0.509	0.798
Took Spanish ITBS test in 2006–07	0.503	0.491	0.798
White	0.022	0.009	0.152
Black	0.222	0.219	0.978
Hispanic	0.743	0.768	0.737
Asian	0.012	0.003	0.142
Other race	0.002	0.001	0.432
Male	0.522	0.509	0.476
Female	0.478	0.491	0.476
Free lunch	0.570	0.586	0.655
English language learner	0.539	0.519	0.705
Special education	0.040	0.048	0.362
Percent black	0.232	0.231	0.996
Percent Hispanic	0.734	0.751	0.819
Percent free lunch	0.582	0.597	0.618
Missing ITBS reading score 2006–07	0.118	0.144	0.087
Missing ITBS math score 2006–07	0.275	0.273	0.980
Missing ITBS reading score 2005–06	0.818	0.796	0.519
Missing ITBS math score 2005–06	0.565	0.545	0.656
Number of students	1,941	1,777	
Fourth grade			
B. NYC			
NYS ELA 2007–08	652.659	654.419	0.545
NYS math 2007–08	677.944	679.198	0.659
NYS ELA 2006–07	598.651	601.579	0.412
NYS math 2006–07	634.834	632.794	0.598
White	0.029	0.043	0.601
Black	0.455	0.442	0.870
Hispanic	0.422	0.438	0.824
Asian	0.085	0.071	0.762
Other race	0.009	0.006	0.243
Male	0.515	0.507	0.597
Female	0.485	0.493	0.597
Free lunch	0.916	0.890	0.361

TABLE III
(CONTINUED)

Fourth grade			
B. NYC			
English language learner	0.153	0.167	0.666
Special education	0.100	0.118	0.319
Percent black	0.442	0.424	0.819
Percent Hispanic	0.424	0.443	0.787
Percent free lunch	0.910	0.889	0.405
Individual-level behavior 2007–08	0.160	0.106	0.132
School-level behavior 2007–08	111.091	65.677	0.060
Missing NYS ELA score 2007–08	0.061	0.070	0.253
Missing NYS math score 2007–08	0.042	0.050	0.147
Missing NYS ELA score 2006–07	0.954	0.958	0.604
Missing ITBS math score 2005–06	0.951	0.958	0.453
Number of students	3,234	3,348	
Seventh grade			
NYS ELA 2007–08	648.805	648.939	0.977
NYS math 2007–08	661.833	662.573	0.923
NYS ELA 2006–07	650.588	651.327	0.910
NYS math 2006–07	663.429	664.819	0.852
White	0.072	0.080	0.904
Black	0.448	0.370	0.415
Hispanic	0.384	0.422	0.671
Asian	0.090	0.126	0.544
Other race	0.006	0.002	0.013
Male	0.497	0.510	0.460
Female	0.503	0.490	0.460
Free lunch	0.913	0.868	0.413
English language learner	0.138	0.137	0.975
Special education	0.098	0.117	0.310
Percent black	0.441	0.370	0.444
Percent Hispanic	0.390	0.420	0.735
Percent free lunch	0.906	0.877	0.482
Individual-level behavior 2007–08	0.122	0.244	0.025
School-level behavior 2007–08	124.637	168.455	0.297
Missing NYS ELA score 2007–08	0.073	0.067	0.668
Missing NYS math score 2007–08	0.085	0.045	0.236
Missing NYS ELA score 2006–07	0.111	0.111	0.957
Missing ITBS math score 2005–06	0.091	0.091	0.981
Number of students	4,696	4,605	

TABLE III
(CONTINUED)

C. Chicago			
ISAT reading 2007–08	240.571	239.930	0.810
ISAT math 2007–08	257.627	257.595	0.992
ISAT reading 2006–07	229.244	228.940	0.925
ISAT math 2006–07	243.949	243.984	0.992
White	0.046	0.049	0.936
Black	0.534	0.572	0.804
Hispanic	0.397	0.368	0.839
Asian	0.024	0.010	0.312
Other race	0.000	0.001	0.407
Male	0.468	0.489	0.250
Female	0.532	0.511	0.250
Free lunch	0.920	0.931	0.683
English language learner	0.008	0.006	0.591
Percent black	0.556	0.566	0.947
Percent Hispanic	0.352	0.349	0.984
Percent free lunch	0.917	0.932	0.619
Missing ISAT reading score 2007–08	0.091	0.093	0.890
Missing ISAT math score 2007–08	0.086	0.088	0.870
Missing ISAT reading score 2006–07	0.125	0.121	0.802
Missing ISAT math score 2006–07	0.126	0.122	0.827
Number of students	4,380	3,275	

Note. Within each panel, the first column presents the mean for control students of the variable indicated in each row. The second column presents the mean for treatment students. The third column presents the p-value of the difference in means between control students and treatment students. To account for possible intraschool correlation, this is calculated by regressing each baseline variable on an indicator for being in treatment, clustering standard errors at the school level, and using the p-value corresponding to the t-statistic for the treatment indicator variable.

Reservations, or the Temporary Assistance for Needy Families Program; (2) the student was enrolled in Head Start on the basis of meeting that program's low-income criteria; (3) the student is homeless; (4) the student is a migrant child; or (5) the student is a runaway child receiving assistance from a program under the Runaway and Homeless Youth Act and is identified by the local educational liaison. Determination of special education and ELL status varies by district.

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 and New York, we

assign each student who was present at the beginning of the year, that is, 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 three 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 the Intrinsic Motivation Inventory described in Ryan (1982). In Dallas we offered up to $2,000 (prorated by size) for schools in which 90% or more of the surveys were completed. Eighty percent of surveys were returned in Dallas treatment schools and 89% were returned in control schools. In the two other cities, survey responses were lower. In Chicago, despite offering $1,000 per school to schools for collecting 90% of the surveys, only 35% of surveys in treatment schools and 39% of surveys in control schools were returned. In New York City, we were able to offer $500 to schools to administer the survey and could not condition the payment on survey response rate. Fifty-eight percent of surveys were returned in the treatment group and 27% were returned in control schools.

Given the combination of administrative and survey data, sample sizes will differ with various outcomes tested, due to missing data. Online Appendix Table I provides an accounting of sample sizes across all outcomes in our analysis. The overall samples are depicted in the first panel. There are 4,008 students in Dallas, 16,449 in NYC, and 10,628 in Chicago. Below that, we provide the number of students with non-missing data for each city and each outcome.

IV.A. 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. We were also concerned that randomizing within schools could prompt some teachers to provide alternative nonmonetary incentives to control students (unobservable to us) that would undermine the experiment. The same procedure was used in each city to randomly partition the set of interested schools into treatment and control schools.

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 set of possibilities—2.113×10^{41} in New York—we randomly selected 10,000 treatment-control designations and estimated equations identical to:

$$(1) \qquad treatment_s = \alpha + X_s\beta + \varepsilon_s,$$

where the dependent variable takes on the value of 1 for all treatment schools and s represents data measured at the school level that were available at the time of randomization. We then selected the randomization that minimized the maximum z-score from Equation (1). This method was chosen with the goal of achieving balance across a set of predetermined subsamples—race, previous year's test score, whether or not a student is eligible for free lunch or an ELL—without forcing exact balance on each, given our small samples.

There is an active discussion on which randomization procedures have the best properties. Treasure and MacRae (1998) prefer the method just described. Imbens and Wooldridge (2009) and Greevy et al. (2004) recommend matched pairs. Results from simulation evidence presented in Bruhn and McKenzie (2009) suggest that for large samples there is little gain from different methods of randomization over a pure single draw. For small samples, however, matched pairs, rerandomization (the method employed here), and stratification all perform better than a pure random draw. Following the recommendation of Bruhn and McKenzie (2009), we have estimated our treatment effects including all variables to check balance. Whether we include these variables or a richer set of controls described above does not significantly alter the results. We choose to include the richer, individual-level, controls.

Table III tests covariate balance by providing means for all pretreatment variables, by city, for students in the experimental sample. For each variable, we provide a p-value for treatment and control differences in the last column. Across all cities, our

randomization resulted in balance across all covariates with the exception of "other race" and behavioral incidences among seventh graders in New York City. To complement Table I, Online Appendix Figures IA and IB 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 each city.

IV.B. Econometric Models

To estimate the causal impact of providing student incentives on outcomes, we estimate intent-to-treat (ITT) effects, that is, 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. All these variables are pretreatment measures. The ITT effect, π_1, is estimated from

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

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 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.[13] 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.

Typically, in the program evaluation literature, there are also estimates of the "treatment on the treated," which captures the effect of actually participating in a program. We focus on ITT because due to the plausibility that our school-level

13. 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.

randomization resulted in spillovers to students who did not enroll in the incentive program through more focused instruction by teachers, general excitement about receiving incentives, and so on. If true, this would be an important violation of the assumptions needed to credibly identify "treatment on the treated" estimates.

V. The Impact of Financial Incentives on Student Achievement

V.A. State Test Scores

Table IV presents ITT estimates for Dallas, NYC, and Chicago, separately, as well as a pooled estimate across all cities. All results are presented in standard deviation units. Standard errors, clustered at the school level, are in parentheses below each estimate.

The impact of offering incentives to students is statistically 0 across all cities individually and pooled. More precisely, as demonstrated in Table IV, the ITT effect of incentives on reading achievement is 0.012σ (0.069) in Dallas, -0.026σ (0.034) for fourth graders in NYC, 0.004σ (0.017) for seventh graders in NYC, and -0.006σ (0.028) in Chicago. Pooling across all cities yields a treatment effect of -0.008σ (0.018). The patterns in math are similar. The ITT effect of incentives on math achievement is 0.079σ (0.086) in Dallas, 0.062σ (0.047) for fourth graders in NYC, -0.031σ (0.037) for seventh graders in NYC, and -0.010σ (0.023) in Chicago. Pooling across all cities yields a treatment effect in math of 0.008σ (0.022).[14]

Due to low power, however, all of the estimates have 95% confidence intervals that contain effect sizes that would have a positive return on investment. Using a cost-benefit framework identical to that in Krueger (2003), one can show that effect sizes as small as 0.0006 in Dallas, 0.004 for fourth grade in NYC, 0.006 for seventh grade in NYC, and 0.016 in Chicago have a 5% return on investment. Thus, due to the low costs of incentive interventions, to the extent that a coefficient is positive, it may have a positive return on investment. Yet we only have enough power to detect 0.15σ effects, so many values that have a high return we are unable to detect.

14. Online Appendix Table II provides first stage and treatment on the treated estimates that are similar in magnitude.

TABLE IV

MEAN EFFECT SIZES (INTENT-TO-TREAT ESTIMATES) ON ACHIEVEMENT

	Dallas	NYC		Chicago	
Outcome	2nd	4th	7th	9th	Pooled
Reading	0.012	−0.026	0.004	−0.006	−0.008
	(0.069)	(0.034)	(0.017)	(0.028)	(0.018)
	3,608	6,497	9,152	7,616	26,873
Math	0.079	0.062	−0.031	−0.010	0.008
	(0.086)	(0.047)	(0.037)	(0.023)	(0.022)
	3,712	6,517	9,221	7,599	27,049

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 0 and a standard deviation of 1 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 effect size 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. It is the intent-to-treat (ITT) estimate on achievement. The second column presents results for second graders who participated in the Earning by Learning experiment in Dallas. The third and fourth columns present results for fourth and seventh graders, respectively, who participated in the Spark experiment in New York City. The fifth column presents results for ninth graders who participated in the Paper Project experiment in Chicago. The sixth column presents results that are pooled across the three sites, and these regressions include site and grade dummies. All regressions include controls for reading and math test scores from the previous 2 years and their squares, race, gender, free/reduced lunch eligibility, English language learner status, the percent of black students in the school, the percent of Hispanic students in the school, and the percent of free/reduced lunch students in the school. For Dallas, regressions also include a control for whether the student took the English or Spanish version of the ITBS/Logramos test in the previous year. For Dallas and New York City, regressions also include an indicator for being in special education. For New York City, regressions also include controls for the number of recorded behavioral incidents a student had in the previous year, as well as the number of recorded behavioral incidents the school had in the previous year. All standard errors, in parentheses, are clustered at the school level. The numbers of observations are located directly below the standard errors.

One might worry that with several school-level covariates we may be overfitting a handful of observations. Online Appendix Table III estimates treatment effects with school-level regressions. The results are strikingly similar to those using individual-level data. The average (across reading and math) pooled estimate when we estimate treatment effects at the individual level is almost exactly 0. The same estimate, using school-level regressions, is −0.008.

Another potential concern for estimation is that we only include students for which we have post-treatment test scores. If students in treatment schools and students in control schools have different rates of selection into this sample, our results may be biased. Online Appendix Table IV compares the rates of attrition of students in treatment schools and students in control schools. The first row regresses whether or not a student switches schools during the school year on a treatment dummy and our parsimonious set of controls. The numbers reported in the table are the coefficients on the treatment indicator. The second row has

whether a student has a nonmissing reading score as an outcome. The final row reports similar results for math scores.

Across all cities there is little evidence that there was differential mobility across treatment and control schools in the year of treatment. Consistent with this, we find similar results for reading and math scores in Dallas and NYC. In Chicago, however, students in treatment schools are 6% more likely to have a nonmissing test score.

In summary, financial incentives for student achievement is not a panacea. Yet due to their low cost and our lack of power, we cannot rule out effect sizes that have a positive return on investment.

V.B. Direct Outcomes, Effort, and Intrinsic Motivation

The previous results reported the impact of incentives on state test scores—an indirect and nonincentivized outcome. Table V presents estimates of the effect of incentives on outcomes for which students where given direct incentives, their self-reported effort, and intrinsic motivation. Outcomes for which students were given direct incentives include: books in Dallas, predictive tests in NYC, and report card grades in Chicago. Treatment students in Dallas read, on average, 12 books in the year of the experiment. Unfortunately, we were unable to determine how many books students in control schools read during the experiment. The predictive tests in NYC is designed to be a good predictor of student achievement on the state tests and is required of all schools. An important benefit of the predictive exams is that they are administered on the last day of our experiment in June. The state tests were administered in January and March, which truncates the length of the treatment. In Chicago, 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 0 and a standard deviation of 1 among students in the same grade across the school district.

Along with the outcomes, Table V also reports results for measures of effort. Data on student effort are 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

TABLE V

MEAN EFFECT SIZES (INTENT-TO-TREAT ESTIMATES) ON DIRECT OUTCOMES, EFFORT, AND INTRINSIC MOTIVATION

A. Dallas

	Effort and intrinsic motivation		
	Attendance rate	Effort index	Intrinsic motivation index
Second grade	−0.040	−0.060	−0.020
	(0.051)	(0.073)	(0.068)
	3,778	1,904	1,746

B. NYC

	Direct outcomes				Effort and intrinsic motivation		
	ELA winter	ELA summer	Math winter	Math summer	Attendance rate	Effort index	Intrinsic Motivation Index
Fourth grade	−0.072	−0.068	−0.042	−0.057	—	—	—
	(0.039)	(0.044)	(0.041)	(0.052)			
	5,942	5,908	5,788	5,872			
Seventh grade	−0.039	−0.053	0.000	−0.115	−0.087	0.042	−0.048
	(0.028)	(0.046)	(0.035)	(0.047)	(0.044)	(0.049)	(0.049)
	7,096	7,242	7,449	7,329	9,655	2,816	2,679

TABLE V
CONTINUED

C. Chicago

	Direct outcomes		Effort and intrinsic motivation				
	GPA	Core GPA	Attendance rate	Credits earned	Credits attempted	Effort index	Instrinsic motivation index
Ninth grade	0.093	0.079	0.152	1.979	0.907	0.000	0.017
	(0.057)	(0.057)	(0.082)	(1.169)	(0.535)	(0.051)	(0.065)
	10,613	10,613	10,628	10,221	10,221	1,875	1,698

Notes. The dependent variable is the variable indicated in each column heading. The attendance rate, predictive exam scores, GPA and core GPA outcomes have been normalized to have a mean of 0 and a standard deviation of 1 within each grade across the entire sample of students in the school district. The effort index and intrinsic motivation index have been normalized to have a mean of 0 and a standard deviation of 1 within each grade across the sample of experimental students in the school district (since surveys were only administered to experimental students). Thus, coefficients for these outcomes are in standard deviation units. The effect size is the difference between mean outcomes of students in schools randomly chosen to participate and mean outcomes of students in schools that were not chosen. It is the intent-to-treat (ITT) estimate on the relevant outcome. Panel A presents results for second graders who participated in the Earning by Learning experiment in Dallas. Panel B presents results for fourth and seventh graders who participated in the Spark experiment in New York City. Panel C presents results for ninth graders who participated in the Paper Project experiment in Chicago. All regressions include controls for reading and math test scores from the previous 2 years and their squares, race, gender, free/reduced lunch eligibility, English language learner status, the percent of black students in the school, the percent of Hispanic students in the school, and the percent of free/reduced lunch students in the school. For Dallas, regressions also include a control for whether the student took the English or Spanish version of the ITBS/Logramos test in the previous year. For Dallas and New York City, regressions also include an indicator for being in special education. For New York City, regressions also include controls for the number of recorded behavioral incidents a student had in the previous year, as well as the number of recorded behavioral incidents the school had in the previous year. All standard errors, in parentheses, are clustered at the school level. The numbers of observations are located directly below the standard errors.

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.[15] Students responded to these questions by selecting answers among "Never," "Some of the Time," "Half of the Time," "Most of the Time," and "All of the Time." We converted these responses to a numerical scale from 1 to 5, where a higher number indicated higher self-reported effort, and then added up all of a student's responses to effort-related questions to obtain an effort index. We then normalized the effort index to have a mean of 0 and a standard deviation of 1 in the experimental sample. See Online Appendix B for further details. We also include attendance as an outcome as we view this as a form of student effort.

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.[16] 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

15. Because participating students in Dallas are only in second grade, they were only asked questions 2 and 4.

16. 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 4). Only students with valid responses to all statements are included in our analysis, as nonresponse may be confused with low intrinsic motivation.[17]

Surprisingly, the treatment effect on predictive tests in NYC (in which treated students were given direct incentives) is negative, two of which are statistically significant. The most relevant predictive tests are those administered at the end of the experiment, labeled ELA summer and Math summer in Table V. For seventh graders, the ITT estimate on these exams is -0.053σ (0.046) in ELA and -0.115σ (0.047) in math. Fourth graders in NYC demonstrate a similar pattern. Paying students for better course grades in core subjects has a modest impact on their grades—an increase of 0.093σ (0.057) in GPA and an increase of 1.979 (1.169) 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 effect of financial incentives on direct outcomes paints a similar picture to that obtained in our analysis of state tests.

The effect of incentives on student effort and intrinsic motivation, also shown in Table V, indicates that there are few differences on the dimensions of effort between those students who received treatment and those who did not, though our estimates are imprecise. The average treatment effect across all sites on the effort index is -0.006, though in Chicago there is some evidence that attendance increased. The average treatment effect on intrinsic motivation is -0.017. It is possible that our experiments provide a weak test of the intrinsic motivation hypothesis given the incentive treatments had very little direct effect, though the vast majority of the intrinsic motivation literature focuses on the use of incentives not their effectiveness.

V.C. Analysis of Subsamples

Table VI investigates treatment effects for subsamples—gender, race/ethnicity, previous year's test score, an income proxy, whethere a student is an English language learner, and, in Dallas only, whether a student took the English or Spanish test.[18] All

17. Fryer (2010) shows that patterns are similar if one estimates treatment effects on each statement independently.

18. To ensure balance on all subsamples, we have run our covariance balance test (identical to Table III) for each subsample. Across all subsamples, the sample

TABLE VI

MEAN EFFECT SIZES (INTENT-TO-TREAT ESTIMATES) ON ACHIEVEMENT BY SUBSAMPLES

		Dallas	NYC		Chicago
Outcome	Subsample	2nd	4th	7th	9th
A. Language proxy					
Reading	English language learner	−0.164	−0.072	0.025	—
		(0.095)	(0.061)	(0.037)	
		1,915	971	1,164	
	Non-English language learner	0.221	−0.014	0.002	—
		(0.068)	(0.036)	(0.017)	
		1,689	5,417	7,872	
Math	English language learner	−0.008	−0.040	−0.054	—
		(0.108)	(0.061)	(0.054)	
		1,963	1,029	1,259	
	Non-English language learner	0.163	0.086	−0.028	—
		(0.098)	(0.052)	(0.039)	
		1,745	5,416	7,883	
Reading	Took Spanish test previous year	−0.118	—	—	—
		(0.104)			
		1,620			
	Took English test previous year	0.173	—	—	—
		(0.069)			
		1,618			
Math	Took Spanish test previous year	0.076	—	—	—
		(0.114)			
		1,656			
	Took English test previous year	0.071	—	—	—
		(0.100)			
		1,674			

TABLE VI
(CONTINUED)

B. Race					
Reading	White	—	−0.115	0.082	−0.112
			(0.126)	(0.057)	(0.054)
			232	701	361
	Black	0.143	−0.003	−0.030	0.010
		(0.073)	(0.048)	(0.022)	(0.032)
		778	2,920	3,775	4,171
	Hispanic	−0.031	−0.020	0.013	−0.009
		(0.084)	(0.034)	(0.023)	(0.035)
		2,742	2,788	3,664	2,943
	Asian	—	−0.066	0.055	0.178
			(0.056)	(0.063)	(0.138)
			501	964	136
Math	White	—	−0.262	0.084	−0.065
			(0.157)	(0.081)	(0.094)
			236	701	361
	Black	0.128	0.116	−0.058	0.006
		(0.128)	(0.064)	(0.037)	(0.026)
		818	2,920	3,760	4,155
	Hispanic	0.049	0.009	−0.069	−0.007
		(0.097)	(0.045)	(0.043)	(0.035)
		2,802	2,798	3718	2,942
	Asian	—	0.192	0.143	−0.138
			(0.067)	(0.073)	(0.131)
			510	998	136

TABLE VI
(CONTINUED)

C. Income proxy					
Reading	Free lunch	−0.042	−0.013	−0.013	−0.005
		(0.075)	(0.034)	(0.021)	(0.028)
		2,087	4,597	6,367	6,975
	No free lunch	0.101	−0.114	0.138	0.022
		(0.073)	(0.077)	(0.055)	(0.065)
		1,517	502	800	569
Math	Free lunch	0.042	0.036	−0.050	−0.007
		(0.091)	(0.047)	(0.039)	(0.024)
		2,141	4,630	6,427	6,961
	No free lunch	0.134	0.124	0.054	−0.064
		(0.095)	(0.106)	(0.055)	(0.044)
		1,567	501	805	567
D. Gender					
Reading	Male	0.042	−0.016	0.013	0.015
		(0.079)	(0.035)	(0.022)	(0.029)
		1,853	3,308	4,601	3,629
	Female	−0.022	−0.043	−0.004	−0.027
		(0.067)	(0.038)	(0.020)	(0.032)
		1,755	3,182	4,541	3,987
Math	Male	0.088	0.060	−0.019	−0.021
		(0.089)	(0.050)	(0.040)	(0.028)
		1,916	3,330	4,638	3,629
	Female	0.064	0.061	−0.041	−0.001
		(0.090)	(0.050)	(0.039)	(0.027)
		1,796	3,183	4,577	3,970

TABLE VI
(CONTINUED)

E. Pretreatment test score					
Reading	Below median previous year score	0.036	0.010	−0.034	−0.005
		(0.074)	(0.046)	(0.019)	(0.027)
		1,577	2,907	4,170	3,605
	Above median previous year score	−0.005	−0.049	0.034	−0.003
		(0.083)	(0.038)	(0.027)	(0.034)
		1,572	3,229	4,455	3,312
Math	Below median previous year score	0.080	0.082	−0.067	0.005
		(0.089)	(0.046)	(0.039)	(0.026)
		1,388	3,001	4,126	3,571
	Above median previous year score	0.003	0.072	0.007	−0.007
		(0.101)	(0.057)	(0.046)	(0.032)
		1,307	3,217	4,504	3,367

Notes. The dependent variable is the outcome indicated in the first column. Reading and math test score outcomes have been normalized to have a mean of 0 and a standard deviation of 1 within each grade across the entire sample of students in the school district. Thus, coefficients for these outcomes are in standard deviation units. The effect size is the difference between mean achievement of students belonging to the subsample indicated in the second column in schools randomly chosen to participate and mean achievement of these students in schools that were not chosen. It is the intent-to-treat (ITT) estimate on achievement. The third column presents results for second graders who participated in the Earning by Learning experiment in Dallas. The fourth and fifth columns present results for fourth and seventh graders, respectively, who participated in the Spark experiment in New York City. The sixth column presents results for ninth graders who participated in the Paper Project experiment in Chicago. All regressions include controls for reading and math test scores from the previous 2 years and their squares, race, gender, free/reduced lunch eligibility, English language learner status, the percent of black students in the school, the percent of Hispanic students in the school, and the percent of free/reduced lunch students in the school. For Dallas, regressions also include a control for whether the student took the English or Spanish version of the ITBS/Logramos test in the previous year. For Dallas and New York City, regressions also include an indicator for being in special education. For New York City, regressions also include controls for the number of recorded behavioral incidents a student had in the previous year, as well as the number of recorded behavioral incidents the school had in the previous year. All standard errors, located in parentheses, are clustered at the school level. The numbers of observations are located directly below the standard errors.

categories are mutually exclusive and collectively exhaustive. Standard errors are clustered at the school level.[19]

Gender is divided into two categories and race/ethnicity is divided into five categories: non-Hispanic white, non-Hispanic black, Hispanic, non-Hispanic Asian, and non-Hispanic other race.[20] We only include a racial/ethnic category in our analysis if there are at least 100 students from that racial/ethnic category in our experimental group. This restriction eliminates whites and Asians in Dallas and other race in all cities. Previous year's test scores are partitioned into two groups, divided at the median. Eligibility for free lunch is used as an income proxy.[21] English language learner is a dichotomous category. The final distinction, whether a student took an English or Spanish test, is only applicable in Dallas. Recall, students in regular classes take the ITBS and students in bilingual classes take the Logramos test. To ensure that treatment does not affect which testing group a student is assigned to, we use the language of the test taken in the year prior to treatment to define this subsample.

Table VI presents ITT estimates across various subsamples. The most informative partition of the data is by language proxies, depicted in Panel A. All other differences across subsamples are statistically insignificant. Students who take the ITBS test in Dallas score 0.173 (0.069) *above* the control group, whereas those who take the Logramos test score score 0.118 (0.104) *below* the control group. This important variation was masked in our combined estimates. Relatedly, students in Dallas who are English language learners, independent of the tests taken, have a treatment effect of -0.164σ (0.095). Non-ELL students increased their reading achievement by 0.221σ (0.068). In NYC, where we also have information on ELL status, we do not see a similar pattern.

is balanced between treatment and control. Results in tabular form are available from the author on request.

19. Fryer (2010) provides an analysis of subsamples where standard errors are adjusted for multiple hypothesis testing using both a Bonferroni correction and the free step-down resampling method detailed in Westfall and Young (1993) and Anderson (2008). These methods simply confirm our results.

20. The 63 students in NYC with missing gender information were not included in the gender subsample estimates. The 66 students in NYC with missing race/ethnicity information are not included in the race/ethnicity subsample estimates.

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

The effects in Dallas are both interesting and surprising and we do not have an air-tight answer. A potential explanation for these results is that providing incentives for reading predominantly English-language books has a negative impact on Spanish speakers by crowding out academic Spanish. Students in regular classes (who took ITBS) read approximately 9.9 books, 9.5 of which were in English. Students in the bilingual class (who took Logramos) read 15.3 books, 6.4 in English and 8.9 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 shown in Fryer (2010), 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 90% 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.

VI. Discussion and Speculation

Our field experiments have generated a rich set of facts. Paying second-grade students to read books significantly increases reading achievement for students who take the English tests or those who are not English language learners and is detrimental to non-English speakers. All other incentive schemes tested in this article had, at best, small to modest effects—none of which were statistically significant.

In this section, we take the point estimates literally and provide a (necessarily) speculative discussion around what broad lessons, if any, can be learned from our set of experiments. Much of the evidence for our discussion relies on cross-city comparisons

of treatment effects, which is problematic. We consider this to be a speculative discussion that may help shape future experimental work.

An obvious interpretation of our results is that all the estimates are essentially 0 and the effects on English speakers in Dallas were observed by chance alone. Yet the size of the results and the consistency with past research on the importance of reading books cast doubt on this as an explanation (Kim 2007; Allington et al. 2010). A second interpretation is that the only meaningful effects stem from English speakers in the Dallas experiment and this is likely due to younger children in the experiment. The lack of results from students of similar ages in Bettinger (2010) and our results from NYC provide evidence against this hypothesis.[22] A broader and more speculative interpretation of the results is that incentives are not a panacea and are more effective if tailored to appropriate inputs to the educational production function.

In what follows, we expand on the latter interpretation and discuss four theories that may explain why incentives for reading books (e.g., inputs) were more effective (for English-speaking students) than our other output-based incentives.

VI.A. 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.[23] If students only have a vague idea of how to increase output, then there may be little incentive to increase effort.[24] In Dallas, students were not required to know how to increase their

22. One might worry that the marginal value of an increase in achievement is not similar across treatments and that might explain the results. To investigate this, we estimated the amount of money a student would earn across treatments for a 0.25-standard-deviation increase in achievement. In Dallas, a 0.25 increase in achievement was associated with earning $13.81. In NYC, a 0.25-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. Thus, the only incentive scheme that produced remotely positive results also had the lowest return on achievement.

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

24. This hypothesis is consistent with the positive results from interventions in which disadvantaged youth are given information on the returns to education (see, for instance, Jensen 2010).

test scores; they only needed to know how to read books. 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, 7 full-time qualitative researchers in New York observed 12 students and their families, as well as 10 classrooms. From detailed interview notes, we gather that students were uniformly excited about the incentives and the prospect of earning money for school performance. 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.

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 10 exams administered in New York, our qualitative team asked students how they felt about the rewards and what they could do to earn more money on the next test. Every student found the question about how to increase his or her scores difficult to answer. Students answering this question discussed test-taking strategies rather than salient inputs into the education production function or improving their general understanding of a subject area.[25] 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.

VI.B. Model 2: Self-Control Problems

Another model consistent with the data is that students know the production function, but either have self-control problems or are sufficiently myopic that they cannot make themselves do

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

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 any time 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.

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.

VI.C. Model 3: Complementary Inputs

A third model that can explain our findings is that the educational production function has important complementarities that are out of the student's control. For instance, incentives may

need to be coupled with good teachers, an engaging curriculum, effective parents, or other inputs to produce output. In Dallas, students could read books independently and at their own pace. 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. 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.

Table VII shows the results of this exercise. For each subject test, the first row reports ITT estimates for the New York sample for all students in treatment and control whose teachers have valid value-added data. This subset comprises approximately 43% of the full sample. The results from this subset of students are similar to those for the full sample. The next two rows 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 0.046σ (0.077) and those with above-the-median quality teachers gain 0.135σ (0.069). The exact opposite pattern is observed for seventh-grade math. The jury is still out as to whether complementary inputs can explain our set of results.

TABLE VII

MEAN EFFECT SIZES (INTENT-TO-TREAT ESTIMATES) ON ACHIEVEMENT: BY TEACHER VALUE-ADDED

		NYC	
Outcome	Subsample	4th	7th
Reading	Nonmissing TVA	−0.015	−0.042
		(0.044)	(0.023)
		3,355	3,404
	Below median TVA	0.010	−0.048
		(0.055)	(0.031)
		1,680	1,793
	Above median TVA	−0.046	−0.072
		(0.069)	(0.029)
		1,675	1,611
Math	Nonmissing TVA	0.095	−0.077
		(0.066)	(0.060)
		3,263	4,237
	Below median TVA	0.046	−0.080
		(0.077)	(0.054)
		1,644	2,133
	Above median TVA	0.135	−0.103
		(0.069)	(0.073)
		1,619	2,104

Notes. The dependent variable is the state assessment taken in New York for the subject indicated in the first column. Outcomes have been normalized to have a mean of 0 and a standard deviation of 1 within each grade across the entire sample of students in the school district. Thus, coefficients are in standard deviation units. The effect size is the difference between mean achievement of students belonging to the subsample indicated in the second column in schools randomly chosen to participate and mean achievement of these students in schools that were not chosen. It is the intent-to-treat (ITT) estimate on achievement. The subsamples are defined according to the Teacher Value-Added score that a student's teacher achieved in the year prior to the experiment. Teacher Value-Added was calculated for New York by the Battelle Institute (http://www.battelleforkids.org). The third and fourth columns present results for fourth and seventh graders, respectively. All regressions include controls for reading and math test scores from the previous 2 years and their squares, race, gender, free/reduced lunch eligibility, English language learner status, the percent of black students in the school, the percent of Hispanic students in the school, the percent of free/reduced lunch students in the school, an indicator for being in special education, the number of recorded behavioral incidents a student had in the previous year, and the number of recorded behavioral incidents the school had in the previous year. All standard errors, located in parentheses, are clustered at the school level. The numbers of observations are located directly below the standard errors.

VI.D. Model 4: Unpredictability of Outputs

In many cases, 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) course grades, and (3) 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.

Further, it is important to remember that our incentive tests in New York were adaptive tests. These exams can quickly move students outside their comfort zone and into material that was not covered in class—especially if they are answering questions correctly. The qualitative team noted several instances in which students complained to their teachers when they were taken aback by questions asked on the exams or surprised by their test results. To these students—and 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. If students do not know how noisy tests are or what influences them, the model is equivalent to Model 1.

Deciphering which model is most responsible for our set of facts is beyond the scope of this article. One or several combinations of the models may ultimately be the correct framework. Future experimentation and modeling is needed.

HARVARD UNIVERSITY
NATIONAL BUREAU OF ECONOMIC RESEARCH

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at QJE online (qje.oxfordjournals.org).

REFERENCES

Allington, Richard, Anne McGill-Franzen, Gregory Camilli, Lunetta Williams, Jennifer Graff, Jacqueline Zeig, Courtney Zmach, and Rhonda Nowak. "Addressing Summer Reading Setback among Economically Disadvantaged Elementary Students." *Reading Psychology*, 31 (2010), 411–427.

Anderson, Michael. "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 (2008), 1481–1495.
Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review*, 92 (2002), 1535–1558.
Angrist, Joshua, Eric Bettinger, and Michael Kremer. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review*, 96 (2006), 847–862.
Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics*, 1 (2009), 136–163.
Angrist, Joshua, and Victor Lavy. "The Effect of High-Stakes High School Achievement Awards: Evidence from a Group-Randomized Trial." *American Economic Review*, 99 (2009), 1384–1414.
Barrera-Osorio, Felipe, Marianne Bertrand, Leigh Linden, and Francisco Perez-Calle. "Improving the Design of Conditional Transfer Programs: Evidence from Randomized Education Experiment in Colombia" *American Economic Journal: Applied Economics*, 3 (2011), 167–195.
Behrman, Jere, Piyali Sengupta, and Petra Todd. "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico." *Economic Development and Cultural Change*, 54 (2005), 237–275.
Bettinger, Eric. "How Financial Aid Affects Persistence." in *College Choices: The Economics of Where to Go, When to Go, and How to Pay For It*, ed. Caroline M. Hoxby (Chicago: University of Chicago Press, 2004).
——. "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores," NBER Working Paper w16333, 2010.
Bruhn, Miriam, and David McKenzie. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics*, 1 (2009), 200–232.
Bullock, Jonathon. "Effects of the Accelerated Reader on Reading Performance of Third, Fourth, and Fifth-Grade Students in One Western Oregon Elementary School." PhD Dissertation, University of Oregon, 2005.
Cameron, Judy, and W. David Pierce. "Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis." *Review of Educational Research*, 64 (1994), 363–423.
Card, David. "The Causal Effect of Education on Earnings." in *Handbook of Labor Economics Vol. 3*, ed. David Card and Orley Ashenfelter (Amsterdam: North Holland, 1999).
Cornwell, Christopher, David Mustard, and Deepa Sridhar. "The Enrollment Effects of Merit-Based Financial Aid: Evidence from Georgia's HOPE Program." *Journal of Labor Economics*, 24 (2006), 761–786.
Deci, Edward. "The Effects of Contingent and Noncontingent Rewards and Controls on Intrinsic Motivation." *Organizational Behavior and Human Performance*, 8 (1972), 217–229.
——. *Intrinsic Motivation* (New York: Plenum, 1975).
Deci, Edward, Haleh Eghrari, Brian Patrick, and Dean Leone. "Facilitating Internalization: The Self-Determination Theory Perspective." *Journal of Personality*, 62 (1994), 119–142.
Dynarski, Susan. "Building the Stock of College-Educated Labor." *Journal of Human Resources*, 43 (2008), 576–610.
Fryer, Roland G. Jr. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." NBER Working Paper w15898, 2010.
Gneezy, Uri, and Aldo Rustichini. "Pay Enough or Don't Pay at All." *Quarterly Journal of Economics*, 115 (2000), 791–810.
Greevy, Robert, Bo Lu, Jeffrey Silber, and Paul Rosenbaum. "Optimal Multivariate Matching before Randomization." *Biostatistics*, 5 (2004), 263–275.
Hahn, Andrew, Tom Leavitt, and Paul Aaron. "Evaluation of the Quantum Opportunities Program: Did the Program Work? A Report on the Post Secondary

Outcomes and Cost-Effectiveness of the QOP Program (1989–1993)." Brandeis University, Heller School, Center for Human Resources, 1994.
Imbens, Guido, and Jeffrey Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, 47 (2009), 5–86.
Jackson, Clement. "A Stitch in Time: The Effects of a Novel Incentive-Based High-School Intervention on College Outcomes." NBER Working Paper w15722, 2010.
Jensen, Robert. "The (Perceived) Returns to Education and the Demand for Schooling," *Quarterly Journal of Economics*, 125 (2010), 515–548.
Kim, James. "The Effects of a Voluntary Summer Reading Intervention on Reading Activities and Reading Achievement." *Journal of Educational Psychology*, 99 (2007), 505–515.
Kohn, Alfie, *Punished by Rewards* (Boston: Houghton Mifflin, 1993).
——. "By All Available Means: Cameron and Pierce's Defense of Extrinsic Motivators." *Review of Educational Research*, 66 (1996), 1–4.
Kremer, Michael, Edward Miguel, and Rebecca Thornton. "Incentives to Learn." *Review of Economics and Statistics*, 91 (2009), 437–456.
Krueger, Alan. "Economic Considerations and Class Size." *Economic Journal*, 113 (2003), F34–F63.
Laibson, David. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics*, 112 (1997), 443–477.
Leuven, Edwin, Hessel Oosterbeek, and Bas van der Klaauw. "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment." *Journal of the European Economic Association*, 8 (2010), 1243–1265.
Mancilla-Martinez, Jeannette, and Nonie Lesaux. "The Gap between Spanish-speakers' Word Reading and Word Knowledge: A Longitudinal Study." Unpublished paper, Harvard University, 2010.
Marshall, Helen, and Lucille MacGruder. "Relations between Parent Money Education Practices and Children's Knowledge and Use of Money." *Child Development*, 31 (1960), 253–284.
Mischel, Walter, Yuichi Shoda, and Monica Rodriguez. "Delay of Gratification in Children." *Science*, 244 (1989), 933–938.
Neal, Derek, and William Johnson. "The Role of Premarket Factors in Black-White Wage Differentials." *Journal of Political Economy*, 104 (1996), 869–895.
Neal, Derek. "Why Has Black-White Skill Convergence Stopped?," in *Handbook of the Economics of Education Vol. 1*, ed. Erik Hanushek and Finis Welch (Amsterdam: North Holland, 2006).
OECD. Education at a Glance 2010: OECD Indicators. http://www.oecd.org/edu/eag2010.
Ross, Steven, John Nunnery, and Elizabeth Goldfeder. "A Randomized Experiment on the Effects of Accelerated Reader/Reading Renaissance in an Urban School District: Preliminary Evaluation Report." University of Memphis, Center for Research in Educational Policy, 2004.
Ryan, Richard. "Control and Information in the Intrapersonal Sphere: An Extension of Cognitive Evaluation Theory." *Journal of Personality and Social Psychology*, 63 (1982), 397–427.
Ryan, Richard, Richard Koestner, and Edward Deci. "Ego-Involved Persistence: When Free-Choice Behavior Is Not Intrinsically Motivated." *Motivation and Emotion*, 15 (1991), 185–205.
Scott-Clayton, Judith. "On Money and Motivation: A Quasi-Experimental Analysis of Financial Incentives for College Achievement." Working Paper, Harvard University, 2008.
Swanson, Christopher. "Cities in Crisis 2009: Closing the Graduation Gap." Editorial Projects in Education, 2009. http://www.americaspromise.org/Our-Work/Dropout-Prevention/Cities-in-Crisis.aspx.
Treasure, Tom, and Kenneth MacRae. "Minimisation: The Platinum Standard for Trials?" *British Medical Journal*, 317 (1998), 317–362.
Westfall, Peter, and S. Stanley Young. *Resampling-Based Multiple Testing* (New York: Wiley, 1993).