Harnessing Complementarities in the Education Production Function

John A. List
University of Chicago and NBER

Jeffrey A. Livingston[a]
Bentley University

Susanne Neckermann
University of Chicago

Abstract

Recent research shows that complementarities between inputs in the education production function might be harnessed by implementing two policies simultaneously. However, budget constraints may make fully funding both policies infeasible, forcing school districts to choose between fully funding one policy or partially funding several. To address this issue, we conduct a field experiment where $90 of financial incentives are provided to either one input (students, parents, or tutors) or spread equally among multiple inputs to improve on a standardized test and meet other standards. The results suggest a limited amount of financial incentives has a greater impact on student achievement when directed at just one input.

JEL: C93 (Field Experiments), D2 (Production and Organizations), I21 (Analysis of Education)

Keywords: field experiments, education production function, complementarities

[a] Contact author. Associate Professor, Bentley University, Department of Economics, 175 Forest Street, Waltham, MA 02452, email: jlivingston@bentley.edu

I. Introduction

There is mixed evidence on whether increasing spending on education improves student outcomes. Some studies find that education expenditures per pupil are uncorrelated with achievement at both the state level (Hanushek, 2016) and national level (OECD, 2012), while others find strong benefits of increased spending (Jackson, Johnson, and Persico, 2016; Jackson, Wigger, and Xiong, 2018). What is clear is that in order for spending to have positive effects, it is critically important to allocate resources in the most effective way possible. As Hanushek (2016) argues, "There now appears to be a general consensus that how money is spent is much more important than how much is spent… What remains to be unpacked is the precise ways in which expenditure needs to be directed and administered if it is to lift student achievement efficiently and effectively."

In this pursuit, recent work examines whether harnessing complementarities between inputs in the education production function might have a greater impact on educational attainment than policies that target only one input at a time. Historically, many studies assumed that the education production function is linear and additively separable in its inputs,[1] but recent work relaxes this assumption.[2] A large and growing body of empirical work finds evidence of

[1] See Rothstein (2010) and Harris, Sass, and Semykina (2014) for reviews of some of these models.
[2] De Fraja, Oliveira and Zanchi (2010) present a model where student achievement enters into the objective functions of students, parents and schools, and each selects their effort level. Student achievement increases with the effort level of each input, and the sign of their best response functions depends on whether the efforts of each input are strategic substitutes or strategic complements.

complementarities,[3] though some studies support the conclusion that inputs are substitutes.[4] Most recently, Mbiti et al. (2018) compare three treatments where the first provides resources to schools in the form of block grants, the second gives bonus payments to teachers if their students meet achievement standards, and the third does both at the same time. They find the greatest impact when both treatments are combined, and that the impact of the third treatment is greater than the sum of the impacts of the treatments that target only one input, suggesting that the policies have complementary effects.

While much of the empirical work addressing this question provides evidence that complementarities between inputs exist, it often does so by comparing the effects of implementing two policies at the same time to the effects of implementing each of those policies individually. As a result, the more successful treatments have larger budgets than the less successful treatments. However, schools frequently face binding budget constraints, so evaluating how to allocate scare resources among different policy alternatives, is a crucial policy question.

[3] For example, Fryer (2012) gives incentives to students, parents, and teachers at the same time to complete mathematics objectives and/or hold and attend parent-teacher conferences. He finds modest impacts on mathematics test scores relative to a control group where no input is given incentives, but larger impacts among high-achieving students. Behrman et al. (2015) find that giving financial rewards to both students and their teachers at the same time if students meet various achievement thresholds on standardized tests has a greater impact than giving the same incentive to only one input or the other. Johnson and Jackson (2017) find that increased Head Start spending has a bigger impact on students who are also later exposed to higher levels of public K-12 school spending. Martorell et al. (2016) conduct an experiment where student attendance of summer school is incentivized. The experiment has two treatment groups, one where students receive non-financial incentives and another where parents additionally receive financial incentives along with the student non-financial incentives. They find that the combined incentives are more effective than the student-only incentives. Finally, Geng (2018) studies the staggered implementation of two policies that were enacted in New York City: a policy that required students to meet a minimum proficiency threshold on standardized tests in reading and mathematics in order to advance to the next grade level, and a school accountability policy whereby poor performing schools risked closure. He finds that the policies had small effects when implemented individually, but the combined policies resulted in larger impacts on mathematics test scores, student absences, and suspension rates.

[4] For example, Das et al. (2013) find in experiments in India and Zambia that households respond to anticipated increases in school resources by decreasing the resources that they provide. Similarly, Houtenville and Conway (2008) find that parents may decrease their effort in response to an increase in school resources.

In this study, we consider this issue with the particular example of financial incentives. We study whether a policy of giving financial incentives to encourage improved performance is more effective when the incentives are directed towards one specific input in the education production function or spread among them. Such incentives have been recognized as a potentially cost-effective way to improve educational outcomes; Fryer (2011), Levitt, List, and Sadoff (2016), and Barrow and Rouse (2018) each contribute to and review much of this literature. As Fryer (2011) notes, students especially may have high discount rates and fail to work hard in school because the benefits come too far in the future; incentives make these rewards more immediate.

To this end, we conduct a randomized field experiment where we provide financial incentives to three key inputs into a student's education (the students themselves, their parents, and tutors who were hired to work with the students) that give rewards if the student improves on a standardized test and meets other academic and behavioral standards. Either a single input or a combination of these three inputs are provided incentives to meet (or to aid the student in meeting) these standards.

In each case, a pool of $90 is made available to the incentivized parties, regardless of how many inputs are targeted. For example, when only the student is given an incentive, the student is paid $90 if all of the achievement standards are met. But when both the student and their parent are incentivized, each is paid $45 if the student meets the standards. Finally, when all three are incentivized, each is paid $30 if the student meets the standards. When more than one input is incentivized, the effects are likely enhanced by complementarities between the inputs as found in the extant literature, but subjects may not respond as strongly to the incentive since the rewards are lower. The overall effect of the incentives may be smaller if this price effect outweighs the benefits of synergies between the inputs.

Our results suggest that when the incentives are lowered but spread across multiple inputs, the price effect outweighs the effect of complementarities. When only one input is incentivized, we observe similar test score gains regardless of who receives the incentive – the student, the parent or the tutor. The effect sizes are substantial, ranging between 0.33 and 0.40 standard deviations. However, when the same budget is split across multiple inputs, the gains relative to control are smaller. For example, when the reward is shared by all three inputs, the estimated impact on test scores is only 0.08 standard deviations and is statistically insignificant.

II. Experimental Design

The experiment was conducted in Chicago Heights, IL, a suburb thirty miles south of Chicago. Each of the city's nine elementary and middle schools were involved in the experiment. While there are some differences in the demographic composition of the schools, the schools as a whole are populated largely by low-income and minority students. 38 percent are African American, 53 percent are Hispanic, and 93 percent are eligible for the district's free lunch program. They also struggle with low rates of success in meeting state achievement standards. Only 53 percent of students met the minimum proficiency standard on both the reading and math portions of the Illinois Standards Achievement Test (ISAT) in 2010, the results of which were applied to the No Child Left Behind Act to identify failing schools.

Our design consists of six treatment groups and one control group. The goal of the design is to investigate whether a limited pool of financial incentives is more effective when they are directed towards one specific input or spread among them. Accordingly, the treatment groups allocate $90 of incentive payments across three inputs in various combinations – the student, the parents, and a tutor with whom the student worked on preparation for the 2011 ISAT. The first three treatment groups incentivize only one of the student, the student's parent(s), or the tutor. If the student meets all of the required achievement and behavioral standards (described below), the

input receives the entire $90. We refer to these treatments as *student only $90, parent only $90,* and *tutor only $90,* respectively. The fourth treatment group investigates whether this $90 is more effective if shared among the student and the parents. Each receives $45 if the standards are met. We refer to this treatment as *student and parent $45 each*. The fifth treatment group splits the incentive across all three parties; each earns $30 if the standards are met. We refer to this treatment as *all three $30 each*. The final treatment incentivizes a single input, students, but pays them only $30 if all standards were met. We refer to this treatment as *student only $30*.

There are four academic and behavioral standards students are required to meet to earn these incentives. They are inspired by those employed by Levitt, List and Sadoff (2016) and include: improving by at least one point (out of 20) on a standardized test that we created, improving the student's grade in the relevant subject (reading or math) or maintaining it at its previous level and above a failing grade of F, having no more than two unexcused absences during the assessment period, and having no all-day suspensions during the assessment period.[5,6]

The standardized tests used in the experiment were created using resources provided by Discovery Education, which make it possible to create "probes" to measure a student's

[5] The original standards were provided by the school leadership where Levitt, List, and Sadoff (2016) conducted their experiment, and are based on what are considered to be the minimum requirements necessary to complete the ninth grade. They include: no more than one unexcused absence and no all-day suspensions in the month, letter grades of C or higher in all classes on the last day of the month, and when available, scoring at grade level or improving upon a standardized school reading assessment taken in the previous month. These standards require students to meet a common threshold; in response, students who are near the threshold react more strongly to the provided incentives. As an alternative, we employ individually-tailored standards to avoid such threshold effects. Also, they conducted monthly evaluations while our assessment periods last roughly two months. Our attendance standard accordingly allows one additional unexcused absence.

[6] In addition to these standards, we wanted to provide parents with a tool for helping their child improve. At the end of each week, tutors were required to create a homework assignment for each of their groups that was designed to be a review of what they had covered that week. The tutors instructed the students to bring these assignments home to work on with their parents. Parents who were part of one of the parent incentive groups (*parent only $90, student and parent $45 each,* and *all three $30 each*) then faced the additional requirement of completing these assignments with their child each week, and having their child return them to the tutor.

preparedness for the ISAT at any time. These exams were created by randomly drawing questions from a test bank of questions that cover the same skills and knowledge that is tested on the ISAT. Separate probes were created for each grade level (K through 8) and subject (reading and math). The probes consist of 20 multiple choice questions, are administered by the tutors,[7] and are taken by computer. We used these exams in the experiment instead of the actual ISAT because the system scores the exams automatically, making the results are available immediately upon completion of the test.

We conduct two experiments. Students took a baseline probe at the beginning of each experiment and an assessment probe at the end of each experiment. Experiment I began on January 10th, 2011 with the announcement of the incentives to the subjects. The baseline probe was taken during the week of January 17th, 2011. Experiment I concluded when the assessment tests were administered between March 21st and March 25th, 2011.[8] Experiment II used the Experiment I assessment tests as the baseline tests. The assessment probe for Experiment II was taken by most students beginning on May 23rd, 2011,[9] but the experiment concluded with the release of final trimester grades on June 6th, 2011.

[7] Because tutors met with their various groups of students at different times throughout the course of the week, it was impossible for the experimenters to administer the exams to the students. We therefore had to have the tutors administer the exams to each of their groups. While this may have allowed tutors to cheat on the exams by providing the students help or even providing answers, it was the only feasible alternative.
[8] Although the experiment did not begin until January 10th at the beginning of the trimester which followed the holiday break, the tutors began meeting with their students in early November 2010. The 2015 tutors however worked with the students from January 2015 to March 2015. Their time was more limited because unlike the 2011 tutors who each worked for 100 days, the 2015 tutors were hired for only 50 days each. Also, they worked five days per week while the 2011 tutors had more irregular schedules.
[9] Near the end of the experiment, several tutors ran out of their 100 work days near the beginning of May, so they had to administer their probes early. The administration of the final ThinkLink exam and other end of the year activities also interfered substantially with the schedules of both the tutors and the students, making a consistent testing window impossible to achieve. As a whole, the final probe was administered beginning on May 5th and throughout the month of May and into the first week of June.

The *student only $30* treatment was not part of the original experiments. It was added in 2015. This additional treatment was designed to be part of Experiment I. Students who participated in 2015 took the same baseline and assessment probes that were used in Experiment I. Also, it was conducted at the same time of year: the baseline probe was conducted during the week of January 26th, 2015 and the final probe was conducted during the week of March 23rd, 2015. Also,

All 2011 subjects were part of both experiments and were assigned to the same treatment groups in each experiment. However, the assessments for the two experiments were independent; those who failed to earn a reward in Experiment I were able to do so in Experiment II, and vice versa.

The tutors with whom the students worked were hired by the school district to assist tier 2 students, who were judged to be at risk of failing to meet state minimum proficiency standards on the ISAT in reading, math or both.[10] In 2011, these tutors worked with 496 students, grades Kindergarten through eighth. Among these students, 82 met with tutors in both reading and mathematics, and 414 met with a tutor in only one subject. Those who met with more than one tutor participated in the experiment more than once. [11] This yielded 579 student-level observations which were organized by the schools into 163 groups who worked with a tutor. Students met with the tutors in groups ranging in size from one to nine; these groups typically consisted of students of the same grade level. We also hired two additional tutors in 2015 to work at one of the nine

[10] During the 2010-2011 academic year, the district hired 32 tutors for 100 days at a wage of $100 per day with the goal of assisting tier two students with learning the skills needed for ISAT preparation in either reading or math. Each of the nine schools was provided with two reading tutors and one math tutor; five English as a Second Language tutors were also employed. The experiment worked with the reading and math tutors. Of these 27 tutors, 23 were involved in the experiment. Two elected not to participate, one was converted to a permanent substitute teacher shortly after the beginning of the experiment, and one was not hired until well after the experiment began.

[11] Among these 82 students, 81 were part of the experiment twice and one worked with all three of the school's tutors and participated in the experiment three times. 8 received the same treatment for both subjects, 21 were in the control group for one subject and a treatment group for the other subject, and 53 received two different treatments for each subject.

schools in order to run the *student only $30* treatment. These two tutors along with six others who were already working at the school worked with another 69 students, 35 of whom received tutoring in both reading and math.[12] This yielded 104 student-level observations which were organized into 24 tutor-groups. A total of 565 students participated in the experiment across both years. Since several students participated more than once, we have 683 student-level observations.

In 2011, we randomized students into the original five treatments or control at the tutor-group level, rather than at the individual level, to make it easier for the tutors to keep track of each student's treatment and to avoid spillover effects. The randomization process was as follows: tutor-groups are randomly assigned to initial treatments blocking on tutor. We then improve the balance on school, homeroom teacher, subject (reading or math), grade level groups (K to 2nd, 3rd to 5th, and 6th to 8th), gender, race/ethnicity, number of meetings per week the group met with the tutor, and baseline test score using the following procedure. We randomly select a pair of tutor-groups to swap treatment assignments, calculate an overall imbalance score which is based on a hypothesis test that the randomization is balanced on each of the above variables,[13] and keep this new assignment if it results in a lower imbalance score. This procedure is then repeated 500 times and we utilize the resulting assignment.[14] The 2015 students were randomized into either the sixth treatment (*student only $30*) or control using the same procedure.

[12] Among these 35 students, five were in the control group for each subject, three were in the same treatment group for each subject, and 27 were in the control group for one subject and a treatment group for the other subject.

[13] These hypothesis tests are Pearson Chi-squared tests of the null hypothesis that the number of subjects assigned to each treatment from each category of the above variables are independent. In other words, we cross-tabulate each variable's categories with treatment assignment, and test the null hypothesis that the rows and columns of that table are independent.

[14] One of the tutors elected to drop out of the experiment shortly after our randomization was conducted and the tutors had already been informed of the treatment groups to which their student groups were assigned. Including the students of this tutor, 620 student-level observations were part of the 2011 randomization.

We first informed the tutors about the experiment in November, and met with them frequently to make sure that they understood all of the program's details and expectations. Students were informed of their incentives and the standards they had to meet by their tutors as well as by a letter which we provided. Parents were informed of the incentives and standards in four ways: by phone when possible,[15] by a letter we sent home with their child, by another copy of this letter which we mailed, and by a weekly letter from the tutor which accompanied the weekly assignments which the tutors sent home. The letters to parents were provided in both English and Spanish since many parents did not speak English. New letters were given to tutors, students and parents at the end of the first assessment, to remind them of the details of the experiment and that everyone was starting with a clean slate for the second assessment. Appendices A through C present examples of the letters provided to the parents, students and tutors, respectively, at the beginning of the experiment. The letters given at the beginning of the second assessment look similar.

For Experiment I (including the administration of the *student only $30* treatment in 2015), grades and information about absences and suspensions were available at the time each student took the assessment probe. Since the tests were administered and graded by computer, we were able to assess immediately which students qualified for their reward at the conclusion of the test, so students who met all four standards were paid immediately upon completion of their exam. Parents were paid two weeks later either at pizza parties we held at the schools, or by mail if they were unable to attend. All parents and their children were invited to attend, and we did not inform parents ahead of time whether they had earned a reward. At the party, we reviewed the performance of each student with their parents, paid those who qualified, and made sure the parents were aware that the incentive program was continuing and that each student started with a clean slate. We

[15] Phone contacts were rather unreliable. Parents in Chicago Heights often rely on pre-paid cell phones, so their numbers change frequently and they often forget to update their contact information with the schools.

attempted to contact parents who were unable to attend by phone, letters sent home with the students, and by mail as we did at the beginning of the experiment. Tutors were paid in person at school following the completion of the testing of all of their groups.

For Experiment II, immediate payment for the students was not possible because the assessment probes had to be administered before final trimester grades were issued on the final day of the school year, June 6th. All students and parents who qualified were paid by mail. Tutors who earned rewards were paid either in person or by mail.

III. Data and Results

III.1 Balance on covariates

Table 1 reports the sample means and standard deviations by treatment group for pre-treatment characteristics and for baseline achievement in our sample.[16] The table indicates significant differences between treatment and control group means, calculated using standard errors clustered by tutor-group and student. Although the 2011 and 2015 samples were randomized separately, subjects that were randomized into control in each case are pooled together into one control group.[17] There are no statistically significant differences in baseline probe score, standardized by grade and subject (reading or math).[18] Subjects in the *student only $30* treatment

[16] The first panel reports baseline probe scores standardized within sample to have a mean of 0 and a standard deviation of 1 by grade (K through 8) and subject (reading or math). All test standardizations were done using the pooled samples from 2011 and 2015. The second panel reports demographic characteristics such as gender and ethnicity as well as the number of tutor meetings the students had per week and whether or not parents received our letter explaining the program and treatments. The last panel reports attrition caused by students leaving the program or tutors dropping out.

[17] All subsequent analysis pools the control groups from 2011 and 2015 and compares all six treatment groups to this pooled control group. The point estimates are very similar and the qualitative conclusions are unchanged if either a 2015 indicator is included in the specifications or the data from each year are analyzed separately.

[18] While the average standardized baseline probe score in the control group is substantially lower than most of the treatment groups, this is driven by differences between the 2011 and 2015 cohorts and the fact that the control group contains many observations from 2015. This difference is not present in either the

answered a smaller percentage of easy questions correctly than control; the difference is significant at the five percent level.

Among demographic characteristics, the only statistically significant differences are driven by differences between the 2011 and 2015 cohorts. The 2015 cohort had a larger percentage of Hispanic students than the 2011 cohort (68 percent versus 40 percent); thus, several treatments that consist of only 2011 students have significantly fewer Hispanic students and significantly fewer than control which includes many 2015 students. Similarly, a larger percentage of students in the 2015 cohort were eligible for reduced price or free lunch than the 2011 cohort (100 percent versus 88 percent). If the control group is restricted to include only 2011 students, these differences are not present, and the analysis presented below is qualitatively unchanged and the point estimates are very similar. Further, as shown below, including controls for pre-treatment characteristics as well as baseline performance does not alter the results. Also, all of the 2015 students were eligible for reduced price or free lunch and all 2015 groups met with their tutors five times per week. Consequently, these characteristics are overrepresented in both the control group (since 52 of the 152 control subjects participated in 2015) and the *student only $30* treatment (all of whom participated in 2015), resulting in several significant differences between some treatments and control in both of these variables.

III.2 Empirical strategy

Our analysis begins with a detailed look at the impact of the treatments on the probes, our incentivized standardized test. Discovery Education classifies each question on these probes as easy, moderate, or difficult. Thus, we are able to examine improvement not only on the overall score, but also on the percentage of each type of question answered correctly to see on which

2011 or 2015 samples, and as noted above, analyzing the subsamples separately yields qualitatively identical results.

margin improvement is occurring. We also examine the other incentivized outcomes: the course grade received in the relevant class (reading or math), the number of unexcused absences and suspensions, and whether the student meets all standards and achieves the reward threshold.

For each of these outcome measures, we estimate variants of the following equation by Ordinary Least Squares:

$$A_{igjrt} = \alpha A_{igjrt-1} + \beta_1 T_{gjrt} + \beta_2 X_{igjrt} + \beta_3 \gamma_g + \beta_4 \theta_j + \varepsilon_{igjrt}, \quad (1)$$

where A_{igjrt} is the achievement of student i in grade g, assigned to tutor j and group r in assessment period t; $A_{igjrt\text{-}1}$ is the baseline assessment from the previous period, T_{jgrt} is a vector of variables indicating the treatments assigned to tutor-group r where the control group is the omitted category, X_i is a vector of individual student characteristics;[19] γ_g and θ_j are grade and tutor fixed effects, respectively; and ε_{igjrt} measures white noise. Standard errors are clustered at two non-nested levels: by tutor-group, which is the level of randomization, and by student since some students were in both the reading and math programs and participated in the experiments as part of more than one tutor-group.

III.3 Results

We begin by examining the effect of our treatment groups on the extent to which student scores improved on the probe exams in the raw data. Figure 1 displays the impacts of the treatments on average differences between raw scores (out of 20) on the baseline and assessment probes in Experiment I. Each treatment results in a larger improvement than experienced by control students. However, larger gains accrue to the treatments where the incentive is targeted at only one input.

[19]These characteristics include gender, race/ethnicity (African-American, Hispanic or Caucasian), the number of meetings the student had each week with her tutor, eligibility for free or reduced price lunch, a dummy variable indicating whether the initial mailing was received by the parents, and dummy variables indicating whether the student was in multiple tutor-groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in only one tutor-group are the omitted category).

Control students improve over their baseline score by an average of 1.49 points, while students in the *parent only $90*, *student only $90*, and *tutor only $90* treatments improve by 2.74 points, 2.26 points, and 2.38 points, respectively. The gains are similar for students in the *student and parent $45 each* and *student only $30* treatments, who improve by 2.28 and 2.27 points, respectively. The gains when the $90 payment is split among all three inputs are smaller, however, at only 1.81 points.

We next examine how the treatments impact the distribution of probe score improvement. Figure 2 presents empirical CDFs of probe improvement. The first six panels compare one of the treatment groups to control. Each treatment appears to shift the distribution of probe improvements to the right, particularly among the weaker students in the left portion of the distributions. The effect of the *parent only $90* treatment appears biggest; it results in the largest shift in the distribution from control and has the largest shift among the better students who improve their scores by larger amounts.

Table 2 presents *p*-values of tests of whether each of these distributions are statistically different from each other.[20] The results indicate that each of the $90 individual incentives and the $45 incentive split among students and parents generate shifts away from the control distribution. However, the *parent only $90*, *student only $90*, and *tutor only $90* distributions are also significantly different from the *student and parent and tutor $30 each* distribution at the 10 percent level at a minimum. The $90 of incentive payments results in a bigger shift in the distribution of score improvements when targeted at one input rather than being split across all three inputs. None

[20] The test statistics are constructed using permutation methods based on Schmid and Trede (1995) and run one-sided tests for stochastic dominance and separatedness of the distributions (see also Imas, 2014). The test statistics identify the degree to which one distribution lies to the right of the other, and take into account both the consistency of the differences between the distributions (i.e. how often they cross) and the size of the differences (i.e., the magnitudes). The *p*-values are computed by Monte Carlo methods with 100,000 replications.

of the other distribution differences are statistically significant, however, including the s*tudent only $90* and *student only $30* distributions. These distributions overlap substantially, as shown in the final panel of Figure 2.

The story is somewhat different when baseline and student characteristics are controlled for. Table 3 presents estimates of equation 1 where the dependent variable is the Experiment I standardized probe score. Column 1 reports the effects of the treatments on the overall probe score, standardized by grade and subject (reading or math).[21] The individual input incentives s*tudent only $90* as well as *parent only $90* and *tutor only $90* each have a statistically significant and sizeable positive effect on probe scores.[22] The individual reward conditions have similar and substantial impacts on performance: an increase in test scores ranging from roughly 0.33 to 0.40 standard deviations. The coefficients for these treatments are not statistically significantly different from one another, so there is no evidence that any one input is more vital than the others.

When the $90 is split between inputs or reduced in size, the estimated effects are smaller. The point estimate of the effect of *student and parent $45 each* is lower at 0.25 standard deviations and significant at only the 10 percent level, but statistically insignificantly different from the individual reward effects. Finally, the estimated effect of splitting the reward three ways (*all three $30 each)* is both much smaller in magnitude (only 0.08 standard deviations) and statistically insignificant. The *all three $30 each* effect is statistically significantly different from the *parent only $90* and *tutor only $90* effects, but all other differences between the effects of the treatments

[21] The regression in column 1 uses only 645 observations rather than 683 because 15 students were absent at the time when either the initial assessment probe or the second assessment probe was administered, 21 students did not have all demographic characteristics available (ethnicity, eligibility for reduced price or free lunch, and/or receipt of mailings), and two students were missing both demographic characteristics and at least one probe score. All 38 of the affected observations are from the 2011 experiment.

[22] In each regression that follows, inference is based on *p*-values that are adjusted for multiple hypothesis testing since we test for the significance of the effects of six treatment groups at the same time. The adjustment controls the false discovery rate following the two-stage "sharpened" procedure proposed in Benjamini, Krieger, and Yekutieli (2006) and reviewed by Anderson (2008).

are statistically insignificant. Overall, while in most cases we cannot reject the hypothesis that each of the treatments has the same impact on student achievement, the evidence is most consistent with the hypothesis that concentrating available resources on one particular input has the greatest impact.

We added the *student only $30* treatment in 2015 to examine the functional form of the response to incentives for a single input. By comparing its impact to that of *student only $90*, we can see whether lowering the reward amount has a smaller effect, at least among students. The point estimate of the effect of *student only $30* is only 0.04 standard deviations, 0.28 standard deviations lower than that of *student only $90*, though the difference is not statistically significant. This suggests that lowering the reward from $90 to $30 can be expected to result in far less impact on a subject's behavior than the larger reward. In the circumstance we study, the impact of complementarities on the effect of the split but lower incentive would have to be enormous in order to be policy relevant.

The incentives have the biggest impact on student performance on the easiest exam questions. Columns 2 through 4 of Table 3 report the results of regressions where the outcome variable is the percentage of easy, moderate and difficult questions that the students answer correctly, respectively. The impacts of the individual incentives on the percentage of easy questions answered correctly range from increases of 6.6 to 9.3 percentage points, which represents an increase of 0.32 to 0.44 standard deviations. *Student and parent $45 each* results in a smaller gain of 5.3 percentage points, and again, there is no statistically significant effect when the reward is split three ways. No such gains from any of the treatments are evident on the more difficult questions.

The s*tudent only $90* treatment also resulted in more long-term learning gains. Column 5 shows the estimated impact of the treatments on performance on the 2012 ISAT, which was taken

in March 2012, approximately one year after the first assessment and nine months after the conclusion of the experiment.[23] The estimate suggests students who were given the $90 incentive improved on the 2012 ISAT by 0.34 standard deviations relative to control. The magnitude of this effect is almost identical to their improvement on the 2011 probe. The point estimates suggest that the *tutor only $90* treatment and the *student and parent $45 each* treatment resulted in sizeable increases in 2012 ISAT scores as well, though neither of these estimates are statistically significant.[24]

Table 4 investigates whether these results are sensitive to specification changes. Columns 1 through 3 show that the estimated impacts of the treatments are robust to changes in the characteristics and types of fixed effects that are included as regressors. Regardless of the combination of student characteristics, tutor fixed effects, and grade level fixed effects that are controlled for, the qualitative results are similar. One difference, as shown in column 1, is that the impact of *all three $30 each* is estimated to be larger (0.21 standard deviations) and statistically significant at the 10 percent level when no controls other than the treatment indicators are included in the regression.

Table 5 examines whether the effect of our treatments on student achievement varies with subject, gender, and ethnicity. Columns 1 and 2 report regressions where reading and mathematics

[23] In List, Livingston, and Neckermann (2018), we examine the impact that the original five treatments had on 2011 ISAT performance. It was taken the week before the first assessment's final probe, but performance on it was not incented. Despite the fact that the 2011 ISAT covers the same set of knowledge and skills as our incented probes and was taken at approximately the same time, we observe no parallel improvements on this test, and the point estimates of the effects of the treatments are typically negative. The evidence suggests that the probe incentives crowded out intrinsic motivation to prepare for and exert effort on the 2011 ISAT.

[24] There are fewer observations available for the 2012 ISAT regression largely because 8th graders in 2011 do not take the test in 2012. The 2012 ISAT regressions thus include data from 3rd through 7th graders only. Scores for many younger students are also unavailable since many moved out of the school district. The results are qualitatively similar if the sample is restricted to include only student observations who took all tests (the baseline and final probes, and the 2011 and 2012 ISATs.

students are examined separately, columns 3 and 4 report regressions where females and males are considered separately, and columns 5 through 7 report results for African American, Hispanic, and White students.[25] Though many estimates are statistically insignificant due to a loss of power when splitting the sample, for the most part, the same qualitative pattern of results observed for the entire sample is present across subject, gender, and the minority ethnicities. For most subgroups, the individual $90 incentives and the split $45 incentive have large estimated impacts, but the $30 incentives split three ways are ineffective. White students, however, do not show statistically significant responses to any treatment and the point estimates of the effects are frequently negative. Interestingly, the estimated impact of the *all three $30 each* treatment, along with the *student only $90* treatment, are the only ones that are positive. No conclusions can be drawn from this result since only 172 observations split across five treatments and control are involved, but the degree to which complementarities might differ by ethnicity is an important subject for future research.

Table 6 shows that the treatments do not have similarly strong effects on the other incentivized outcomes. Columns 1, 2, and 3 report the results of regressions where the dependent variables are standardized class grade, number of unexcused absences, and number of suspensions, respectively.[26] No treatment effects are statistically significant. The lack of improvement on grades is perhaps not surprising since the achievement standard merely required that the student maintain their grade at its previous level. There are also no statistically significant effects on both unexcused

[25] There is no estimated effect of the *student only $30* treatment for white students in column 7 because only two white students participated in the 2015 experiment.

[26] The grades regression uses only 633 observations rather than 683 because either first or second trimester grades were not available for 27 students, 18 students did not have all demographic characteristics available (ethnicity and eligibility for reduced price or free lunch), and five students were missing both demographic characteristics and at least one grade. All 50 of the affected observations are from the 2011 experiment. Similarly, unexcused absences and suspensions data are unavailable for three students; with missing demographic characteristics those regressions are left with 657 observations. Finally, data on at least one standard, demographic characteristic or both are not available for 64 students, leaving 619 observations for the threshold regression.

absences and suspensions, although the point estimates of the effects of the $90 individual input incentives are largely consistent with the hypothesis that these treatments reduce both of these indicators of poor behavior.

Finally, column 4 of Table 6 reports the results where the dependent variable indicates whether the student met the achievement threshold to qualify for a reward. While the point estimates suggest that the individual $90 incentives and the *student only $30* treatment each result a sizeable increase in the probability that all of the four standards required to receive a reward are met, none of these effects are statistically significant.[27]

Table 7 displays the results for Experiment II. This assessment did not include the *student only $30* treatment since the 2015 experiment only covered the original first assessment period. The results are quite different from what we observe in Experiment I. None of the treatment groups exhibit statistically significant differences from control on any of the observed outcomes.[28] The estimated impacts of the treatments on probe performance are generally positive, and in this case the smaller rewards spread across inputs have larger estimated impacts than the *student only $90* and *parent only $90* treatments. However, particularly after correcting *p*-values for the false discovery rate, none of the estimates are significantly different from control or from each other.

[27] The results are similar if the equation is estimated instead as a probit rather than a linear probability model.

[28] One crucial difference between our two assessments is a loss of students from our sample which occurred because several tutors reached the end of their 100 days early in May and had to leave the schools. Others failed to administer the probes before they left their jobs, either because they were unable to do so or they decided that doing so was not worth the effort. Accordingly, there is a substantial loss in the number of observations for the second assessment. This raises the possibility that the different pattern of results is due merely to attrition bias. The remaining students may be those who are less susceptible to treatment. As a check, we reran the first assessment regressions using only the subsample of students who are part of the second assessment. The qualitative results are the same as those reported in Table 3, so the attrited students do not appear to have been more impacted by incentives than those who remain in the sample for the second assessment. Results using this subsample are available from the authors by request.

Finally, we pool the observations from Experiment I and Experiment II and re-estimate the impact of the treatment groups on probe performance, adding a dummy variable indicating the experiment from which the observation came to the specification. The results are presented in Table 8. While only the effect of the *tutor only $90* treatment is statistically significant at the 10 percent level after correcting *p*-values for the false discovery rate, the pattern of the estimates is similar to what is found in Experiment I. The individual rewards result in improvements in probe scores of 0.20 to 0.29 standard deviations. The smaller reward amounts spread across multiple result in smaller estimated effects (0.17 standard deviations for *student and parent $45 each*, and 0.12 standard deviations for *student and parent and tutor $30 each*), though the gap is not as large as found when considering Experiment I alone.

There are a number of reasons that may explain why the results of the second assessment are different from those of the first. First, there is substantial non-random sample attrition in the second assessment period, as noted above. Second, most students took their final ThinkLink probe between May 23rd and June 3rd, the last day of school. For these students, the probe was the sixth standardized test that the students had taken since January. Each of these tests asked similar questions. Students may have grown tired of taking these repetitive tests and begun to take them less seriously. Third, students may not have taken the exam seriously because it was so close to the end of the school year. For example, we received anecdotal reports from some tutors that students were finishing the probe in less than five minutes because they were anxious to attend end-of-the-year field day activities. This includes some students who were a part of the student only $90 treatment. Fourth, as previously mentioned, many end-of-the-year activities interfered with the tutors' schedules in the month of May, substantially reducing the amount of treatment the students received. These activities include the final ThinkLink exam which took two weeks to administer, field trips, and outdoor field days and barbeques.

Finally, for the first assessment, the test result was the last standard to be evaluated, and since the exam was conducted and graded by computer, results were known the moment the exam was completed. Students were made aware by their tutors that if they passed the testing standard and met all other standards, they would immediately be given their reward. However, grades for the second assessment were not available until after the school year had concluded. Accordingly, they were not available at the time the students took the test for the second assessment, so we were unable to pay students at the conclusion of the exam. Students were made aware beforehand that payment would be mailed to those who earned a reward once we had information about their final grades on June 6th, three days after the end of the school year. Since the final testing began as early as May 5th, some students had to wait over a month to receive payment; most had to wait approximately two weeks.

IV. Conclusion

A number of previous studies find that combining policies can harness complementarities between inputs in the education production function, resulting in greater impact than the sum of the individual policies. But to date, no study that we are aware of has examined a situation often faced by policy makers: whether a limited budget is best spent on one policy or spread among them. We contribute to this pursuit by examining whether limited resources are better allocated when dedicated to improving the effort of only one of students, parents, or tutors in the education production function or spread among them.

While we make no claims that our result might generalize to other important policy domains, we can conclude that in this particular case, large rewards for an individual input have bigger effects than smaller rewards for multiple inputs. The large rewards have a substantial and robust impact on student performance on the incentivized test. There are no statistical differences between coefficients; hence, there is no evidence that it matters which party receives the reward.

Pure redistribution can explain why *student only $90* and *parent only $90* might have the same effect. For example, incentivized parents might have promised their student that they would give her the money if she earned the reward. Indeed, at the pizza parties following the first assessment where parents were paid in cash, we observed many parents giving their reward to their child.

However, incentivizing multiple parties with the same total reward shared among the inputs may reduce the effectiveness of the reward. Keeping the budget constant creates two factors that may cause the effect of incentivizing multiple inputs to diverge from the effect of incentivizing a single input. While complementarities may be harnessed, the magnitude of the individual effects may be smaller since the rewards for each input are smaller. Our results are most consistent with the conclusion that any improvements resulting from complementarities are overwhelmed by the impact of reduced effort from the individual inputs. Improvements are smaller when multiple parties are incentivized, and are very small and statistically insignificant when the incentive payments are divided among all three inputs. Hence, from a policy perspective, we can conclude that given a certain budget, it is better to incentivize individual parties than to split the money between multiple parties.

Our study may be the first to raise the issue of whether targeting a limited budget at one input (or policy) is more or less effective than spreading that same budget around multiple input or policies. We contend that this general issue has crucial policy relevance since school districts frequently work under constrained budgets. However, we consider the issue only in the limited context of the provision of financial incentives. While these incentives may change the optimal effort level exerted by different inputs in similar ways to other policies, we have no evidence of whether our conclusions would hold when considering other policy menus. More substantial and expensive policy initiatives such as the combinations studied by Mbiti et al. (2018) may work together in ways that our study cannot address. Investigations of how to best spend limited budgets

on other available policy interventions, and to best harness complementarities using these policies, remains an important avenue for future research.

Acknowledgements

Many thanks to the administration, principals, staff and faculty of SD 170 in Chicago Heights, IL, without whom this project would have been impossible. Special thanks to Superintendent Tom Amadio, Mary Kay Entsminger, and especially the tutors who participated in the study who went above and beyond their call of duty to help make the study a success. We thank the Kenneth and Anne Griffin Foundation for generous financial support. Alec Brandon, Eran Flicker, Justin Holz, Jennie Huang, Dan Li, Ryan Malitz, and Phuong Ta provided excellent research assistance. All remaining errors are our own.

References

Anderson, M.L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103(484):1481-1495.

Barrow, L. and Rouse, C.E. 2018. "Financial Incentives and Educational Investment: The Impact of Performance-based Scholarships on Student Time Use." *Education Finance and Policy* 13(4): 419-448.

Behnman, J.R., Parker, S.W., Todd, P.E. and Wolpin, K.I. 2015. "Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools." *Journal of Political Economy* 123(2):325-364.

Benjamini, Y., Krieger, A., and Yekutieli, D. 2006. "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika* 93: 491–507.

Das, J., Dercon, S. Habyarimana, J., Krishnan, P., Muralidharan, K., and Sundararaman, V. 2013. "School inputs, household substitution, and test scores." *American Economic Journal: Applied Economics* 5(2):29-57.

De Fraja, G., Oliveira, T. and Zanchi, L. 2010. "Must try harder. Evaluating the role of effort in educational attainment." *Review of Economics and Statistics* 92(3):577-597.

Fryer, R.G. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics* 126(4): 1755-1798.

Fryer, R.G. 2012. "Aligning student, parent, and teacher incentives: Evidence from Houston public schools." NBER working paper no. 17752.

Geng, T. 2018. "The complementarity of incentive policies in education: Evidence from New York City." Unpublished mimeo, Columbia University.

Hanushek, E.A. 2016. "What matters for student achievement." *Education Next* 16(2):18-26.

Harris, D.N., Sass, T.R., and Semykina, A. 2014. "Value-added models and the measurement of teacher productivity." *Economics of Education Review* 38:9–23.

Houtenville, A.J., and Conway, K.S. 2008. "Parental effort, school resources, and student achievement." *Journal of Human Resources* 43(2):437-453.

Imas, A. 2014. "Working for the 'warm glow': On the benefits and limits of prosocial incentives." *Journal of Public Economics*, 114:14-18.

Jackson, C.K., Johnson, R.C. and Persico, C. 2016. "The effects of school spending on educational and economic outcomes: evidence from school finance reforms." *Quarterly Journal of Economics* 131(1):157-218.

Jackson, C.K., Wigger, C. and Xiong, H. 2018. "Do school spending cuts matter? Evidence from the great recession." NBER working paper No. 24203.

Johnson, R.C. and Jackson, C.K. 2017. "Reducing inequality through dynamic complementarity: evidence from head start and public school spending." NBER Working Paper No. 23489.

Levitt, S.D., List, J.A., Neckermann, S. and Sadoff, S., 2016. The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy* 8(4):183-219.

Levitt, S.D., List, J.A. and Sadoff, S. 2016. "The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment." NBER Working Paper No. 22107.

List, J.A., Livingston, J.A. and Neckermann, S. 2018. "Do financial incentives crowd out intrinsic motivation to perform on standardized tests?" Unpublished mimeo, University of Chicago.

Martorell, P., Miller, T., Santibañez, L. and Augustine, C.H. 2016. "Can incentives for parents and students change educational inputs? Experimental evidence from summer school." *Economics of Education Review* 50:113-126.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C. and Rajani, R. 2017. "Inputs, incentives, and complementarities in primary education: Experimental evidence from Tanzania." Unpublished mimeo, University of Virginia.

OECD, 2012. "Does money buy strong performance in PISA?" *PISA in Focus* 13:1-4.

Rothstein, J. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1):175–214.

Schmid, F. and Trede, M. 1995. "A distribution free test for the two sample problem for general alternatives." *Computational Statistics & Data Analysis*, 20(4):409-419.

Figure 1. Experiment I: Impact of treatments on probe improvement

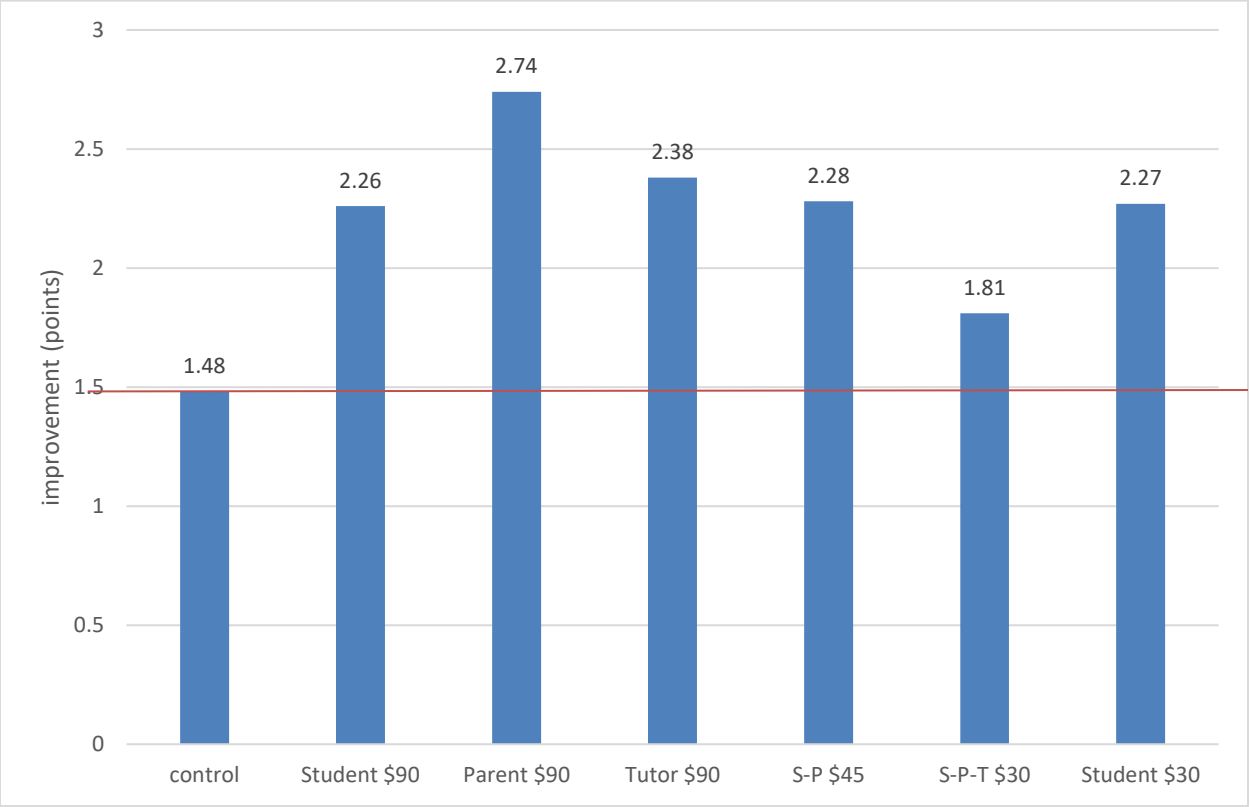

Figure 2. Experiment I: Empirical CDFs of probe score improvement

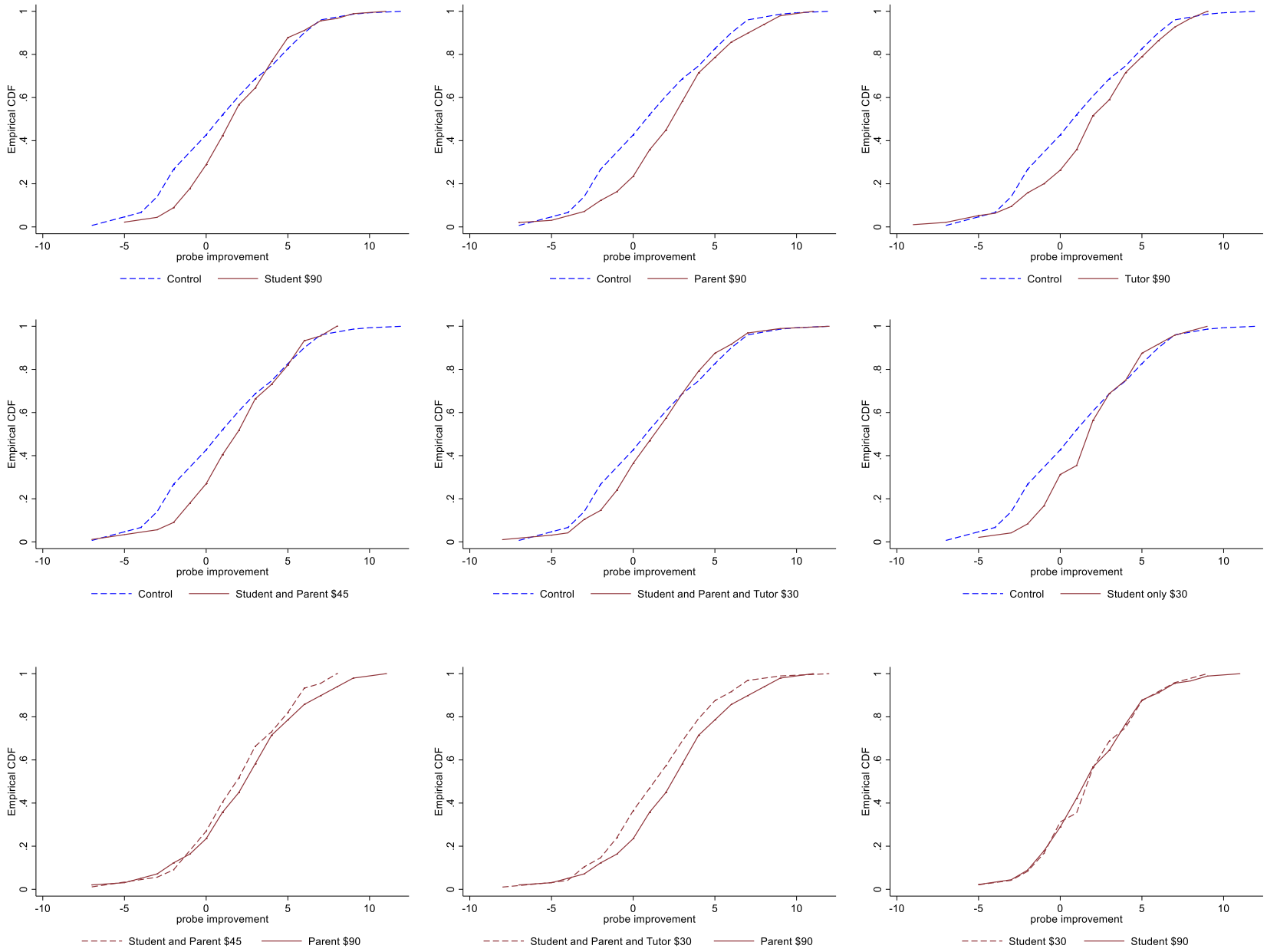

Table 1. Summary statistics by treatment group: baseline assessment and student characteristics

	Control	Student only $90	Parent only $90	Tutor only $90	Student and Parent $45 each	All three $30 each	Student only $30
	N = 151	N = 96	N = 100	N = 101	N = 89	N = 98	N = 48
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Standardized baseline exam	-0.06	0.18	0.08	0.02	-0.06	0.06	-0.21
	(0.97)	(1.04)	(1.05)	(0.90)	(1.01)	(0.90)	(1.05)
Percent of easy questions correct	45.51	49.57	50.39	48.20	41.94	47.54	34.67**
	(23.28)	(22.94)	(22.80)	(24.01)	(22.58)	(24.03)	(19.07)
Percent of moderate questions correct	37.95	44.70	42.04	41.63	39.14	40.40	37.79
	(21.49)	(23.42)	(22.53)	(18.26)	(21.46)	(20.33)	(18.06)
Percent of difficult questions correct	38.39	34.89	39.95	36.42	43.14	33.20	35.60
	(24.09)	(21.94)	(23.07)	(21.76)	(21.50)	(19.83)	(22.48)
Reading or math, 1 = reading	0.66	0.60	0.63	0.71	0.64	0.56	0.67
	(0.48)	(0.49)	(0.49)	(0.45)	(0.48)	(0.50)	(0.48)
Gender, 1 = female	0.46	0.44	0.52	0.56	0.47	0.49	0.50
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.51)
Reduced or free lunch, 1 = yes	0.96	0.81**	0.85**	0.88	0.87	0.94	1.00**
	(0.20)	(0.40)	(0.36)	(0.33)	(0.34)	(0.23)	(0.00)
African-American, 1 = yes	0.30	0.30	0.24	0.36	0.36	0.25	0.29
	(0.46)	(0.46)	(0.43)	(0.48)	(0.48)	(0.43)	(0.46)
Hispanic, 1 = yes	0.58	0.38	0.39	0.32**	0.28***	0.53	0.69
	(0.50)	(0.49)	(0.49)	(0.47)	(0.45)	(0.50)	(0.47)
Number of meetings with tutor per week	3.92	3.31*	3.48	3.61	3.37	3.46	5.00***
	(1.29)	(1.24)	(1.16)	(1.20)	(1.47)	(1.24)	(0.00)
First assessment attrition	1	1	4	8	0	2	0
First assessment attrition (percent)	0.66	1.00	4.17	7.92	0.00	2.04	0.00
Second assessment attrition	15	12	14	13	13	11	n/a
Second assessment attrition (percent)	15.63	12.00	14.58	12.87	14.61	11.22	

Note: The table reports means and standard deviations in parentheses. The asterisks indicate statistical significance from the control group at 10/5/1 percent level calculated using robust standard errors clustered by tutor group and student. All but the *student only $30* treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *parent only $90* treatment, students in the *student only $90* and *student only $30* treatments, and tutors in the *tutor only $90* treatment. Both students and parents received incentives in the *student and parent $45 each* treatment while everyone received incentives in the *all three $30 each* treatment. First assessment attrition reports the number of students who took the baseline assessment, but did not take the first assessment. Second assessment attrition reports the number of students who took the first assessment, but did not take the second assessment. Baseline exam scores are standardized within sample.

Table 2. Experiment I: Distribution shift tests

	Control (1)	Student $90 (2)	Parent $90 (3)	Tutor $90 (4)	S-P $45 (5)	S-P-T $30 (6)
Control						
Student $90	0.018 **					
Parent $90	0.005 ***	0.447				
Tutor $90	0.007 ***	0.503	0.849			
S-P $45	0.055 *	0.618	0.202	0.246		
S-P-T $30	0.333	0.086 *	0.029 **	0.061 *	0.277	
Student $30	0.211	0.388	0.167	0.222	0.482	0.662

Table 3. Experiment I: Impacts of treatments on test performance

	Probe	Easy	Moderate	Difficult	2012 ISAT
	(1)	(2)	(3)	(4)	(5)
Student only $90	0.330	6.619	3.436	8.987	0.343
(standard error)	(0.132)	(2.640)	(3.583)	(4.280)	(0.130)
[FDR corrected *p*-value]	[0.041]	[0.017]	[1.000]	[0.276]	[0.042]
Parent only $90	0.395	9.287	4.573	5.909	0.053
	(0.152)	(2.768)	(3.357)	(4.014)	(0.121)
	[0.041]	[0.007]	[1.000]	[0.545]	[0.700]
Tutor only $90	0.333	8.087	0.603	3.020	0.255
	(0.148)	(2.783)	(3.918)	(4.455)	(0.155)
	[0.041]	[0.011]	[1.000]	[1.000]	[0.254]
Student and parent $45 each	0.247	5.315	3.849	-0.279	0.193
	(0.141)	(2.718)	(3.387)	(4.125)	(0.166)
	[0.064]	[0.040]	[1.000]	[1.000]	[0.338]
All three $30 each	0.077	2.445	-3.312	2.276	0.011
	(0.167)	(2.781)	(3.033)	(4.135)	(0.171)
	[0.347]	[0.179]	[1.000]	[1.000]	[0.700]
Student only $30	0.041	-0.070	-1.687	-1.195	
	(0.260)	(5.103)	(6.923)	(5.332)	
	[0.412]	[0.492]	[1.000]	[1.000]	
N	645	599	599	599	289
Adjusted R^2	0.269	0.247	0.216	0.158	0.469

Note: The table reports coefficient estimates over robust standard errors clustered by tutor group and student in parentheses. Inference is based on *p*-values, reported in brackets, which are adjusted for multiple hypothesis testing of the effects of the six treatment groups. The adjustment controls the false discovery rate following the two-stage sharpened procedure proposed in Benjamini, Krieger, and Yekutieli (2006) and reviewed by Anderson (2008). All but the *student only $30* treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *parent only $90* treatment, students in the *student only $90* and *student only $30* treatments, and tutors in the *tutor only $90* treatment. Both students and parents received incentives in the *student and parent $45 each* treatment while everyone received incentives in the *all three $30 each* treatment. Probes and grades are standardized within sample. The Easy, Moderate, and Difficult columns represent regressions with the percent of easy, moderate, or difficult questions answered correctly on the first assessment as the dependent variable, respectively. All regressions also control for tutor fixed effects, grade level, ethnicity, gender, reduced-lunch status, the subject in which the student was tutored, dummy variables indicating whether the student was in multiple treatment groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in the experiment only once are the omitted category), whether the parent received mail, and the number of meetings with the tutor per week. Probe, Easy, Moderate, and Difficult use the respective score on the Baseline Assessment as its baseline.

Table 4. Experiment I: Control variables

	(1)	(2)	(3)
Student only $90	0.348	0.311	0.330
(standa rd error)	(0.134)	(0.129)	(0.132)
[FDR corrected *p*-value]	[0.025]	[0.048]	[0.041]
Parent only $90	0.440	0.383	0.395
	(0.153)	(0.142)	(0.152)
	[0.025]	[0.048]	[0.041]
Tutor only $90	0.317	0.303	0.333
	(0.150)	(0.141)	(0.148)
	[0.036]	[0.051]	[0.041]
Student and parent $45 each	0.284	0.249	0.247
	(0.130)	(0.138)	(0.141)
	[0.036]	[0.067]	[0.064]
All three $30 each	0.206	0.113	0.077
	(0.156)	(0.161)	(0.167)
	[0.080]	[0.307]	[0.347]
Student only $30	0.028	0.064	0.041
	(0.199)	(0.276)	(0.260)
	[0.286]	[0.373]	[0.412]
Baseline	Yes	Yes	Yes
Characteristics	No	No	Yes
Tutor FE	No	Yes	Yes
Grade Level FE	No	Yes	Yes
N	666	666	645
Adjusted R^2	0.146	0.246	0.269

Note: The table reports coefficient estimates over robust standard errors clustered by tutor group and student in parentheses. Inference is based on *p*-values, reported in brackets, which are adjusted for multiple hypothesis testing of the effects of the six treatment groups. The adjustment controls the false discovery rate following the two-stage sharpened procedure proposed in Benjamini, Krieger, and Yekutieli (2006) and reviewed by Anderson (2008). All but the *student only $30* treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *parent only $90* treatment, students in the *student only $90* and *student only $30* treatments, and tutors in the *tutor only $90* treatment. Both students and parents received incentives in the *student and parent $45 each* treatment while everyone received incentives in the *all three $30 each* treatment. First assessment attrition reports the number of students who took the baseline assessment, but did not take the first assessment. Second assessment attrition reports the number of students who took the first assessment, but did not take the second assessment. Baseline exam scores are standardized within sample.

Table 5. Experiment I: Impact of treatments by student characteristics

	Subject		Gender		Ethnicity		
	Reading	Math	Female	Male	Black	Hispanic	White
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student only $90	0.469	0.353	0.267	0.341	0.580	0.155	0.095
(standard error)	(0.172)	(0.281)	(0.199)	(0.182)	(0.265)	(0.203)	(0.386)
[FDR corrected *p*-value]	[0.038]	[0.419]	[0.220]	[0.137]	[0.211]	[0.889]	[1.000]
Parent only $90	0.440	0.505	0.315	0.543	0.443	0.376	-0.051
	(0.206)	(0.283)	(0.198)	(0.201)	(0.318)	(0.213)	(0.385)
	[0.087]	[0.419]	[0.174]	[0.044]	[0.373]	[0.301]	[1.000]
Tutor only $90	0.384	0.337	0.393	0.394	0.182	0.473	-0.062
	(0.194)	(0.290)	(0.185)	(0.219)	(0.266)	(0.243)	(0.383)
	[0.087]	[0.419]	[0.114]	[0.137]	[0.589]	[0.301]	[1.000]
Student and parent $45 each	0.216	0.473	0.405	0.204	0.348	0.184	-0.076
	(0.147)	(0.340)	(0.187)	(0.187)	(0.213)	(0.300)	(0.320)
	[0.107]	[0.419]	[0.114]	[0.260]	[0.347]	[0.889]	[1.000]
All three $30 each	0.374	-0.177	0.151	0.093	-0.100	0.017	0.045
	(0.228)	(0.272)	(0.210)	(0.225)	(0.270)	(0.235)	(0.341)
	[0.107]	[0.419]	[0.370]	[0.599]	[0.746]	[1.000]	[1.000]
Student only $30	0.138	-0.469	-0.015	-0.092	0.085	0.185	
	(0.254)	(0.329)	(0.396)	(0.288)	(0.529)	(0.342)	
	[0.208]	[0.419]	[0.478]	[0.599]	[0.774]	[0.889]	
Observations	405	240	322	323	186	287	172
Adjusted R^2	0.303	0.296	0.315	0.328	0.415	0.360	0.371

Note: The table reports coefficient estimates over robust standard errors clustered by tutor group and student in parentheses. Inference is based on *p*-values, reported in brackets, which are adjusted for multiple hypothesis testing of the effects of the six treatment groups. The adjustment controls the false discovery rate following the two-stage sharpened procedure proposed in Benjamini, Krieger, and Yekutieli (2006) and reviewed by Anderson (2008). All but the *student only $30* treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *parent only $90* treatment, students in the *student only $90* and *student only $30* treatments, and tutors in the *tutor only $90* treatment. Both students and parents received incentives in the *student and parent $45 each* treatment while everyone received incentives in the *all three $30 each* treatment. Probe scores were standardized using our sample. All outcomes use the first probe as their baseline. All regressions also control for tutor fixed effects, grade level, ethnicity, gender, reduced-lunch status, the subject in which the student was tutored, dummy variables indicating whether the student was in multiple treatment groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in the experiment only once are the omitted category), whether the parent received mail, and the number of meetings with the tutor per week.

Table 6. Experiment I: Effect of treatments on other standards

	Grade	> 2 Unexcused	> 0 Suspensions	Threshold
	(1)	(2)	(3)	(4)
Student only $90	-0.173	-0.016	-0.031	0.103
(standard error)	(0.099)	(0.035)	(0.024)	(0.086)
[FDR corrected *p*-value]	[0.924]	[1.000]	[1.000]	[0.538]
Parent only $90	-0.009	0.004	0.004	0.153
	(0.074)	(0.038)	(0.034)	(0.089)
	[1.000]	[1.000]	[1.000]	[0.538]
Tutor only $90	0.017	0.014	-0.014	0.090
	(0.080)	(0.038)	(0.024)	(0.075)
	[1.000]	[1.000]	[1.000]	[0.538]
Student and parent $45 each	-0.071	0.062*	0.012	0.032
	(0.069)	(0.037)	(0.032)	(0.070)
	[1.000]	[1.000]	[1.000]	[0.538]
All three $30 each	-0.049	0.018	0.015	-0.000
	(0.071)	(0.029)	(0.027)	(0.077)
	[1.000]	[1.000]	[1.000]	[0.538]
Student only $30	0.138	-0.034	0.025	0.179
	(0.148)	(0.033)	(0.020)	(0.138)
	[1.000]	[1.000]	[1.000]	[0.538]
N	631	654	654	621
Adjusted R^2	0.737	0.238	0.171	0.211

Note: The table reports coefficient estimates over robust standard errors clustered by tutor group and student in parentheses. Inference is based on *p*-values, reported in brackets, which are adjusted for multiple hypothesis testing of the effects of the six treatment groups. The adjustment controls the false discovery rate following the two-stage sharpened procedure proposed in Benjamini, Krieger, and Yekutieli (2006) and reviewed by Anderson (2008). All but the *student only $30* treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *parent only $90* treatment, students in the *student only $90* and *student only $30* treatments, and tutors in the *tutor only $90* treatment. Both students and parents received incentives in the *student and parent $45 each* treatment while everyone received incentives in the *all three $30 each* treatment. Grades are converted to numerical equivalents and are standardized within sample by grade and subject. Threshold is a dummy variable that equals 1 if students met all four academic and behavioral standards required to earn the incentive payment. All regressions also control for tutor fixed effects, grade level, ethnicity, gender, reduced-lunch status, the subject in which the student was tutored, dummy variables indicating whether the student was in multiple treatment groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in the experiment only once are the omitted category), whether the parent received mail, and the number of meetings with the tutor per week. Grade controls for the student's baseline grades in the trimester before the experiment began.

Table 7. Experiment II: Impact of treatments on all outcome variables

	Probe	Easy	Moderate	Difficult	Grade	> 2 Unexcused	> 0 Suspensions	Threshold
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Student only $90	0.061	-5.523	0.756	-3.062	0.193	0.047	-0.028	-0.019
(standard error)	(0.146)	(3.210)	(4.077)	(6.989)	(0.117)	(0.045)	(0.030)	(0.069)
[FDR corrected *p*-value]	[1.000]	[0.740]	[0.571]	[1.000]	[0.672]	[1.000]	[1.000]	[1.000]
Parent only $90	0.099	-0.610	3.710	4.810	-0.030	0.002	0.012	-0.049
	(0.159)	(4.124)	(4.053)	(6.670)	(0.134)	(0.045)	(0.029)	(0.071)
	[1.000]	[1.000]	[0.370]	[0.891]	[0.672]	[1.000]	[1.000]	[1.000]
Tutor only $90	0.257	3.477	5.234	11.888	-0.176	0.067	0.011	0.039
	(0.154)	(4.187)	(4.249)	(6.149)	(0.150)	(0.049)	(0.030)	(0.071)
	[0.924]	[1.000]	[0.279]	[0.361]	[0.672]	[1.000]	[1.000]	[1.000]
Student and parent $45 each	0.130	-4.471	7.973	8.439	-0.030	-0.053	-0.004	0.091
	(0.161)	(4.151)	(4.047)	(7.054)	(0.123)	(0.051)	(0.039)	(0.077)
	[1.000]	[1.000]	[0.109]	[0.866]	[0.672]	[1.000]	[1.000]	[0.880]
All three $30 each	0.175	0.519	10.323	3.798	0.140	-0.020	-0.024	0.133
	(0.143)	(3.359)	(3.833)	(5.180)	(0.115)	(0.046)	(0.031)	(0.073)
	[0.924]	[1.000]	[0.037]	[0.891]	[0.672]	[1.000]	[1.000]	[0.504]
N	462	399	399	399	540	550	550	461
Adjusted R^2	0.353	0.211	0.274	0.310	0.388	0.338	0.144	0.164

Note: The table reports coefficient estimates over robust standard errors clustered by tutor group and student in parentheses. Inference is based on *p*-values, reported in brackets, which are adjusted for multiple hypothesis testing of the effects of the six treatment groups. The adjustment controls the false discovery rate following the two-stage sharpened procedure proposed in Benjamini, Krieger, and Yekutieli (2006) and reviewed by Anderson (2008). All treatments had bi-monthly monetary incentives for student performance. Parents received incentives in the *parent only $90* treatment, students in the *student only $90* treatment, and tutors in the *tutor only $90* treatment. Both students and parents received incentives in the *student and parent $45 each* treatment while everyone received incentives in the *all three $30 each* treatment. Probes and grades are standardized within sample. The Easy, Moderate, and Difficult columns represent regressions with the percent of easy, moderate, or difficult questions answered correctly on the first assessment as the dependent variable, respectively. Threshold is a dummy variable that equals 1 if students met all four academic and behavioral standards required to earn the incentive payment. Probe, Easy, Moderate, and Difficult use the respective score on the baseline assessment as its baseline, while Grade controls for the student's baseline grades in the trimester before the experiment began. All regressions also control for tutor fixed effects, grade level, ethnicity, gender, reduced-lunch status, the subject in which the student was tutored, dummy variables indicating whether the student was in multiple treatment groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in the experiment only once are the omitted category), whether the parent received mail, and the number of meetings with the tutor per week.

Table 8. Pooled samples from Experiment I and Experiment II:
Impacts of treatments on test performance

	Probe	Easy	Moderate	Difficult
	(1)	(2)	(3)	(4)
Student only $90	0.201	1.226	2.507	4.687
(standard error)	(0.112)	(2.190)	(2.981)	(3.468)
[FDR corrected *p*-value]	[0.139]	[1.000]	[1.000]	[0.373]
Parent only $90	0.248	4.660	3.545	4.959
	(0.125)	(2.819)	(2.822)	(3.538)
	[0.137]	[0.325]	[1.000]	[0.373]
Tutor only $90	0.293	6.117	1.782	7.255
	(0.113)	(2.574)	(2.953)	(3.613)
	[0.064]	[0.092]	[1.000]	[0.370]
Student and parent $45 each	0.172	1.450	5.797	3.840
	(0.110)	(2.493)	(2.539)	(3.694)
	[0.169]	[1.000]	[0.153]	[0.427]
All three $30 each	0.125	1.195	2.370	3.511
	(0.129)	(2.681)	(2.338)	(2.843)
	[0.218]	[1.000]	[1.000]	[0.373]
Student only $30	0.006	-1.139	-3.367	-1.667
	(0.246)	(5.090)	(7.116)	(6.292)
	[0.486]	[1.000]	[1.000]	[0.483]
N	1,106	997	997	997
Adjusted R^2	0.237	0.160	0.141	0.142

Note: The table reports coefficient estimates over robust standard errors clustered by tutor group and student in parentheses. Inference is based on *p*-values, reported in brackets, which are adjusted for multiple hypothesis testing of the effects of the six treatment groups. The adjustment controls the false discovery rate following the two-stage sharpened procedure proposed in Benjamini, Krieger, and Yekutieli (2006) and reviewed by Anderson (2008). All but the *student only $30* treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *parent only $90* treatment, students in the *student only $90* and *student only $30* treatments, and tutors in the *tutor only $90* treatment. Both students and parents received incentives in the *student and parent $45 each* treatment while everyone received incentives in the *all three $30 each* treatment. Probes and grades are standardized within sample. The Easy, Moderate, and Difficult columns represent regressions with the percent of easy, moderate, or difficult questions answered correctly on the first assessment as the dependent variable, respectively. All regressions also control for tutor fixed effects, grade level, ethnicity, gender, reduced-lunch status, the subject in which the student was tutored, dummy variables indicating whether the student was in multiple treatment groups (an indicator for being in the same treatment group twice and an indicator for being in different treatment groups; students who were in the experiment only once are the omitted category), whether the parent received mail, the number of meetings with the tutor per week, and a dummy variable indicating which experiment the observation comes from (I or II). Probe, Easy, Moderate, and Difficult use the respective score on the Baseline Assessment as its baseline.

Appendix A. Example Letter to Students

Dear Student,

We are excited to be able to conduct this study with you. You will have the chance to earn money if you do several things:

1. You must have no more than two unexcused absences during an assessment period.
2. You must have had zero all-day suspensions (either in school or out of school) during an assessment period.
3. Your grade in either reading or math, depending on the subject that you are working on with your tutor, must either remain where it was on your last report card or improve. It must not get worse. It also must be above a grade of F.
4. Your must have an improved score on a Discovery Education Thinklink exam in either reading or math, depending on the subject that you are working on with your tutor.

If all of these standards are met, **you will be paid $90.**

The evaluations will occur two times over the course of the rest of the school year, so you will have a chance to earn this reward two different times. The dates of the evaluations are based on when report cards are issued:

March 17th, 2011
June 6th, 2011

Thank you very much for participating!

Appendix B. Example Letter to Parents

Dear Parent,

We are excited to be able to conduct this study on the academic achievement of elementary school children with you. As part of the study, you, your child, and your child's reading or math tutor may have the chance to earn money if your child, FULL NAME HERE, meets a set of behavioral and achievement standards.

The standards that must be met for you to receive the reward are:

1. Each Friday, the tutor will give your child a package of materials or an assignment to work on together with you. You must complete the materials or assignment with your student, and keep a record of what material has been covered each week on the sheet that we will provide to you. Any completed materials and the record sheet should be sent back to school and returned by your child to their tutor a week later, on the Friday after you receive them.
2. Your child must have no more than two unexcused absences during an assessment period.
3. The student must have had zero all-day suspensions (either in school or out of school) during an assessment period.
4. Your child's grade in the relevant subject (either reading or math, depending on the subject that the tutor is teaching your child) must either remain at its previous level or improve. It must not decline. It also must be above a grade of F.
5. Your child must have an improved score on a Discovery Education Thinklink exam in the relevant subject (reading or math).

If all of these standards are met, **you will be paid $45**. Your child will also be paid $45 if he or she avoids unexcused absences and all-day suspensions as mentioned, maintains his or her grade in the relevant class, and improves his or her score on the Discovery Education Thinklink exam in the relevant subject.

The evaluations will occur two times over the course of the rest of the school year, so you will have a chance to earn rewards on two different occasions. The dates of the evaluations are based on when report cards are issued:

March 17th, 2011
June 6th, 2011

Thank you very much for participating, If you have any questions, please do not hesitate to contact me. My contact information is:

Jeff Livingston
Email: jlivingston@bentley.edu
Phone: (XXX) XXX-XXXX

Appendix C. Example Letter to Tutors

Hi Tutors,

We are excited to be able to conduct this study on the academic achievement of elementary school children with you. As part of the study, you, your students, and the students' parents may have the chance to earn extra money if the student meets a set of behavioral and achievement standards.

Here is how the study will work. Each of your groups of students will be randomly assigned to one of six possible incentive programs. These programs include:

1) Only you are eligible for a reward.
   If all of the standards are met, ***you*** will be paid **$90**.

2) Only the student is eligible for a reward.
   If all of the standards are met, the ***student*** will be paid **$90**.

3) Only the student's parents are eligible for a reward.
   If all of the standards are met, the ***student's parents*** will be paid **$90**.

4) Both the student and his or her parents are eligible for a reward.
   If all of the standards are met, the ***student*** and the ***student's parents*** will be paid **$45 each.**

5) Both you, the student and the student's parents are eligible for a reward.
   If all of the standards are met, ***you***, the ***student*** and the ***student's parents*** will be paid **$30 each**.

6) Nobody is eligible for a reward.

Your group assignments to the incentive programs are described in the attached letter. Every student in one of your groups will be part of the same incentive program. So, for example, if you have a group of six students that you meet with, that group is assigned to incentive program 1, and the standards below are met for all six students, you would be paid $540. If three of the six students meet the standards, then you would be paid $270.

The standards that must be met for you to receive the reward are as follows:

1. Create a package of materials on that week's areas covered for the student to bring home and work on with their parent(s). This should be done at the end of each week, **beginning the week of January 10th, 2011**. Your materials should be sent home with the students on Friday, and should consist of a review of the material you went over with them in your sessions that week.

   Important note: **this should only be done for students whose parents are getting a financial incentive**. **So, this should be done for your student groups that are assigned to incentive program 3, 4 or 5 only.** As long as the materials are provided to the parents and a copy is given to us, this standard is met.

   You do not need to collect the materials back from the parents and keep track of whether they actually used them if you do not want to. Keeping a record of what was done and returning the materials to me will be one of the conditions that the parents have to meet in order to receive their incentive payment.

2. Keep a record of what material has been covered with each group of students each week. As long as a record is provided to me each week, this standard is met.
3. The student must have had no more than two unexcused absences since the last evaluation.
4. The student must have had zero out of school suspensions since the last evaluation.
5. The student's grade in the relevant subject (Reading or Math) must either remain at its previous level or improve. It must not decline. It also must be above a grade of F.
6. For third graders through eighth graders, the student must have an improved score on a Discovery Education Thinklink probe exam in the relevant subject (reading or math). For first and second graders, improvement must be shown on a similar exam.

The evaluations will occur two times over the course of the rest of the school year, so you will have a chance to earn rewards on two different occasions. The dates of the evaluations are based on when report cards are issued:

March 17th
June 6th

Thank you very much for participating, If you have any questions, please do not hesitate to contact me. My contact information is:

Jeff Livingston
Email: jlivingston@bentley.edu
Phone: (XXX) XXX-XXXX