Paying to Learn:
The Effect of Financial Incentives on Elementary School Test Scores

Eric P. Bettinger
Case Western Reserve and NBER

April 28, 2008

Abstract:
In recent years, policymakers and academics have become increasingly interested in applying financial incentives to individual decision-making in education. This paper presents evidence from a pay for performance program which took place in Coshocton, Ohio. Since 2004, the Coshocton City Schools has provided 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 are eligible for these rewards. Coshocton determined eligibility for the program using randomization. By exploiting the randomization of Coshocton's incentive program, this paper attempts to identify the effects of student incentive programs on students' academic behavior. Additionally, the structure of Coshocton's program creates some "kinks" in student incentives. These kinks reveal places where students who were eligible for the incentive program no longer had any incentive for subsequent "high stakes" work. The results present evidence that young children respond to incentives. Math scores improved about 0.15 standard deviations higher for elementary school students who were eligible for the program relative to the control group. We find little evidence that reading, social science, and science test scores changed in response to the incentive program. However, students' behavior at the specific discontinuities in the cash incentive program suggests that students respond to incentives even in ways which may not be desirable to educators.

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.

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. In developing countries, 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,[1] 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 2007, Angrist, Lang, and Oreopoulos 2007, Kremer, Miguel and Thornton 2008, Leuven, Oosterbeek, and van der Klaauw 2003) 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

[1] The funding for these cash payments comes from a directed grant from the Simpson Family Foundation.

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. Unlike other incentive programs, Coshocton conducted separate lotteries in consecutive years, so many students were eligible one year for the program but not the next. As a result, we can identify both the contemporaneous effects of the incentive program and the "year after" effects when the incentive program is no longer available. Additionally, the structure of Coshocton's program creates some "kinks" in student incentives. These kinks occur when students who were eligible for the incentive program no longer had any incentive for subsequent "high stakes" work. This feature allows us to identify ways in which subtle features of the incentive design actually reduced academic effort.

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. While the contemporaneous effects were positive in the first year, we find little evidence that the effect persists in subsequent years when students are no longer eligible for the cash incentive. We also find that students' intrinsic motivation is not significantly lower as a result of participating in the program. Finally, students' behavior at the specific discontinuities in the cash incentive program suggests that students respond strongly to incentives even in ways which may not be desirable to educators.

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,[2] and in the last decade, policymakers throughout the world have experimented with programs that pay students for academic performance.

Israel, for example, 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 (2007) find that cash incentives in

[2] 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.

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 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).[3] 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

[3] 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.

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). More recent work examined the implications and existence of extrinsic motivation crowding out intrinsic motivation in education and in economic contexts (e.g. Gneezy & Rustichini 2000, Cameron & Pierce 1994, Eisenberger & Cameron 1996, Frey 1994, Frey & Oberholzer-Gee 1997).

Researchers in psychology have found that the efficacy of external motivators is context specific (e.g. Deci 1978, Csikszentmihalyi 1978). The efficacy of external rewards depends on the type of behavior being incentivized and the type of reward, and the efficacy may even vary from student to student.

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, students' academic performance, especially after the incentives are no longer present, is a potential indicator of intrinsic motivation. Because of the multi-year nature of Coshocton's program, we can also examine the effects of the program over multiple years and once students were no longer eligible. Finally, 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. Eligible students receive $15 for each test on which they score proficient or better. On any test for which a student scores in the "Accelerated" or "Advanced" 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. The school district mails students' rewards in early June after the release of testing results.

In spring 2004, the school board unanimously adopted the Coshocton Incentive Program. To determine the specific details of the program and to educate and resolve concerns that the community might have about the program, the school board set up an advisory and implementation committee. This committee consists of representatives of various constituencies throughout Coshocton. Members included the district superintendent, the director of curriculum, a member of the school board, a representative from the Simpson Family Foundation, the president of the local teacher's union, one teacher from each school involved in the program, one parent from the parent-teacher associations of each participating school, a principal from one of the participating schools, the school district's special education coordinator, two representatives from the business community, and one expert in research design and evaluation. Broad

[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.

representation of parents, educators, and community leaders has enabled them to solicit feedback and to be sensitive to concerns throughout the community. It also created the political momentum to allow the program to continue. In 2005, a similar program in New York City lacked political momentum and grass roots support. The program was cancelled although in 2007, it was revived using only privately donated funds (Medina 2007).

As a condition for Mr. Simpson's donation, Coshocton City Schools had to agree to implement the program using randomization. The 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. 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.[5]

---

[5] 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-

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 2007). 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.[6] We used a bingo-cage and ping-pong balls (one for each grade-school combination) to conduct the drawing 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

---

trials often conducted across hospitals or communities. Group-randomized trials are attractive in places where randomization at the student or patient level is impractical.

[6] 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.

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

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

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 previous year 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.[7]

[7] 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 2001, and Wooldridge 2003). Our results do not change when we correct our clustered standard errors in the way prescribed by Wooldridge (2003).

Our outcome of interest will be student test scores. Students generally take five tests – mathematics, reading, writing, science, and social science.[8] We examine each score separately. 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.

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.

---

[8] We exclude writing from the analysis for two reasons. First, 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.

[9] In every case, we use the scale scores as the primary test score.

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 in 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. Figures 1A and 1B report the differences in the treatment and control distributions of pre-program math and reading test scores. We find almost no difference in the CDF's.[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.

[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] Appendix Figures 1-4 show the math, reading, social science and science distributions once we regression-adjust for grade, year, and school interactions. All of the distributions looks similar except social sciences where there appears to be some pre-program differences.

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.

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.

Figure 2A shows the regression adjusted differences in the cumulative densities for the treatment and control groups in terms of the math test scores.[13] The treatment group's distribution has shifted to the right of the control group's distribution with noticeable differences at both the bottom and top of the distribution.

The results in Table 3 and Figures 5A and 5B 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. 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, accelerated, and advanced. Students were paid $15 if they made it to proficient and an additional $5 if they scored accelerated or advanced. In

[13] The covariates in the regression-adjustment include age, gender, race, and controls for grade, year, school, and publisher of test.

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 Column 2, we present a linear probability model where the dependent variable is whether students scored proficient or higher.[14] In these results, we get a point estimate of 3 percentage points, and the estimated effect is not significant. The results suggest little impact on the proportion of students scoring over this margin. In Column 3, 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 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 4 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

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

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).

By allowing a separate treatment effect depending on where students were in previous years, we can show where the effects of the program are greatest. Given that Table 4 shows large effects at the top of the distribution, we should expect positive treatment effects among students previously at the top of the distribution (assuming the distribution is stable over time). The proficient margin is near the 30th percentile, so treatment effects among the 2nd and 3rd quartiles may reflect any effects for students in this range, but given the results in Table 4, there appears to be little effect here. Finally, the bottom quartile was not represented in Table 4.

The estimates of equation 2 are reported in Table 5. They are almost exactly as predicted. 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. These positive effects were evident in Figure 2A. 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. 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. Figure 2B tells a similar story in comparing the test score distributions of treatment and control groups. 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.

Why are there effects in math but not reading? One explanation is that math scores are more elastic than reading scores. Educational interventions often increase math scores with little to no impact on reading scores (e.g. Reardon, Cheadle, and Robinson 2008).[15] Another explanation is that the incentive program did not have the "right" incentives for reading.

The design of the reading incentive differed from the other subjects in one important way. It is the one subject where a “kink” in the incentive program exists. Third graders (who made up one-third of the treatment sample) take two reading tests each year. The first test is administered in October and the second is administered in March. Reading is the only subject where this takes place. Eligible third grade students were informed that the district would reward them for their highest test score.

About 17 percent of students scored in the highest category in the first administration. These students were guaranteed a payment of $20 regardless of their performance on the second administration. Hence these students no longer had an external incentive to succeed on the second exam.

Before considering the impact of scoring high in the fall on students' performance in the subsequent spring exam, it is useful to examine whether there was an impact of the incentive on students' performance in the fall. The lottery was held in September each year and the fall reading exam was administered to third graders in October. The incentives may have been tangible for these students in this short time frame. However, if we repeat the estimation in Table 6 focusing only on third graders fall reading exam, we find that there was no significant effect. The point estimate is negative (-0.198) and insignificant (s.e. = .215). This is a very noisy estimate but the sign and lack of significance may suggest that the fall performance in reading was unaffected by the incentive program.

[15] 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.

Assuming that the incentive program did not affect student's fall performances, we now examine whether a student scoring in the highest level influenced students' behavior in the spring test. To examine whether this second exam made a difference, we focus only on the third graders in the district. We use all available data (4 years spanning from the 2002-03 school year to the 2005-06 school year). We augment Equation 1 by including a dummy variable for each of the five levels (Level 1=Lowest Level) in which a student might have scored in the October administration of the exam. Instead of including an overall treatment effect, we interact treatment status with each of the five dummy variables for the respective levels. The dummy variables for score levels control for the average March test score for students who scored in this level during the fall administration while the interactions show the differential effect by treatment status. Columns 1 and 2 of Table 7 report these interactions. We report standard errors that cluster at the school-grade-year-fall achievement level.

The most interesting and only significant result is the interaction between Level 5 and treatment. These Level 5 students learned in November that they had earned all of the money that they could for reading. The estimated effect for these students is negative and significant suggesting that they performed worse in the spring administration than other students who had scored similarly but were not eligible for the treatment.[16]

One explanation for this result is that students regressed to the mean. As a specification check for the third grade results, we also estimated the effects of these different reading level designations on math scores. Math only has one administration per year and it occurs in the spring. We use students' levels from the reading test and interact those with the treatment variable to see if fall reading designations impacted spring test scores in math. By contrast to the reading scores, every estimated effect is large and for the most part statistically significant. The estimated effect for the highest level students is positive and marginally significant.[17]

Another interpretation of the reading results is that students respond to early signals. Students who are below the threshold for the cash reward (Level 1 students) no longer try as hard and do not respond to the incentive program. Students at the top of the distribution respond to

[16] Students in Level 4 had also scored high enough to qualify for the highest reward in the incentive program as well. School administrators reported that many students were unsure if the $20 reward came at Level 4 or Level 5.
[17] When we test whether the estimated effects are similar across math and reading, we find that the differences are statistically significant for Levels 1, 3, and 5. The Level 5 and to a lesser extent the Level 3 differences are compatible with a story of students trying less in subjects where they knew they were "in the money." The Level 1 results are consistent with student discouragement after learning that they were far from the mark in reading.

the information by slacking off. In contrast, students received no early signal in math, and all students, regardless of signals in other disciplines, seemed to respond to the incentive program.

In sum, we find no overall effect on reading. The third grade results, however, are consistent with a story that third grade students respond to incentives in a sophisticated yet predictable way.

Alternative Subjects

Students were also tested in social science and science. Table 8 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.[18] 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 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.

Multi year Treatment and Year After Effects

Another potential indicator of students' intrinsic motivation is their subsequent performance in the subject. Given the multi-year nature of the program, we can measure the

[18] The SRQ-A was validated in Ryan and Connell (1989). Detailed descriptions are available at http://psych.rochester.edu/SDT/measures/selfreg_acad.html.

effects of being in the treatment for consecutive years or being in the treatment one year but not the next. If the effect of the program is cumulative, then we might expect the program to continue to improve student test scores from year to year. There are five grade-school-year combinations where the students won the lottery for multiple, consecutive years of the program. There were nine grade-school-year combinations where students were eligible in one year but not the next. In our empirical specification we can include a main treatment effect, a treatment effect for individuals who are in their second year of treatment, and a separate control for students who are a year removed from the treatment. In Table 9, we present these results focusing on the sample of students ever involved in the lottery. As before, we find treatment effects in math between 0.12 and 0.18 standard deviations that are significant. For students in the second year of the treatment, the overall treatment effect is positive but not significant. This is somewhat different from the estimated effect in reading. In reading, as before, the overall treatment effect is not significant; however, the effect in the second year is positive and significant. The major caveat here is that we only have only a five treatment cells that have had treatment over multiple years.

In our specification we also include an indicator for whether students had previously been on the program but were now no longer eligible. In math, the estimate is not significant and small in magnitude. In reading, the estimated effect is strongly negative and significant. The effects are larger than any other effects measured in the paper. As before, the one concern with these results is that there are a limited number (nine) of grade-school cohorts where students were eligible one year but not the next. In a large sample, the randomization would assure balance between this grade-school cohort and others who had not won the lottery. In the small sample, there could be some imbalance. For example, these nine cohorts were 20 percentage points more likely to be participating in the free/reduced lunch program than students currently in the treatment, students who had never been selected, and students who had been selected in one year but not the next. While we control for free/reduced lunch in Table 9, there may be important unobservable characteristics among this group.

While we are cautious about the "year after" results because of the small number of cohorts, the results may also be an indicator of students' post-program intrinsic interest in math and reading. The math results suggest that there was little drop-off in intrinsic motivation for math. The reading results suggest that there were potential declines in test scores once the program was removed.

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 2007) 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 10. 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.

Figures 3A and 3B show the cumulative distribution functions of math scores for both females and males. In both cases, it is clear that there was an increase in student test scores throughout the distribution if students were eligible for the incentive program. Figures 4A and 4B show analogous results for reading test scores. The results here do not show any significant shift in the distribution of test scores for either females or males who were eligible for the incentive program.

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 11. 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. Even in reading, however, we find some evidence that students responded to incentives in meaningful and predictable ways.

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.

We 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, 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.[19]

While the short-run effects suggest a positive impact of the program, the data have less to say about the long-run impacts of the program. We observe "year after" effects in math that are small and not significant. The reading results, by contrast, were negative and significant. However, we caution that we may not have adequate sample to accurately measure these "year after" effects. Future research and additional data collection may shed additional light on this.

Coshocton's program, however, was highly cost effective relative to other 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.[20] By contrast, in Project STAR class size dropped from the around 24 to around 15, and the average test score gain from small classes was 0.25. Given the findings on class size, the projected drop in average class size in Coshocton would not generate a 0.15 standard deviation effect. The Coshocton incentive program was thus 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.

[19] 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.

[20] Coshocton would have had to hire 7.6 new teachers to reduce class size to 15.

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 (2007) "The Effects of High Stakes School Achievement Awards: Evidence from a Group-Randomized Trial." MIT mimeo.

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.

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, Stephen, and Kevin Lang (2001), "Inference with Differences-in-Differences and Other Panel Data," Boston University Department of Economics, mimeo.

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.

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.

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 (2003), “The Effect of Financial Rewards on Students’ Achievement: Evidence from a Randomized Experiment,” CEPR Discussion Paper no.3921.

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).

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 time of test)	Free/Reduced Lunch Participation	White
	(1)	(2)	(3)	(4)
Treatment	0.013	-13.61	0.012	-0.0005
	[0.026]	[8.37]	[0.018]	[0.0114]
Grade, Year, School Controls	Yes	Yes	Yes	Yes
N	1527	1504	1527	1527
N (students)	893	887	893	893
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 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
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 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 Proficient (lvl>=3) Mean=.61	Over Advanced (lvl>=4) Mean=.33	Accelerated (lvl=5) Mean=.16
	(1)	(2)	(3)	(4)
Treatment	.203 [.074]	.038 [.028]	.092 [.026]	.052 [.024]
Age, Female, Race, and Socioeconomic Control	Yes	Yes	Yes	Yes
Grade, Year, School Controls	Yes	Yes	Yes	Yes
N	1248	1248	1248	1248
N (students)	840	840	840	840
N (grade- school-year)	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)
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
Grade, Year, School Controls	Yes	Yes
School by Year Fixed Effects	No	Yes
N	1615	1615
N (students)	873	873
N (grade- school-year)	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 3rd Grade March Exams
by October Reading Exam Performance

	Reading Test Scores March Administration		Math Test Scores March Administration	
	(1)	(2)	(3)	(4)
Treatment*Level 1 on Fall Reading Test (Level 1 = lowest level)	.026 [.105]	-.027 [.099]	.379 [.086]	.330 [.087]
Treatment*Level 2 on Fall Reading Test	.135 [.118]	.133 [.125]	.229 [.144]	.260 [.156]
Treatment*Level 3 on Fall Reading Test	.121 [.111]	.095 [.100]	.373 [.115]	.349 [.113]
Treatment*Level 4 on Fall Reading Test	.165 [.106]	.163 [.110]	.338 [.108]	.353 [.106]
Treatment*Level 5 on Fall Reading Test (Level 5 = highest level)	-.309 [.108]	-.313 [.104]	.229 [.149]	.270 [.157]
Grade, Year, School Controls	Yes	Yes	Yes	Yes
Age, Gender, Race Controls	No	Yes	No	Yes
N	462	462	560	560
N (students)	462	462	450	450
N (grade- school-year-lvl)	76	76	76	76

Sample includes all 3rd Graders from 2002-03 to 2005-06. Standard errors in brackets control for clustering across grade-school-year-fall achievement level combinations.

Table 8. 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 9. OLS Estimates of Effects of Pay to Learn in 2nd Year by First Year Treatment Status

	Math Scores				Reading Scores			
	(1) Full Sample	(2) Full Sample	(3) Cluster Avg	(4) Donald-Lang	(5) Full Sample	(6) Full Sample	(7) Cluster Avg	(8) Donald-Lang
Treatment Effect	.182 [.056]	.123 [.059]	.230 [.070]	.215 [.068]	-.037 [.045]	-.076 [.052]	-.013 [.064]	-.010 [.062]
Difference in Treatment Effect During the Second Year of Treatment	.008 [.102]	.054 [.096]	-.142 [.147]	-.137 [.144]	.151 [.070]	.200 [.080]	.009 [.105]	.003 [.103]
Difference in Test Scores the Year After Being on the Treatment	-.004 [.068]	.029 [.057]	-.030 [.083]	-.011 [.084]	-.309 [.058]	-.310 [.065]	-.272 [.120]	-.252 [.109]
Age, Gender, Race, FRL Controls	Yes	Yes	No	Yes	Yes	Yes	No	Yes
School by Year FE	No	Yes	No	No	No	Yes	No	No
N	2107	2107			3006	3006		
N (students)	921	921			937	937		
N (grade-school-year)	63	63	63	62	63	63	63	62

Notes: Sample includes students in 3rd through 6th grade for the 2004-05 to the 2006-07 school years. It also includes 7th graders in 2006 and 7th and 8th graders in 2008. Standard errors in brackets control for clustering across grade-school-year combinations which is the level at which treatment was assigned.

Table 10. 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 11. 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 [.050]	.114 [.052]	-.003 [.036]	-.002 [.039]
Sibling was Eligible for Treatment (but student was not)	-.055 [.054]	-.066 [.056]	-.048 [.091]	-.043 [.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.

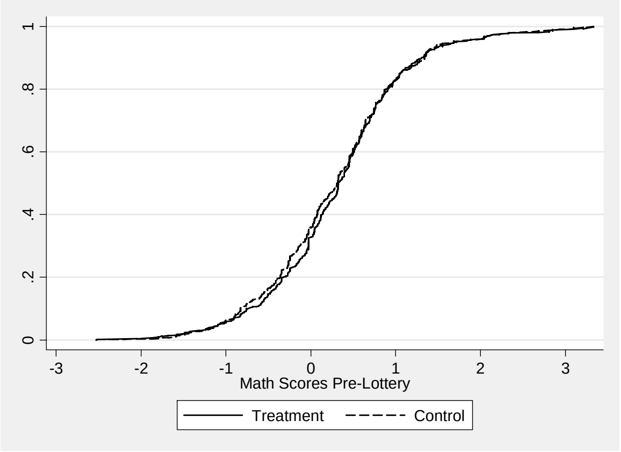


Figure 1A. CDF's of 2004, Pre-Lottery, Unadjusted Math Test Scores by Treatment Status

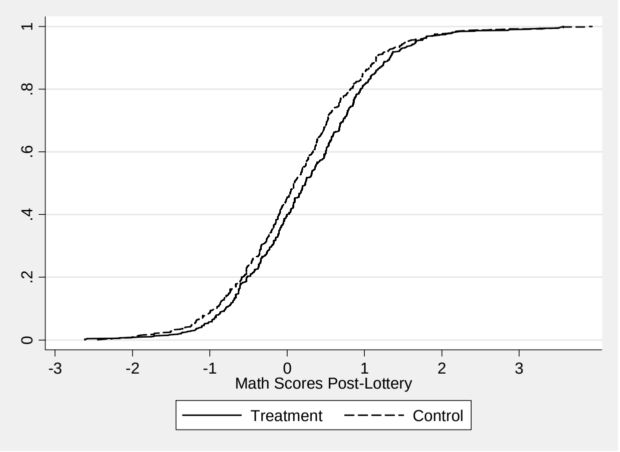


Figure 2A. CDF's of Post-Lottery, Unadjusted Math Test Scores by Treatment Status

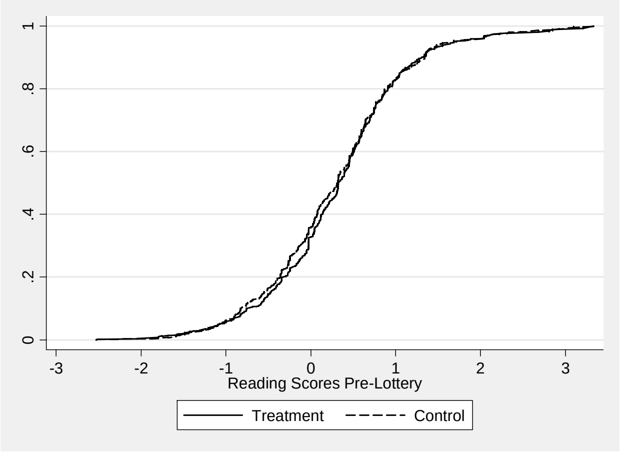


Figure 1B. CDF's of 2004, Pre-Lottery, Unadjusted Reading Test Scores by Treatment Status

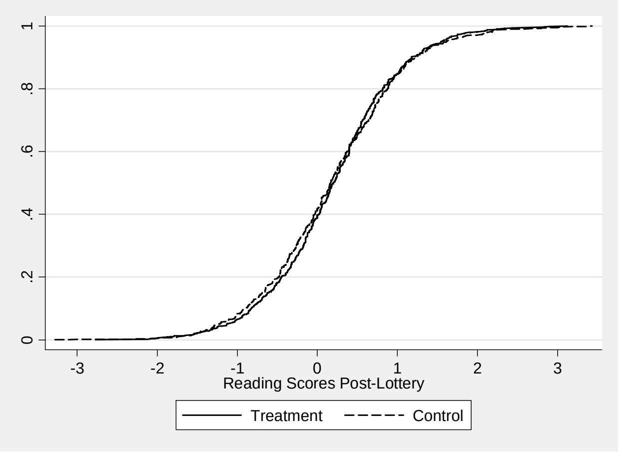


Figure 2B. CDF's of Post-Lottery, Unadjusted Reading Test Scores by Treatment Status

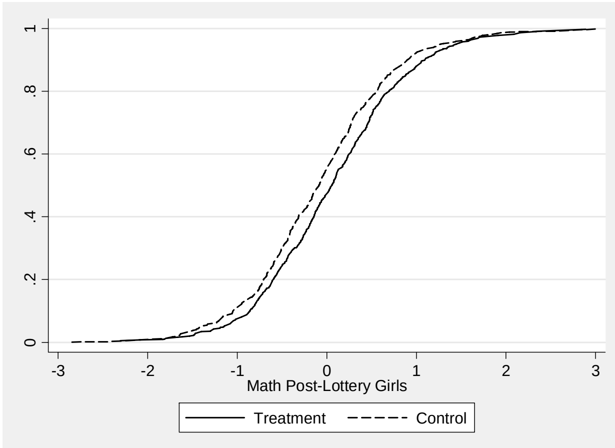


Figure 3A. CDF's of Post-Lottery, Regression-Adjusted Math Test Scores by Treatment Status, Girls

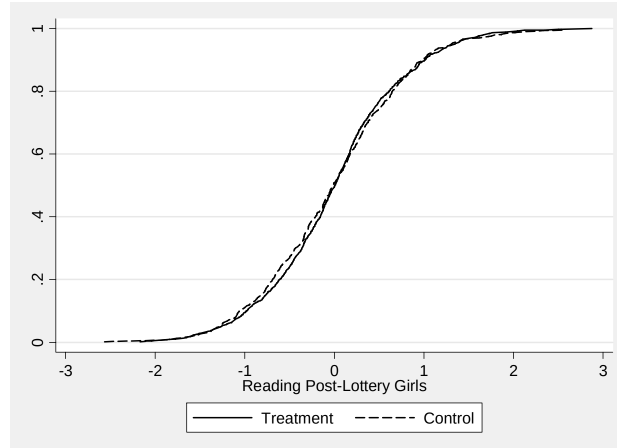


Figure 4A. CDF's of Post-Lottery, Regression-Adjusted Reading Test Scores by Treatment Status, Girls

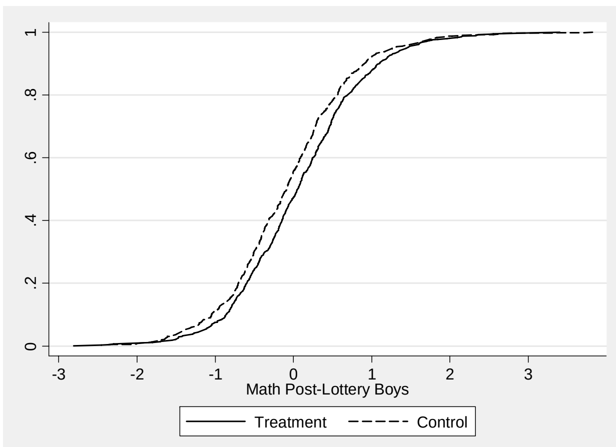


Figure 3B. CDF's of Post-Lottery, Regression-Adjusted Math Test Scores by Treatment Status, Boys

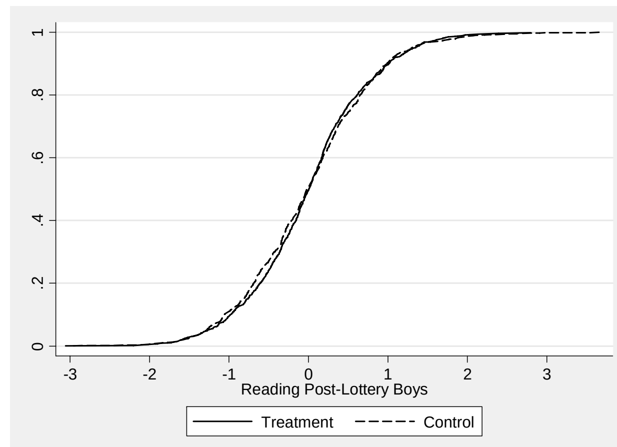


Figure 4B. CDF's of Post-Lottery, Regression-Adjusted Reading Test Scores by Treatment Status, Boys

Appendix. Coshocton Incentive Winners

2004-2005 School Year

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

2005-06 School Year

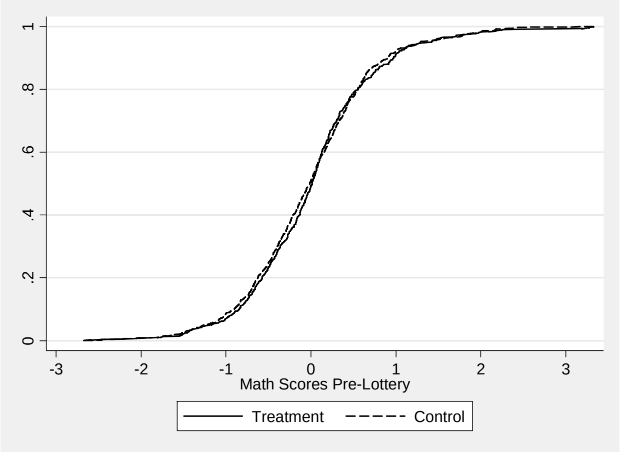
Washington	5th
Central	3rd, 5th
South Lawn	3rd, 4th, 6th
Lincoln	4th, 6th

2006-2007 School Year

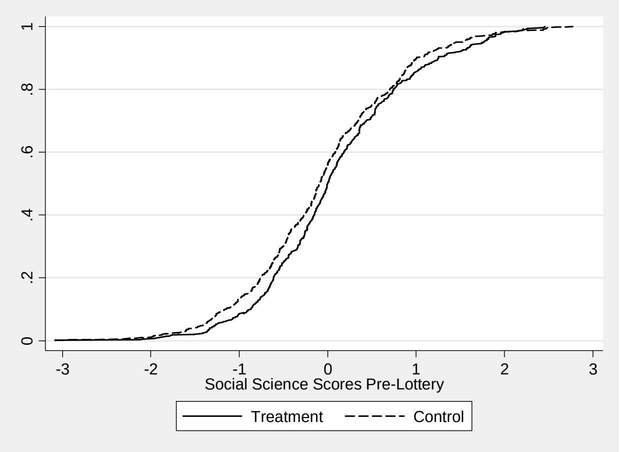
Washington	3rd, 4th, 6th
Central	4th, 6th
South Lawn	5th
Lincoln	3rd, 6th

Treatment Years By School Cohort:

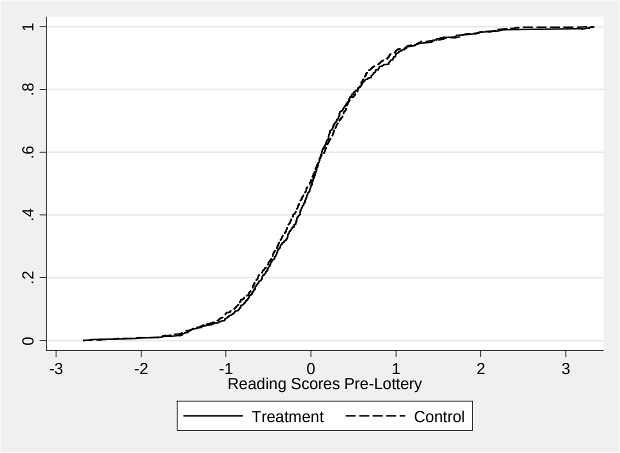
School	Grade in 2004-05	Years Won Lottery
Washington	1	2007
	2	2007
	3	2005
	4	2005, 2006, 2007
	5	
	6	2005
Central	1	
	2	2006, 2007
	3	2005
	4	2006, 2007
	5	
	6	2005
South Lawn	1	
	2	2006
	3	2005, 2006, 2007
	4	
	5	2006
	6	2005
Lincoln	1	2007
	2	
	3	2005, 2006
	4	2007
	5	2006
	6	



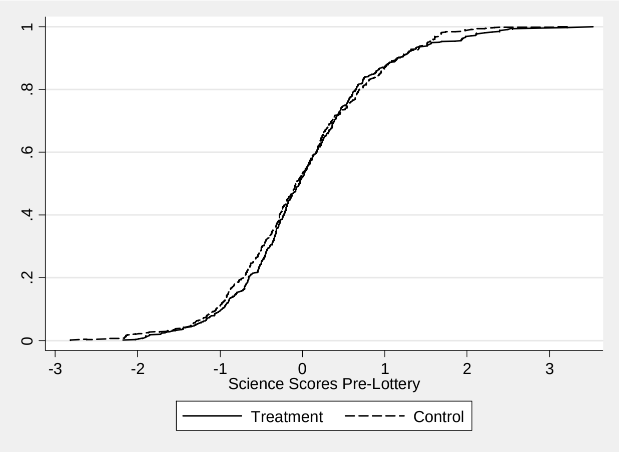
Appendix 1. CDF's of 2004, Pre-Lottery, Regression-Adjusted Math Test Scores by Treatment Status



Appendix 3. CDF's of 2004, Pre-Lottery, Regression-Adjusted Social Science Test Scores by Treatment Status



Appendix 2. CDF's of 2004, Pre-Lottery, Regression-Adjusted Reading Test Scores by Treatment Status



Appendix 4. CDF's of 2004, Pre-Lottery, Regression-Adjusted Science Test Scores by Treatment Status