NBER WORKING PAPER SERIES

PAYING TO LEARN:
THE EFFECT OF FINANCIAL INCENTIVES ON ELEMENTARY SCHOOL TEST SCORES

Eric P. Bettinger

Working Paper 16333
http://www.nber.org/papers/w16333

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2010

The author thanks Bob Simpson and the Coshocton City Schools, especially Patty Cramer, Wade Lucas and David Hire, for help throughout the project. The author also thanks Jim Rebitzer, David Cooper, Michael Kremer, Bridget Long, Ted Miguel, Phil Oreopoulos, and seminar participants at Case Western Reserve University, University of British Columbia, University of Arizona, University of Pittsburgh, Stanford University, UCLA, Georgetown, and Ohio State University for helpful comments. All opinions and mistakes are my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores
Eric P. Bettinger
NBER Working Paper No. 16333
September 2010
JEL No. I2,I20,I21

ABSTRACT

Policymakers and academics are increasingly interested in applying financial incentives to individuals in education. This paper presents evidence from a pay for performance program taking place in Coshocton, Ohio. Since 2004, Coshocton has provided cash payments to students in grades three through six for successful completion of their standardized testing. Coshocton determined eligibility for the program using randomization, and using this randomization, this paper identifies the effects of the program on students' academic behavior. We find that math scores improved about 0.15 standard deviations but that reading, social science, and science test scores did not improve.

Eric P. Bettinger
Stanford School of Education
CERAS 522, 520 Galvez Mall
Stanford, CA 94305
and NBER
ebettinger@stanford.edu

I. Introduction

In recent years, economists, policymakers, and education researchers have become increasingly interested in the role that incentives play in education. In the United States, for example, *No Child Left Behind* and other recent educational reforms have changed the incentives surrounding state-mandated test scores by creating penalties if students' academic performance is not improving or if it does not meet a certain level. Outside the United States, parents and children can receive cash rewards for school attendance and regular check-ups (e.g. Progresa in Mexico), for student test scores (e.g. recent experiments in Kenya and Israel), or for college attendance (e.g. England's EMA program).

In Fall 2004, Coshocton City Schools (Coshocton, Ohio) developed a financial incentive program focused on improving students' academic performance in primary school. With the financial support of a local foundation, Coshocton began making cash payments to students for successful completion of their standardized testing. Students in third, fourth, fifth, and sixth grade who passed district and state-mandated standardized exams became eligible for these rewards.

The rationale for these cash payments is straightforward. While academic achievement in early grades can greatly improve students' long-run success – facilitating college attendance and greater job opportunities (e.g. Schweinhart and Weikart 2002), these long-run benefits are intangible to many young children. Few third, fourth, fifth, or sixth graders actively think about college attendance or employment. Additionally, children are inherently impatient, and many studies in education, psychology, and economics document how children are often more motivated by short-run rewards than less tangible long-run rewards (e.g. Chelonis, Flake, Baldwin, Blake and Paule 2004, Harbaugh and Krause 1998, Bettinger and Slonim 2007).

The Coshocton experiment makes a unique contribution to previous research in economics, education, and psychology. Recent studies in economics (e.g. Angrist and Lavy 2009, Angrist, Lang, and Oreopoulos 2007, Kremer, Miguel and Thornton 2008, Leuven, Oosterbeek, and van der Klaauw 2010, Fryer 2010) have largely focused on the effects of external incentives on students' academic achievement among students in secondary and post-secondary schools. Coshocton is the first study in economics to focus on financial incentives for student achievement in primary schools. The Coshocton experiment also builds on prior literature in psychology. Psychologists have been particularly active in the development of

theory and evidence on the role of financial incentive programs. Most of the early literature on "token economies" and incentives focused on the effects of external or extrinsic motivators on individual's intrinsic motivation (e.g. Ayllon & Azrin 1968, Kazdin 1975a & 1975b, Kazdin & Bootzin 1972, Deci 1975, Lepper, Greene, & Nisbett 1973). In Coshocton, we were able to gather data on intrinsic motivation among students and can measure the impact of the program on intrinsic motivation. In our discussion of the results, we attempt to reconcile the recent findings in economics with previous literature from psychology.

One of the most important features of the Coshocton incentive program is the fact that the district assigned eligibility for the program using randomization. This was a condition mandated by the sponsoring foundation. The unit of randomization is a grade-school level (i.e. grade i at school j), and Coshocton city schools conducted lotteries at the beginning of the 2004-05, 2005-06, and 2006-07 school years. In each year, half of all students in grades three through six became eligible for the financial incentive which can be as much as $100 per student.

Coshocton may be an ideal place to study financial incentives for several reasons. First, Coshocton is a disadvantaged, poor community at the foot of Appalachia. It may be a perfect place to measure the extent and potential of financial incentives to improve academic achievement among disadvantaged students. Second, Coshocton was nearly in a state of academic emergency in 1999. District leaders were willing to try non-traditional ways of improving student achievement, and this willingness set the stage for the program's adoption. Finally, Coshocton's small yet intimate community afforded us unique access to students, teachers, and parents. As a result, not only can this research show quantitative evidence on the overall effect of the incentive program, but it may also yield insights into specific mechanisms by which the incentive program affected students.

The primary focus of this paper is on measuring the effects of Coshocton's program on student achievement. We find that math scores improved about 0.15 standard deviations higher for students who were eligible for the program relative to the control group. This effect occurs throughout the distribution of math test scores. In contrast to math, the estimated effects on reading, social science, and science test scores are both small and imprecise. We find that students' intrinsic motivation is not significantly lower as a result of participating in the program. In the final section of the paper, we present some evidence on the cost-effectiveness of Coshocton's program. In short, the Coshocton incentive program was a highly cost-effective

way of boosting test scores proving much cheaper at generating effects in math than other interventions.

II. Background

Previous Research on Financial Incentives and Student Achievement

School administrators and parents have long believed that students are motivated by short-run stimuli. As early as 1820, New York City introduced a system of financial rewards for students who performed well at school (Ravitch 1974). There are also many anecdotes of teachers or principals offering pizza parties, visits to museums, and other forms of entertainment to students who pass standardized exams. Additionally, many parents offer their children cash or other rewards for good grades,[1] and in the last decade, policymakers throughout the world have experimented with programs that pay students for academic performance. For example, New York City, Washington, D.C., Chicago, and Dallas are currently experimenting with providing cash rewards based on student performance (Fryer 2010).

Israel implemented two types of student incentive programs in 1999-2000. The first program provided cash payments to high school students who took high school completion exams. Students were paid for taking the exam and for their performance on the exam. Students were chosen to participate through a lottery. The second incentive program randomly chose high schools and provided cash incentives to students within those schools who took the high school completion exam. Angrist and Lavy (2009) find that cash incentives in Israel improved both the number of students taking high school completion exams and student test scores, particularly when the randomization involved entire schools.

There are other incentive programs throughout the world that focus on helping low-income families and children succeed in primary and secondary schooling. College students in the United Kingdom are eligible for money from the Educational Maintenance Allowance (EMA). The EMA provides cash payments – up to thirty pounds a week – to college students who remain in school. Following a successful pilot stage covering a third of the country (Deardon et al 2001), the EMA became available nationwide in September 2004. Similarly, a

[1] In a May 2007 survey of students in our sample, 65 percent reported that their parents were paying them money for their school performance. This percentage did not differ across treatment and control groups. Similarly 74 percent of students reported being paid for doing chores at home.

large Canadian university started a cash reward tied to student performance in 2005. Angrist, Lang, and Oreopoulos (2007) evaluate the program finding a small improvement, particularly among women, in their first semester grade point average but no effect after a year.

Kenya has also implemented a program rewarding attendance and test scores with cash payments. The program focused on female students in an effort to increase female participation in schooling. Evidence in Kremer, Miguel, and Thornton (2008) suggest that the effects were large and positive for girls. Additionally, poor families in Mexico, Colombia, Brazil and other countries receive cash payments if their children get regular check-ups and attend school. Though these programs were paired with incentive to improve health as well, the effect on education is well documented (e.g. Behrman, Sengupta, & Todd 2000).

Although many of the recent experiments were studied by economists, research on the effects of external incentive on students' outcomes has a long pedigree. Psychologists were particularly active in examining the role of incentives and token economies during the early 1970's. Many of the papers focused on the effects of external or extrinsic incentives on contemporaneous performance (e.g. Ayllon & Azrin 1968, Kazdin 1975a & 1975b, Kazdin & Bootzin 1972).[2] Other papers focused on the effects of external motivators on extrinsic and intrinsic motivation (e.g. Deci 1975, Lepper, Greene, & Nisbett 1973). The studies demonstrated that there are certain contexts in which external incentives can improve student outcomes (see reviews by Lepper & Greene 1978, Cameron & Pierce 2002). For example, when students lack intrinsic motivation, external rewards can improve outcomes such as academic achievement and subsequent intrinsic motivation (e.g. Lepper, Greene, & Nisbett 1973). By contrast, external rewards may reduce intrinsic motivation in students who already possess intrinsic motivation for learning a subject like math (e.g. Greene, Sternberg, & Lepper 1976) or art (e.g. Greene & Lepper 1974). Other research in psychology has found that the efficacy of external motivators depends on the type of behavior being incentivized and the type of reward, and the efficacy may even vary from student to student (e.g. Deci 1978, Csikszentmihalyi 1978). More recent work examined the implications and existence of extrinsic motivation crowding out intrinsic

[2] As Cameron and Pierce (2002) outline, these early studies differed in the actions that students had to do to receive the rewards, in the expectations that students had about their potential compensation, and in the populations studied. Many studies rewarded students for solving a number of puzzles, engaging in a specific activity, finishing a task, or students' absolute or relative performance on some assessment.

motivation in education and in economic contexts (e.g. Gneezy & Rustichini 2000, Cameron & Pierce 1994, Eisenberger & Cameron 1996, Frey 1994, Frey & Oberholzer-Gee 1997).

As we mentioned in the introduction, one of the aims of the paper is to reconcile some of the recent findings in economics with the established literature in psychology on the impacts of external incentives. To help do this, we gathered data on intrinsic motivation in students participating in the program using established metrics in psychology. Additionally, in extending the economics literature on financial incentives in education, we can not only measure the impact on primary school kids, but, similar to recent studies, we can also test whether the effects of the program differed by gender or generated spillover effects on non-participating students.

Coshocton Incentive Program

Coshocton is a poor, Appalachian community located in Eastern Ohio. The economically depressed community is characterized by high unemployment and low manufacturing and agricultural wages. According to the 2000 Census, the average income in Coshocton ($24,000) is significantly less than that of Ohio as a whole ($31,000). Coshocton is a predominantly white community (94 percent), and over 55 percent of students in the district qualify for free/reduced lunch. Additionally, as recent as 1999, Coshocton City Schools was performing so poorly that the state of Ohio was threatening to intercede and "take-over" the schools.

In November 2003, Robert Simpson, long-time resident of Coshocton and the owner of one of Coshocton's manufacturing plants, read an editorial in Forbes magazine about paying students for academic performance (Miguel 2003). The editorial highlighted results from the incentive program evaluated in Kremer, Miguel, and Thornton (2008). Robert Simpson subsequently contacted the Coshocton City Schools, and in Spring 2004, the Simpson Family Foundation offered a gift of $100,000 to the Coshocton City Schools to be used to establish a financial incentive program for Coshocton's elementary schools. Mr. Simpson specified that the district implement the program using randomization so that the district could rigorously evaluate the program to determine its overall effect.

The program aims at improving achievement for all students in five core subjects. Each year students in grades three through six take five different achievement tests in math, reading,

writing, science, and social studies.[3] Eligible students receive $15 for each test on which they score proficient or better. On any test for which a student scores in the "Advanced" or "Accelerated" designation under Ohio's state testing program, the student receives $20 instead of just $15.[4] Thus, an eligible student who scores proficient on all five tests would receive $75 and an eligible student who scores advanced on all five tests would receive $100. Even if a student passes just one test, he or she receives a financial reward. The relevant exams vary by grade depending on whether state-mandated proficiency and achievement exams are required.[5] The school district mails students' rewards in early June after the release of testing results. As a reference, over the life of the experiment about 61 percent of students typically scored above "Proficient" on the math tests; 33 percent scored above "Advanced;" and 16 percent scored "Accelerated."

As a condition for Mr. Simpson's donation, Coshocton City Schools had to agree to implement the program using randomization. The district-appointed advisory committee elected to randomize across grade levels within the district's elementary schools. The unit of randomization is the grade level at each school. In Coshocton, there are four elementary schools and four eligible grade levels (third through sixth grade) at each school. In each year, eight of these 16 eligible grade-school combinations would be selected via lottery. In each year, the randomization is repeated amongst the eligible grade-school combinations.

As an example, 3rd grade at Washington Elementary School and 6th grade at South Lawn Elementary School were among the grade-school combinations chosen in the first year as treatment schools. All students in all classes in that grade level at the respective schools were eligible for the incentive program in the 2004-05 school year. Fifth graders at South Lawn and 4th graders at Washington were not chosen in that same lottery, and so these grades at these schools are part of the control group in the first year. In September 2005, Coshocton City Schools conducted a second lottery in which eight new grade-school combinations were selected.

[3] One worry about a program like this is that students are not fully rational. Harbaugh and Krause (2001) show that students younger than eight consistently make decisions that appear irrational given prior decisions. As a result we focus on students in third grade (normally age 8) and above. Numerous studies (e.g. Bettinger and Slonim 2007) find high discount rates among children which would suggest that they overvalue small amounts of money but have less foresight for distant consumption.

[4] In 2005-06, the fifth and sixth grade students only took four exams (omitting writing). They were compensated $20 for proficient and $25 for more advanced designations.

[5] Appendix Table 2 outlines the specific tests taken in each subject by each grade in each year. We include controls in the regression specifications for the type of test (either Terra Nova or Ohio Proficiency). We do not find that treatment effects vary by test taken.

In the second year, all of the third and fourth grades at Washington and fifth and sixth grades at South Lawn had the same chances of being selected in the lottery during the second year. The Appendix provides a list of which grades at which schools were eligible for the incentive program in each of the three lotteries and across cohorts.[6]

The lottery was structured as follows. First the district randomly selected one grade per school. This ensured that each school would participate in the program. This stratification also helps ensure that control and treatment groups are balanced (see Angrist and Lavy 2009). After these four drawings (one per school), Coshocton conducted a fifth drawing in which they chose four additional grade-school combinations from amongst the remaining possibilities. This stratification does not impact the use of randomization as a means for identifying the effects so long as lotteries in subsequent years are similarly performed. We refer to those students who won the lottery as the "treatment" group because these students were eligible for the financial incentive. We refer to those students who participated in but lost the lottery as the "control" group. Because of the repeated nature of the lottery, a student could be in the treatment group in one or more years and similarly in the control group during other years. The lotteries were conducted at open school board meetings. In the second and third years, the district brought all of the eligible students to the school board meeting and besides conducting the lottery held a pep assembly for academics featuring the high school marching band and cheerleaders.[7] We used a bingo-cage and ping-pong balls (one for each grade-school combination) to conduct the drawing

[6] The advisory committee decided on this level of randomization for a number of reasons. First, randomization at the school level was impractical given the number of schools in Coshocton (4) and Mr. Simpson's desire to keep the money in Coshocton. Second, Coshocton did not want to randomize at the student or class level. Teachers did not want to have some students in a particular class participating in the program and others not participating as it would make it difficult (and perhaps psychologically damaging) to use it as a motivational tool. Additionally, principals did not want classrooms within grades at the same school to be the unit of randomization. Principals in Coshocton did not want a competitive environment across classrooms within the same grade and were worried that the randomization could end up pitting classrooms within the same grade and the same school against each other. Also, many teachers in a given grade at each of the schools have collaborative teaching arrangements where one teacher teaches math to all students in the grade level at the school while another teacher teaches reading. In these team-teaching assignments, it might be difficult for teachers to remember which students are eligible. As noted in Angrist and Lavy (2007), randomizing over grades within schools is similar to the research design in group-randomized-trials often conducted across hospitals or communities. Group-randomized trials are attractive in places where randomization at the student or patient level is impractical.

[7] One worry about the public lottery was that students would be disappointed if they lost. If the loss of the lottery discouraged students from trying, then treatment effects could be because of negative effects on the control rather than positive effects on the treatment. Part of the motivation for conducting the lottery early in the year was to allow time to pass so that students might forget any disappointment. Additionally, each year we surveyed teachers and asked them to report on a five-point scale whether students who lost the lottery were "disappointed" or whether they were "less willing to take tests." The average response was low, and teachers "somewhat disagreed" with these statements.

because it was more intuitive to students and community members than a random number generator.

Rather than pay students in cash, the advisory committee elected to pay students with "Coshocton Children's Bucks." The advisory committee was reluctant to give children cash since parents could easily take their children's cash and spend it on themselves rather than their children. As a result, Coshocton's Chamber of Commerce agreed to print children's gift certificates redeemable at any store in Coshocton. The gift certificates say "Children" on them and must be redeemed for children's items. Importantly, local retailers enforced this restriction. For example, cashiers at Walmart were instructed to ask the children and their parents if the chosen item was for the child. The use of Coshocton Bucks helped mitigate concerns of parental misconduct and provided some assurance that the incentive program would benefit the children directly.

III. Empirical Methodology

Empirical Specification

Because of randomization, simple t-tests or regression-based comparisons between the treatment and control groups can provide an unbiased estimate of the causal effect of the program (Angrist and Krueger 1999). We have data for all students between 1st and 6th grade starting in the 2002-03 school year and going through the 2006-07 school year. Our most simple regression model is a simple difference-in-differences type regression model

$$(1) \qquad y_{ijkt}=a+b*Treat_{jkt}+ grade_k + school_j + time_t+e_{ijkt}$$

where y_{ijkt} represents the outcome for student i at school j in the grade k at time t, $Treat_{jkt}$ is an indicator for whether the students at school j in grade k won the lottery and was hence eligible for the incentive program at time t. The variables $grade_k$, $school_j$, and $time_t$ are fixed effects controlling for just grade, school, and time. We can augment Equation 1 to include student covariates such as age, gender, race, free/reduced lunch status, and pre-program test scores. Since student can take tests in different years from different manufacturers, we also include dummy variables for the manufacturer of the test (Terra Nova or Ohio Department of Education). Finally, e_{ijkt} is an individual specific standard error per year. We can interpret the coefficient b as the effect of the incentive program. We can focus the sample only on the three

years in which the program was available or we can extend the sample to include pre-program years.

In estimating our standard errors, we cluster them at the level of treatment. All of the students in a specific grade in a specific school were facing similar incentives and teachers in these grades used assignments and other motivational reminders of the incentive program. So in practice, the sample over three years may be as low as 48 – 24 "treatment" grades and 24 "control" grades. We correct our standard errors using the standard cluster correction.[8]

Our outcome of interest will be student test scores. Students generally take five tests – mathematics, reading, writing, science, and social science. In these multiple exams, students may take tests from different test manufacturers within the same year. To make these scores comparable, we normalized all of them according to the population mean and variance for the appropriate test. For example, the third grade reading test in 2004 was published by the Ohio Department of Education and administered to all students in the state of Ohio. We normalized Coshocton students test scores using the mean and variance for this test across the universe of students who took this test (i.e. the entire state of Ohio).[9] The empirical specifications include year and test manufacturer controls in case there are other systematic differences that normalizing does not account for.

We examine each score separately for each subject except writing. We exclude writing from the analysis for two reasons. First, in any given year, some grades took the Ohio test and some grades took the Terra Nova. The state-administered tests assign one of 8 possible values as the test score. The Terra-Nova assigns test scores over a 307 point range. Normalizing these test scores for comparison purposes was difficult. Second, in the 2006-2007 school year, the state stopped administering writing exams to fourth graders, and Coshocton chose not to adopt a separate exam for this subject. Fourth graders in this year were offered $20 per subject for passing and $25 for an advanced distinction.

We also report various interactions between treatment status and individual characteristics. We pay particular attention to interactions between gender and treatment status and interactions which may result in programmatic spillovers. These interactions have proven to

[8] Forty-eight clusters is slightly above the threshold where we would need to worry about cluster corrections in the face of a small number of clusters (e.g. Angrist and Lavy 2005, Donald and Lang 2007, and Wooldridge 2003). Our results do not change when we correct our clustered standard errors in the way prescribed by Wooldridge (2003).
[9] In every case, we use the scale scores as the primary test score.

be important in other incentive literature. We do not report interactions based on SES as we find no difference in treatment effects associated with free/reduced lunch participation.

Verifying the Randomization

We first set out to determine whether the randomization yielded similar control and treatment groups. While the randomization was tightly controlled so that there were no violations, the small number of units of randomization (24 treatment and 24 control cohorts across the three years) may make it so that there could be small imbalances in the randomization.

Table 1 shows some basic regressions attempting to demonstrate that the randomization yielded comparable treatment groups. In each column of Table 1, we estimate Equation 1 using an individual characteristic as the dependent variable. For example, in Column 1 we regress an indicator that an individual was female against their treatment status including controls for grade, school, and year. We report the difference by treatment status in the likelihood of being female. The estimated effect is close to zero. Below the estimated difference, we report the standard error controlling for correlation within a specific grade at a specific school in a specific year (i.e. the unit of randomization). The difference is insignificant. In Column 2, we perform a similar analysis with students' ages with similar results. In Columns 3 and 4, we repeat the analysis for free/reduced lunch status and race. In each case, we find no difference between lottery winners and losers in this characteristic. The final column includes an indicator for whether or not the student wont the lottery in the previous year. In the last two years of the experiment, about 28 percent of students who won the lottery in one year, won the lottery in the next year. Given that 3rd graders could not have won the lottery in the prior year, the percent was higher (39 percent) among students who participated in the prior year lottery. The estimate suggests that winners were about 17 percentage points more likely to win the lottery if they participated in the prior year. This difference is not statistically significant although the standard error bands are large. Given the method of randomization was such that winning in one year was orthogonal to winning in the prior year, any deviation is the result of random imbalance over the small number of cells over which we are randomizing.

Another dimension to evaluate the randomization is to examine the test scores of students prior to the start of the program. In the 2003-2004 school year, the year prior to the program

beginning, almost all of the students in the school district were tested.[10] For each lottery, we assemble the data to include both lottery winners' and losers' pre-program test scores. We then stack the respective lotteries to test for balance across control and treatment across all lotteries.[11] As in the previous results, the lotteries look balanced. In Table 2, we estimate differences of 0.0647 standard deviations in math and 0.0607 standard deviations in reading without including covariates. Once we control for covariates, these differences fall to 0.0146 and 0.0210 standard deviations respectively.[12]

Finally, another way to see the balance in the lottery is to observe the distribution of the lottery winners across grades and schools. Over the three years, each school was guaranteed at least one winner per year because of the stratification of the lottery. We would expect that the other winners would be equally distributed across schools. In the end, the 24 winning grade-school-year combinations were distributed as follows: two schools had six winners each; one had seven; and one had five. Across grades, the distribution included the following: 8 winners from third grade, 5 winners from fourth grade, 3 winners from fifth grade, and 8 winners from 6th grade. The distribution is somewhat more skewed across grades than across schools, but given that differences across students and socioeconomic status is larger across schools than across grades, the unequal distribution across grades is not too troubling.

IV. Baseline Results

Mathematics

Table 3 shows the baseline results for math test scores. Each column in Table 3 is a separate estimate of Equation 1. The sample focuses exclusively on students who participated in the lottery. This includes students in third through sixth grade in the 2004-05 through 2006 -07 school years. The data are longitudinal so that a student could appear multiple times depending on their grade level. The sample size for math is slightly higher than in the previous tables in

[10] Students who were in 3rd grade in 2007 were in kindergarten at this time. They were the only students not tested in 2004. For this group, we use the 2006 test scores from their second grade year. This is the first time that they were administered tests. They were not eligible for the incentive program in 2006.
[11] Alternatively, we could run each lottery year individually. We do this without finding any statistically significant differences although with clustering we do not have statistical power in evaluating one lottery individually.
[12] The sample is larger for reading since 3rd graders take both the standardized exam in both fall and spring. We have included all previous reading tests. The results are similar if we focus only on the maximum score in the pre-program test scores.

that we have included students for which demographic data were missing. We have included indicator variables to note when demographic variables are missing.

When we just compare math scores for students eligible for the payments and for students who were not eligible, we find that eligible students' math scores were about 0.19 standard deviations higher. This is a significant difference. These baseline regressions include controls for grade, year, school, and test type (i.e. manufacturer). Given that at least half of the treatment group were chosen in school-year specific lotteries, we can also include school by year effects to control for systematic differences across schools in each year. When we also add school by year fixed effects, the estimated effect is about 0.14 standard deviations and remains significant. In Column 3, we include additional controls for age, gender, and race. With these additional controls, the difference stays roughly the same (0.18 standard deviations) and remains significant. When we add additional school by year fixed effects, the estimated effect is about 0.13 standard deviations and the estimate remains significant.

The results in Table 3 seem to suggest a significant, positive effect of the incentive program in math on math scores. In Table 4, we examine the effects of the program on students' passage rates which were the thresholds for which students were rewarded. The state of Ohio assigns students to one of five categories based on students' scale score in the respective grades. These five categories are from lowest to highest, deficient, basic, proficient, advanced, and accelerated. Students were paid $15 if they made it to proficient and an additional $5 if they scored advanced or accelerated. In Table 4, we show the estimated effects of the program on different test score measures based on the five-point categorization used by the state.

In the first column, we show the basic results using the five point distribution as the dependent variable. Here we find that students eligible for the awards score, on average, 0.2 levels higher than other students. In Columns 2 and 3, we present a linear probability model where the dependent variable is whether students scored basic or higher or proficient or higher respectively.[13] In these results, we get point estimates of 2-3 percentage points, and the estimated effect is not significant. The results suggest little impact on the proportion of students scoring over these margins. In Column 4, we repeat this analysis except we focus on students scoring advanced (level 4) or higher. Here we find significant results. Students who were eligible for the incentive program were 9.2 percentage points more likely to score above this

[13] In Columns 2-4, we find similar results when we estimate Probit models instead of linear probability models.

threshold. Given that about one-third of students score advanced or better, the estimated results suggest a sizeable increase in students' test scores. In Column 5 of Table 4, we repeat this exercise focusing on whether students scored accelerated (level 5). About 16 percent of students scored in this range, and the program increased the likelihood that students scored in this range by 5.2 percentage points.

The results in Table 4 suggest that the program was not effective in moving students over the proficient/non-proficient margin. The estimated effect is small and insignificant. By contrast, the program was quite successful in helping students move from scoring proficient to scoring advanced or accelerated.

Another way to verify theses results is to examine how the treatment effect varied with prior achievement. Assuming that the ranking of students' test scores is similar over time, interactions with pre-program achievement may show whether the estimated effect is strong at the top of the distribution. To capture the potential effect, we estimate equation 2:

$$(2) \qquad y_{ijkt}=a+\sum_{q=1}^{4} b_q * Treat_{jkt} * 1(Quartile=q)_{i(t=2004)} + \sum_{q=1}^{4} c_q * 1(Quartile=q)_{i(t=2004)} + grade_k + school_j + time_t + e_{ijkt}$$

where q indexes the quartile of achievement for students in pre-program test scores (i.e. 2004 test scores) and the *1(Quartile=q)* is a series of indicator variables for whether the student was in the specific quartile in 2004. We also include an additional category for students for whom there is no test score in 2004 (e.g. students moving in the district).

The estimates of equation 2 are reported in Table 5. For reference, we include the baseline specification with controls for the lagged score in the first column. The results in Columns 2 and 3 show the results by previous test score. We find significant positive effects for students who had previously been identified at the top of the test score distribution. We find small, positive, insignificant estimates for the students in the middle of the distribution. Interestingly, we find positive, significant effects for students at the bottom of the distribution. Given the results in Table 4, the fact that the bottom of the distribution improves suggests that students in that group are improving their test scores but not enough to make a significant change in the proportion of students scoring greater than the proficient threshold.

In sum, we find positive, significant effects on math test scores particularly for students at the top of the distribution. These effects served to move students over thresholds (advanced and accelerated) that are considered significant by the state of Ohio. This is similar to Leuven, Oosterbeek, and van der Klaauw (2010) which finds that high scoring students were more responsive to financial incentives. We find very little movement in the middle of the distribution and a positive but insignificant effect on the proportion of students scoring proficient. We also find positive effects at the bottom of the distribution but these effects did not seem to push students over the proficient threshold.

Reading

Table 6 shows the estimated effects on reading test scores. The specifications are identical to the previous table except they focus on differences in reading test scores. While all of the point estimates are positive, none of the estimated treatment effects are statistically significant. The standard errors are similar to the previous table, but the estimated treatment effects are much smaller ranging from 0.01 to 0.02 standard deviations. Additionally, although we do not report the estimates in the tables, we find no significant effects when we examine the how the program affected students at significant thresholds defining whether students are proficient, advanced, or accelerated.

Finding an effect in math but not reading is similar to findings in previous studies on educational interventions (e.g. Reardon, Cheadle, and Robinson 2008, Rouse 1998). Educational interventions often increase math scores with little to no impact on reading scores. Math particularly in the grades studied seems to be more elastic than other subjects.[14] Similarly, the early psychological research on extrinsic rewards found that extrinsic motivators were more effective as tasks were less conceptual in nature (e.g. Lepper and Greene 1978). Math is less conceptual than reading in early grades. Students can memorize a series of facts in math that can adequately prepare them for most tests. By contrast, it is much more difficult for students to prepare for a specific reading text.

[14] One theory as to why math is more elastic than reading focuses on parents' contributions to students' educations. In primary grades, much of students' reading experiences take place in the home while math instruction occurs largely in the schools. Some evidence of this is that the falloff in test scores between spring and fall are greater in math than in reading since students have less contact with math than reading over the course of the summer (Cooper, Nye, Charlton, Lindsay, and Greathouse 1996).

While Table 6 shows no significant change in reading test scores, we do detect some change in students' efforts in reading. Coshocton City Schools participates in a program called Accelerated Reader (AR). AR assigns point values to books (e.g. *Harry Potter and the Goblet of Fire* is worth 32 points). Point totals are assigned according to the difficulty, length, and importance of the book. For the students who were in 3rd or 4th grade during 2005, we can track their accelerated reader points for three years. The average amount of points was 62.7 with a standard deviation of 56.5. When we compare the point totals of winners and losers (using Equation 1), we find that lottery winners earned about 13 accelerated reader points more than students who were not eligible. While this may only be the equivalent of reading just one extra book per year, it does suggest increased effort in reading.

Fryer (2010) provides some evidence that students' achievement is more likely to increase when students are incentivized for inputs to educational production rather than outputs. For example, when students were incentivized to read books which pushed them to expand their vocabulary and reading comprehension, their achievement increased more than it did in other sites where students were compensated purely for their reading test scores. In Coshocton, students may not have realized that that additional reading may have improved test scores or they may not have been as aggressive as they were in other incentive programs where students were rewarded for pushing themselves to read more difficult books.

Alternative Subjects

Students were also tested in social science and science. Table 7 reports the estimated effects in each of these disciplines. Our specifications are identical to the baseline model.

The social studies results (Panel A) do not show any effect of the incentive program on test scores. We find positive effects around 0.05 standard deviations; however, the estimates are never significant. In science, the results (Panel B) are similar. We do not find significant estimates in any of the specifications and the point estimates are close to zero.

V. Relationship to Previous Research

Intrinsic versus Extrinsic Motivation

One of the most controversial aspects of the program was its potential impact on intrinsic motivation. Research in psychology has debated for over three decades on whether external incentive programs inhibit students' subsequent intrinsic motivation and performance (e.g. Lepper & Greene 1978, Cameron & Pierce 2002, Deci, Koestner & Ryan 2001). Deci et al (2001) make the claim that the consensus in psychology is that extrinsic rewards somehow inhibit students' subsequent intrinsic motivation. Cameron and Pierce (2002) argue that this conclusion is limited to specific payment schemes (e.g. rewards for participation versus rewards for absolute or relative achievement) and the nature of the reward (unexpected versus expected). In their meta-analysis of 145 studies, they find 11 studies in which participants were paid for exceeding a specific score on a task – similar to the Coshocton incentive program. Across those studies, they find no effect on intrinsic motivation as measured by observing students' subsequent choices, and they find an *increase* in intrinsic motivation coming from students' self-reported interest measures.

In May 2007, the school district attempted to gather data on the intrinsic motivation of students using two methods. First, 432 students completed the Academic Self-Regulation Questionnaire (SRQ-A) which measures students' intrinsic and extrinsic motivation for academic tasks.[15] Studies that have argued that incentives crowd out intrinsic motivation have used the SRQ-A to measure intrinsic motivation (Ryan and Connell 1989, Grolnick, Ryan, and Deci 1991, Miserandino 1996). Second, teachers rated on a five-point scale the degree to which students possessed an "internal desire to do well for the sake of doing well or learning" in math and in reading.

In the SRQ-A measure of intrinsic motivation, we find no significant difference across treatment and control groups. The mean measure is 2.48 with a standard deviation of 0.80 (min=1, max=4). The raw difference between treatment and control groups was -0.05 (s.e.=0.08) and the difference after controlling for school and grade effects was -0.08 (s.e.=0.09). Similarly, we find no statistically significant differences in measures of external regulation where the raw difference was 0.03 (s.e.=0.06) and the regression-adjusted difference was 0.09 (s.e.=0.07). The estimated differences are all small and not statistically significant.

Teachers' ratings of students presented similar results. The average rating for students' math was 3.20 with a standard deviation of 1.19 and the average rating for reading was 3.18 with

[15] Detailed descriptions are available at http://psych.rochester.edu/SDT/measures/selfreg_acad.html.

a standard deviation of 1.13. When we compare students who were eligible for the cash incentive versus non-eligible students, the raw difference in the math intrinsic motivation score was statistically significant suggesting greater levels of intrinsic motivation among the treatment (difference=0.24 with a standard error of 0.12), but this difference disappears once we control for school and grade (regression-adjusted difference=0.02 with s.e. = 0.13). In reading, we never find significant effects with the regression-adjusted difference being 0.005 (s.e.=0.13).

The direct measures of intrinsic motivation do not suggest any significant drop-off in students' interest as a result of the program. The estimated differences are small and not precisely estimated. Additional data might shed more light on the potential effects of the program on intrinsic motivation, but we find no measurable change in these behaviors between treatment and control groups in our study.

Another potential indicator of students' intrinsic motivation is their subsequent performance in the subject. For example, consider the experiment where the treatment includes two consecutive lotteries. There are four separate possibilities: students could win in both years, win in first but not second, win in second but not first, or lose in both periods. If students' intrinsic motivation declined, then students who were eligible in one year but not the next should experience a decline in test scores. Because of the multi-year nature of the Coshocton experiment, our data includes a number of these consecutive lottery experiments.[16] We caution that in the consecutive lottery experiment, parsing the data into small groups may compromise the randomization. In our large sample, we would expect balance, but given the low number of clusters in our study, estimates using the consecutive lottery experiment are noisy.[17]

To estimate the effects of the consecutive lottery experiment, we replace the treatment indicator in Equation 1 with indicator variables for when students were treated in the consecutive lottery experiment (i.e win-loss, loss-win, win-win). The comparison group includes students who never won. We report these results in Table 8. By extending our data to include seventh and eighth graders who ever participated in the experiment and by adding the 2007-2008 data, we increase the data by about 1200 student observations or by nearly 40 clusters. While we have

[16] We could even have an experiment of three consecutive lotteries, but this would divide the data into eight partitions with only a small sample of clusters in each partition. Our sample size is not sufficiently large to argue that these eight partitions would have been balanced by randomization.

[17] With the additional data from 2007-2008 and from expanding the sample to include 7th and 8th graders, we end up with 85 clusters divided among the four treatments.

more clusters, our statistical power is still limited given that we are separating the sample into four treatment groups instead of two.

Row 1 shows the test scores of students who won the lottery in the prior year but not this year. The estimates are very noisy, yet when we look at the point estimates, these students test scores are virtually identical to those of the control group. Any gains that they may have experienced from winning the lottery in the prior year do not persist into the next year. This is different from previous research by Leuven, Oosterbeek, and van der Klaauw (2010) which finds persistent effects once the incentives are no longer available. Additionally and more relevant for our discussion of intrinsic motivation, the results in Row 1 suggest that students' test scores do not fall below that of the control group. If students' intrinsic motivation for learning is lower after experiencing the incentives in the prior year, then we would have expected a decline in test scores.

Rows 2 and 3 report the effects of the intervention on students who won the lottery in the current year but either won or lost in the prior year. In both cases, the point estimate is similar, and in the case of students who had won the lottery for the first time, the coefficient is marginally significant.

In sum, we find no evidence that the incentive program eroded students' intrinsic motivation. This is true when we measure intrinsic motivation directly. It is also true when we use subsequent performance as a measure of intrinsic motivation.

Effects by Gender

We can also test whether there are significant differences between the responses of boys and girls to the incentive program. Previous studies (e.g. Angrist, Lang, and Oreopoulos 2007, Angrist and Lavy 2009) have found that the effects of incentives on females have been larger than those for males. To test this, we can also augment our basic specification by interacting gender with the treatment effect to detect whether there is a statistically significant difference between the treatment effects for boys and girls. These results appear in Table 9. In these estimates, the treatment effect for boys is between 0.12 and 0.17 standard deviations in math and negligible in reading. The coefficient on the interaction between females and the treatment shows the difference between the treatment effects for boys and girls. Here we always estimate a

positive difference although it is never significant. The standard errors are fairly generous on the interaction term so it is difficult to put bounds on what the difference in the treatment effects may be.

Spillover Effects

We can also test whether the incentive program had spillover effects within families. Previous research by Kremer, Miguel and Thornton (2008) shows a large spillover effect among boys in response to an incentive program focused on girls. In the Coshocton Incentive Program, about 14 percent of the control group had siblings who were eligible for the program. If the incentive program leads to greater effort for an eligible child, siblings may try harder as well. In focus groups with parents, some parents reported that they had provided the incentive program for their children who were not selected to be part of the incentive program in one year.

To test for spillovers, we augment our basic model by including an indicator variable for cases where students have siblings who are eligible but the student him/herself is not. The results of this exercise appear in Table 10. The treatment effects are nearly identical to those treatment effects reported in other tables. The effect of the incentive program in math is between 0.11 and 0.17 standard deviations. The effect of the incentive program in reading is not significant and the point estimate is small. If spillovers exist within families, we should see significant estimates for the sibling indicator. However, we fail to find any significant effect. The point estimates are always negative and the results are not statistically significant. As before, the standard errors are large enough that we cannot reject that there could be spillover effects of some magnitude, but we do not find any significant results in the Coshocton experiment.

VI. Discussion and Conclusion

The paper presents evidence from the Coshocton Incentive Program. The Coshocton Incentive Program offered students between grades three and six financial incentives to perform well on standardized tests. We identify the results of the program by taking advantage of the randomization of which grades at which schools were eligible for the cash award. The results are positive and significant in math but not in reading, social science, or science.

Was it really the incentives or was some other force at work? Because of the research design, we cannot identify whether the effects arise from teachers performing differently in years when their students were eligible or whether students were actively responding directly to the incentive. Annual teacher surveys suggest that teachers used different tools in years that their students were eligible. For example, a popular writing assignment focused on how students would spend "their" money. One teacher decorated the room with paper $100 bills, and a couple of teachers used the rallying cry "Show me the Money" to start math instruction. There were also no changes in teachers' use of other student incentive programs (e.g. pizza party, video game rewards) regardless of whether they were in the control group or the treatment group. While our research suggests that math scores improved in the program, over time teachers became less convinced of the program's efficacy. When asked to rate the program's efficacy on a five-point scale (5=best), teachers' average responses fell from 4.2 in 2005 to 3.8 in 2006 to 3.1 in 2007.

Another piece of evidence suggesting that teachers' behavior affected the outcome comes from a complimentary experiment conducted in the 2007-2008 school year. In the 2007-2008 school year, Coshocton offered a math and reading incentive of up to $25 to 7th and 8th grade students for the state-mandated math and reading results. In this case, however, Coshocton permitted us to randomize eligibility at the student level. The purpose was to see whether students would respond when teachers were not reminding them. Teachers did not know which students were eligible for the rewards in their classes. Among this group, we find that there is no treatment effect. After controlling for individual characteristics, lottery winners scored .070 *lower* than other students (with a standard error of .103). There is a lot of noise here, but the treatment effects reported in Table 3 are outside the 95 percent confidence interval surrounding the estimated effect on 7th and 8th graders. There are two possible explanations for this result. First these students are older and may respond differently to incentive programs than the younger students. Fifty dollars may be less valuable to older children who have a more understanding of the value of money. Second, these students had no reinforcement from the teachers which would support the hypothesis that the treatment effect worked through teachers.

Yet we also have some evidence that students may have increased their effort in response to the program. We asked teachers to rate on a five point scale whether their "students were more motivated to perform well." There were statistically significant differences suggesting that students in the treatment were more likely to be motivated. Additionally, the Coshocton City

Schools conducts special extra-curricular workshops to help students prepare for Spring test administrations. When asked to report whether students were willing "to participate in extra help," teachers whose students were eligible for the reward program agreed with this statement more than teachers whose students were not eligible for the reward.[18] One possibility is that students who attended these special after school sessions learned test-taking strategies; however, when we examine students' item responses, we find that the probability that a student left a question blank was identical across control and treatment groups. We find no statistically significant evidence that students in the treatment had better "test-taking" strategies.

Finally, we turn to the cost effectiveness of Coshocton's program. Coshocton's Incentive Program was highly cost effective relative to other educational interventions. Across the three years, Coshocton's program cost about $52,000, and math scores improved by about 0.15 standard deviations. The overall cost of Coshocton's program was similar to the average teacher salary in Coshocton which was $50,704 in 2007. Suppose instead of using the incentive program that Coshocton had hired an extra teacher to work 1/3 of the year for each of the three years of the experiment. If Coshocton had used the money to hire another teacher, the average class size in third to sixth grade would have only fallen from 19.4 in 2007 to 19.2.[19] By contrast, in Project STAR class size dropped from the around 24 to around 15, and the average test score gain in math and reading from small classes was 0.25. If the gain from class size is linear, then the reduction in class size that would have happened in Coshocton (0.2 students per class) would have generated a 0.006 standard deviation increase in math and reading test scores. The point estimate for the reading gain (.010) and the observed 0.15 standard deviation gain in math exceed any projected gain from class size.

Additionally, the overall expenditure on the program was only 0.15 percent of the district's overall instruction expenditure. The average instructional expenditure per student in 2006-2007 was $5,469. The average expenditure per student eligible for the treatment was about $53 – about one percent of the overall instructional expenditure, and in Coshocton's case, all of the money came from private donations. Hence, in summation, the Coshocton incentive program

[18] In May 2007, the district surveyed students about their study habits. There was no statistically significant difference in the number of hours students reported studying across treatment and control in both reading and math. Also, students in the treatment actually reported that they were less likely to participate in extra-curricular study sessions. This difference was statistically significant and in direct contrast to teachers' perceptions.
[19] Coshocton would have had to hire 7.6 new teachers to reduce class size to 15.

was a cost effective program which led to substantial math test score gains, especially for students at the bottom and top of the test score distribution.

References

Angrist, Joshua, Daniel Lang and Philip Oreopolous (2007) "Incentives and Services for College Achievement: Evidence from a Randomized Trial." IZA Discussion Paper Number 3134.

Angrist, Joshua and Victor Lavy (2009) "The Effects of High Stakes School Achievement Awards: Evidence from a Randomized Trial." *American Economic Review* 99(4): 301-331.

Angrist, Joshua D. and Alan B. Krueger, 1999. "Empirical strategies in labor economics", in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics, Volume 3*.

Ayllon T. and N. Azrin (1968) *The Token Economy*. Prentice-Hall: New Jersey.

Behrman, Jere R., P. Sengupta, and P. Todd, (2000) *Final Report: The Impact of PROGRESA on Achievement Test Scores in the First year*, International Food Policy Research Institute, Food Consumption and Nutrition Division.

Bettinger, Eric and Robert Slonim (2007), "Patience in Children: Evidence from Experimental Economics." *Journal of Public Economics 91*(1-2): 343-363.

Cameron, J., and W. D. Pierce (1994) ''Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis,'' *Review of Educational Research, 64*: 363–423.

Cameron, Judy and W. David Pierce (2002) *Rewards and Intrinsic Motivation: Resolving the Controversy*. Bergin & Garvey: London.

Chelonis, John J., Rebecca Flake, Ronald Baldwin, Donna Blake, Merle Paule, (2004) "Developmental aspects of timing behavior in children," *Neurotoxicology and Teratology*, 26 (3), pp. 461-476.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., and Greathouse, S. The Effects of Summer Vacation on Achievement Test Scores: A Narrative and Meta-analytic Review. *Review of Educational Research,* 66: 227-268, 1996.

Csikszentmihalyi, Mihaly (1978) "Intrinsic Rewards and Emergent Motivation," in *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation* edited by M. Lepper and D. Greene. LEA Press: New Jersey. pp. 205-216.

Dearden, Lorraine, C. Emmerson, C. Frayne, A. Goodman, H. Ichimura, and C. Meghir, (2001) *Education Maintenance Allowance: The First year, A Quantitative Evaluation*, Department for Education and Evaluation Research Report RR257.

Deci, Edward (1975) *Intrinsic Motivation* Plenum Publishing: New York.

Deci, Edward (1978) "Applications of Research on the Effects of Rewards," in *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation* edited by M. Lepper and D. Greene. LEA Press: New Jersey. pp. 193-203.

Deci, E., R. Koestner, and R. Ryan (1999) "A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation." *Psychological Bulletin, 22*: 627-668.

Donald, S. and K. Lang. (2007) "Inference with Difference-in-Differences and Other Panel Data" *Review of Economics and Statistics 89*(2): 221-233.

Eisenberger, R., & Cameron, J. (1996). "Detrimental effects of reward: Reality of myth?" *American Psychologist, 51*, 1153-1166.

Frey, B. S. (1994) ''How Intrinsic Motivation Is Crowded in and out,'' *Rationality and Society, 6:* 334–352.

Frey, B. S., and F. Oberholzer-Gee, (1997) ''The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-out,'' *American Economic Review, 87:* 746-755.

Fryer, Roland (2010) "Financial Incentives and Student Achievement: Evidence from Randomized Trials" NBER Working Paper Number 15898.

Gneezy, U., and A. Rustichini (2000) "Pay Enough or Don't Pay At All." *Quarterly Journal of Economics*, 791-810.

Greene, D., B. Sternberg, and M. Lepper (1976) "Overjustification in a token economy." *Journal of Personality and Social Psychology, 34*: 1219-1234.

Greene, D. and M. Lepper (1974) "Effects of extrinsic rewards on children's subsequent intrinsic interest." *Child Development, 34:* 1141-1145.

Grolnick, W. S., Ryan, R. M., & Deci, E. L. 1991. The inner resources for school performance: Motivational mediators of children's perceptions of their parents. *Journal of Educational Psychology*, *83*, 508-517.

Harbaugh, William, and Kate Krause, 1998. "Economic Experiments that you can Perform at Home on your Children," Working paper, University of Oregon.

Kazdin, A. (1975a) "Recent advances in token economy research" in *Progresss in behavior modification* edited by M. Hersen, R. Eisler, and P. Miller. Academic Press: New York.

Kazdin, A. (1975b) *Behavior modification in applied settings.* Dorsey Presss: Homewood, Illinois.

Kazdin, A. and R. Bootzin (1972) "The token economy: An evaluative review" *Journal of Applied Behavior and Analysis, 5*: 343-372.

Kremer, Michael, Edward Miguel and Rebecca Thornton (2008) "Incentives to Learn" forthcoming in *Review of Economics and Statistics.*

Lepper, Mark, David Greene, and R. Nisbett (1973) "Undermining children's intrinsic interest with extrinsic rewards: A test of the 'overjustification' hypothesis." *Journal of Personality and Social Psychology*, *28*:129-137.

Lepper, Mark R. and David Greene (1978) *The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation*. LEA Publishers: New Jersey.

Leuven, E., H. Oosterbeek and B. van der Klaauw (2010), "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment," *Journal of the European Economic Association.* Forthcoming.

Medina, Jennifer (2007, June 19), "Schools Plan to Pay Cash for Marks," *The New York Times.*

Miguel, Edward (2003) "Cash Talks" *Forbes Magazine* (11/24/2003).

Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology, 88*, 203-214.

Ravitch, D (1974) *The Great School Wars*. Basic Books: New York.

Reardon, Sean F. Jacob E. Cheadle, and Joseph P. Robinson (2008) "The effect of Catholic schooling on math and reading development in kindergarten through fifth grade." Stanford University mimeo.

Ryan, R. M., & Connell, J.P. (1989). "Perceived locus of causality and internalization: Examining reasons for acting in two domains." *Journal of Personality and Social Psychology, 57*: 749-761.

Schweinhart, Lawrence and David Weikart (2002) "The Perry Preschool Project: Significant Benefits" *Journal of At-Risk Issues 8:*5-8.

Wooldridge, Jeffrey (2003) "Cluster-Sample Methods in Applied Econometrics," *American Economic Review 93*:133-138.

Table 1. Differences Between Treatment and Control Groups in Pre-Lottery Characteristics.

	Female	Age (in days at test time)	Free/Reduced Lunch Participation	White	Won Lottery in Prior Year
	(1)	(2)	(3)	(4)	(5)
Treatment	0.013	-13.61	0.012	-0.0005	.172
	[0.026]	[8.37]	[0.018]	[0.0114]	[.112]
Grade, Year, School Controls	Yes	Yes	Yes	Yes	Yes
N	1527	1504	1527	1527	991
N (students)	893	887	893	893	680
N (grade-school combinations)	48	48	48	48	32

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned. Sample drops in Column 5 since we only report results for students in 2005-06 and 2006-2007 which are the only years in which a student may have won the lottery in the prior year.

Table 2. Differences Between Treatment and Control Groups in Pre-Program Test Scores

	Pre-Program Math Scores		Pre-Program Reading Scores	
	(1)	(2)	(3)	(4)
Treatment	.0647	.0146	.0607	.0210
	[.0741]	[.0706]	[.0511]	[.0547]
Grade, Year, School Controls	Yes	Yes	Yes	Yes
Controls for Age, Gender, and Race	No	Yes	No	Yes
N	1572	1572	2637	2637
N (students)	817	817	844	844
N (grade-school combinations)	48	48	48	48

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned.

Table 3. OLS Estimates of Effects of Pay to Learn on Math Test Scores

	Lottery Sample, 3rd-6th Grade from 2004-05 to 2006-2007			
	(1)	(2)	(3)	(4)
Treatment	0.1896 [0.0496]	.1400 [.0485]	0.1802 [0.0487]	0.1328 [0.0485]
Age			-0.0005 [0.0001]	-0.0005 [0.0001]
Female			-0.0428 [0.0426]	-0.0427 [0.0433]
Caucasian			-0.0407 [0.1048]	-0.0525 [0.1055]
Free/Reduced Lunch			-0.3220 [0.0583]	-0.3250 [0.0578]
Grade, Year, School Controls	Yes	Yes	Yes	Yes
School by Year Interactions	No	Yes	No	Yes
R-squared	.144	.155	.190	.201
N	1615	1615	1615	1615
N (students)	873	873	873	873
N (grade- school-year)	48	48	48	48

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned.

Table 4. Effect of Incentive Program on
Proficient Rates for Ohio Mathematics Achievement Test

	Lottery Sample, 3rd-6th Grade from 2004-05 to 2006-2007				
	Pass Level (1=min, 5=max) Mean=2.8 Stdev=1.4	Over Basic (lvl>=2) Mean=.72	Over Proficient (lvl>=3) Mean=.61	Over Advanced (lvl>=4) Mean=.33	Accelerated (lvl=5) Mean=.16
	(1)	(2)	(3)	(4)	(5)
Treatment	.203 [.074]	.021 [.013]	.038 [.028]	.092 [.026]	.052 [.024]
Age, Female, Race, and Socioeconomic Control	Yes	Yes	Yes	Yes	Yes
Grade, Year, School Controls	Yes	Yes	Yes	Yes	Yes
N	1248	1248	1248	1248	1248
N (students)	840	840	840	840	840
N (grade- school-year)	40	40	40	40	40

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned. The sample does not include third and fifth graders in 2005. These students took the Terra Nova exam rather than the Ohio Achievement Test in that year.

Table 5. Distributional Effects of Incentive Program on Math Achievement

	Lottery Sample, 3rd-6th Grade from 2004-05 to 2006-2007		
	(1)	(2)	(3)
Treatment	.136 [.047]		
Treatment*Lagged Score in Lower 25% of Population		.361 [.124]	.304 [.094]
Treatment* Lagged Score in 25-50%		.084 [.092]	.015 [.084]
Treatment* Lagged Score in 51-75%		.061 [.082]	.010 [.077]
Treatment* Lagged Score in top 25%		.236 [.089]	.218 [.178]
Treatment* Lagged Score Missing		.049 [.094]	.017 [.111]
Age, Female, Race, and Socioeconomic Control	Yes	Yes	Yes
Grade, Year, School Controls	Yes	Yes	Yes
School by Year Fixed Effects	No	No	Yes
Lagged Achievement Quartile Fixed Effects	Yes	Yes	Yes
N	1615	1615	1615
N (students)	873	873	873
N (grade- school-year)	48	48	48

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned.

Table 6. OLS Estimates of Effects of Pay to Learn on Reading Test Scores

	Lottery Sample, 3rd-6th Grade from 2004-05 to 2006-2007			
	(1)	(2)	(3)	(4)
Treatment	0.0222 [0.0468]	.0182 [.0489]	0.0095 [0.0425]	0.0103 [0.0454]
Age			-0.0004 [0.0001]	-0.0004 [0.0001]
Female			0.1076 [0.0343]	0.1085 [0.0343]
Caucasian			-0.0521 [0.0983]	-0.0436 [0.1009]
Free/Reduced Lunch			-0.3138 [0.0526]	-0.3121 [0.0527]
Grade, Year, School Controls	Yes	Yes	Yes	Yes
School by Year Interactions	No	Yes	No	Yes
N	2341	2341	2341	2341
N (students)	887	887	887	887
N (grade- school-year)	48	48	48	48

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned. The sample increases relative to Table 3 because third grade students take two exams per year. We have included both exams. Results do not change if we include on the spring exam or the highest exam score for each third grader.

Table 7. OLS Estimates of Effects of Pay to Learn on Other Test Scores

	Lottery Sample 3rd-6th Grade from 2004-05 to 2006-07		
	(1)	(2)	(3)
A. Social Science			
Treatment Effect	0.056	0.048	0.023
	[0.055]	[0.053]	[0.041]
Age, Gender, Race, FRL Controls	No	Yes	Yes
School by Year FE	No	No	Yes
N	1488	1488	1488
N (students)	866	866	866
N (grade- school-year)	48	48	48
B. Science			
Treatment Effect	0.011	0.003	-0.048
	[0.058]	[0.058]	[0.039]
Age, Gender, Race, FRL Controls	No	Yes	Yes
School by Year FE	No	No	Yes
N	1488	1488	1488
N (students)	866	866	866
N (grade- school-year)	48	48	48

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned.

Table 8. Math Scores in Year After Treatment

	All Students who Ever Participated in Lottery from 2004-05 to 2007-2008	
	(1)	(2)
Won Lottery Last Year but Not This Year	.042 [.060]	.078 [.062]
Won Lottery This Year But Not Last Year	.126 [.079]	.066 [.082]
Won Lottery This Year but Not Last Year	.251 [.118]	.236 [.115]
Age, Gender, Race, FRL Controls	Yes	Yes
Grade, Year, School Controls	Yes	Yes
School by Year Interactions	No	Yes
N	2846	2846
N (students)	1106	1106
N (grade- school-year)	85	85

Notes: Sample includes all students who participated in the lottery at any time. Previous tables excluded students once they were not eligible for the lottery.

Table 9. Estimated Treatment Effects by Gender

	Lottery Sample, 3rd-6th Grade from 2004-05 to 2006-2007			
	Math Test Scores		Reading Test Scores	
	(1)	(2)	(3)	(4)
Main Treatment	.168 [.066]	.115 [.067]	.007 [.061]	.008 [.063]
Treatment*Female	.025 [.085]	.036 [.086]	.006 [.073]	.005 [.073]
Age, Gender, Race, FRL Controls	Yes	Yes	Yes	Yes
Grade, Year, School Controls	Yes	Yes	Yes	Yes
School by Year Interactions	No	Yes	No	Yes
N	1615	1615	2341	2341
N (students)	873	873	887	887
N (grade- school-year)	48	48	48	48

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned.

Table 10. Estimated Spillover Effects

	Lottery Sample, 3rd-6th Grade from 2004-05 to 2006-2007			
	Math Test Scores		Reading Test Scores	
	(1)	(2)	(3)	(4)
Main Treatment	.165	.114	-.003	-.002
	[.050]	[.052]	[.036]	[.039]
Sibling was Eligible for Treatment	-.055	-.066	-.048	-.043
(but student was not)	[.054]	[.056]	[.091]	[.093]
Age, Gender, Race, FRL Controls	Yes	Yes	Yes	Yes
Grade, Year, School Controls	Yes	Yes	Yes	Yes
School by Year Interactions	No	Yes	No	Yes
N	1615	1615	2341	2341
N (students)	873	873	887	887
N (grade- school-year)	48	48	48	48

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned.

Appendix Table 1. Coshocton Incentive Winners

2004-2005 School Year

Washington	3rd, 4th, 6th
Central	3rd, 6th
South Lawn	3rd, 6th
Lincoln	3rd

2005-06 School Year

Washington	5th
Central	3rd, 5th
South Lawn	3rd, 4th, 6th
Lincoln	4th, 6th

2006-2007 School Year

Washington	3rd, 4th, 6th
Central	4th, 6th
South Lawn	5th
Lincoln	3rd, 6th

Treatment Years By School Cohort:

School	Grade in 2004-05	Years Won Lottery
Washington	1	2007
	2	2007
	3	2005
	4	2005, 2006, 2007
	5	
	6	2005
Central	1	
	2	2006, 2007
	3	2005
	4	2006, 2007
	5	
	6	2005
South Lawn	1	
	2	2006
	3	2005, 2006, 2007
	4	
	5	2006
	6	2005
Lincoln	1	2007
	2	
	3	2005, 2006
	4	2007
	5	2006
	6	

Appendix Table 2. Incentivized Tests by Grade Year

	Math	Reading	Science	Social Science	Writing
2004-05					
Grade 3	Terra Nova (TN)	Ohio Achievement (OAT)	TN	TN	TN
Grade 4	OAT	OAT	OAT	OAT	OAT
Grade 5	TN	TN	TN	TN	TN
Grade 6	OAT	OAT	OAT	OAT	OAT
2005-06					
Grade 3	OAT	OAT	TN	TN	TN
Grade 4	OAT	OAT	TN	TN	OAT
Grade 5	OAT	OAT	TN	TN	--
Grade 6	OAT	OAT	TN	TN	--
2006-07					
Grade 3	OAT	OAT	TN	TN	TN
Grade 4	OAT	OAT	TN	TN	OAT
Grade 5	OAT	OAT	OAT	OAT	--
Grade 6	OAT	OAT	TN	TN	TN

Notes: Test manufacturer for each test in each subject administered in each year of the incentive program.