Harnessing Complementarities in the Education Production Function*

John A. List
University of Chicago and NBER

Jeffrey A. Livingston[a]
Bentley University

Susanne Neckermann
ZEW and University of Mannheim

Abstract

Studies which seek to estimate components of the education production function, especially those that employ the "value-added" approach, almost universally assume that the production function is linear and additively separable in its inputs. This strict functional form assumes that there are no complementarities between inputs, though there are compelling intuitive reasons to think they might exist. This study conducts a randomized field experiment to evaluate whether complementarities between students, their parents, and tutors who aid the students in specific subjects can be harnessed using financial incentives. No evidence emerges in support of this hypothesis. The results suggests that a given level of financial resources have a far greater impact on student achievement when directed at just one input rather than being spread over multiple inputs. The evidence also suggests that students do not reach their effort frontier on standardized tests in which they have no personal stake, calling into question their usefulness as a measure of student achievement and as an evaluation tool for policy makers.

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

* 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 and Phuong Ta provided excellent research assistance. All remaining errors are our own.

I. Introduction

There is an extremely rich literature regarding the nature of the *education production function* (EPF), that is, the manner by which inputs from students, parents, teachers, schools, and many other sources are translated into a student's academic achievement. Hanushek (2002) reviews much of the extant literature. Recent examples of empirical studies include Krueger (1999), Angrist and Lavy (1999) and Hoxby (2000) who each present evidence on the effect of class size; Hanushek et al. (2003) who address peer effects; Rivkin et al. (2005), who focus on identifying variation in teacher quality; De Fraja et al. (2010) who examine the effects of student, parent and teacher effort; and Houtenville and Smith Conway (2008) who concentrate on the importance of parental involvement.

Despite this richness, one critical issue that the extant literature has largely left in the background is whether inputs in the EPF might be complements. This possibility is intuitively appealing. For example, increased teacher effort might be more effective if parents are also more supportive of their child's academic pursuits. While some studies such as Todd and Wolpin (2003) have recognized that complementarities may exist, most empirical research has assumed that the EPF takes a functional form that precludes the possibility of complementary inputs. In particular, studies that employ the popular "value added" approach typically assume that the production function is linear and additively separable in its inputs, thereby restricting inputs to be perfect substitutes (see, e.g., the papers cited above as well as Aaronson et al. 2007, Jacob and Lefgren 2008 and Rothstein 2010).[1]

[1] Houtenville and Smith Conway (2008) do note that they attempted regressions which included interactions between parental effort and school resources. Although they do not report the results, they do comment that they found almost no significant interaction effects, suggesting the two inputs are indeed substitutes.

This assumption has vital policy implications. If correct, policy makers can target one specific input and expect to see significant returns in student achievement. However, if strong complementarities between inputs exist, strategies which target multiple inputs should be more successful, while strategies that focus on only one input might be doomed to failure. For example, providing teachers with merit pay increases might have few effects unless parents are supportive of their child's studies.

In this paper, we offer a field experiment specifically designed to explore the effects of potential complementarities between inputs in the EPF.[2] In so doing, we also present one of the first studies to investigate whether varying the input that is incentivized results in different impacts on student achievement.[3] We leverage a program in Chicago Heights, IL, elementary and middle schools made up of largely low-income and minority students with low achievement. This program used a grant financed by federal stimulus funds to hire tutors to work with students who the school administrations identified as needing extra help in either reading or math. Our randomized field experiment provides financial incentives to three key inputs into a student's education: the students themselves, their parents, and the students' tutors. Either a single input or a combination of these three inputs are provided incentives to meet (or to aid the student in meeting) a variety of academic and behavioral standards. If complementarities between inputs

[2] Fryer (2012) also presents an experiment designed to harness complementarities by aligning the incentives of all inputs – students, parents and teachers. However, there are crucial differences in the approaches taken by his study and ours. While he incentivizes inputs, e.g. by providing rewards for completing more math objectives in a software program designed to teach math skills, we directly incentivize outputs such as grades and test scores. Also, the Fryer study includes only one treatment group where all inputs are given incentives at the same time, while we also have treatment groups that give incentives to only a subset of the inputs (including individual inputs).

[3] In a closely related study whose design was used as much of the inspiration for the design of our approach, Levitt, List and Sadoff (2011) also vary the reward recipient, incentivizing either students or parents (but never both, as we do here).

are sufficiently important, larger improvements should be observed when providing multiple parties with incentives than when only one input is addressed.

However, even in a controlled experiment, it is difficult to isolate the effect of complementarities and verify their existence.[4] We therefore take a more pragmatic approach and consider whether complementarities are strong enough to influence policy. The experiment is designed to answer the following question: given a set budget that policy makers have at their disposal, what distribution of incentives results in the largest improvement in student achievement? 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 our achievement standards are met. But when both the student and their parent are incentivized, each is paid $45 if the student meets the standards.

Our results provide little evidence that complementarities between inputs are important enough to influence policy. When only one input is incentivized, we observe similar gains regardless of who receives the incentive – the student, the parent or the tutor. The effect sizes are substantial, ranging between 0.3 and 0.5 standard deviations. However, when the same budget is used to incentivize multiple inputs, the gains relative to control are smaller and statistically insignificant.

Our experimental design also allows us to examine whether incentivized achievement improves human capital or merely encourages students to exert more effort on the exam that is used to measure their progress. One standard which the students must meet in order for the

[4] Fryer (2012) reviews several of the competing mechanisms that may cause the kind of financial incentives we employ here to drive behavior in one direction or the other. They include harnessing complementarities along with the degree to which students, parents and teachers are motivated; how the future is discounted; a potential lack of information about the returns to education; the ways in which incentives change the optimal allocation of effort by each input, and the crowding out of intrinsic motivation.

incentivized parties to be paid is to improve on a standardized test that we designed. This test served no purpose outside the experiment, and the results were not reported to the school district, so the only incentive to improve on the test was the financial incentive we provided. These tests were designed to assess the same skills and knowledge that official standardized tests examine, and drew the questions from test banks created by the same organization that develops the standardized tests used by the schools. Indeed, the school district administered an official standardized test at approximately the same time as each of our tests. Should students take both tests seriously and experience true gains in knowledge and skills in response to treatment, we would expect similar gains on both the experimental tests and the official standardized tests. However, if improvements are observed only on the tests for which the inputs are incentivized, then students are likely not reaching their effort frontier on tests in which they have no personal stake.

The answer to this question also has crucial policy implications. Standardized tests are now being widely implemented as a measure of the effectiveness of both schools and teachers. This includes measures such as the No Child Left Behind Act of 2001, which withheld federal funding for states who failed to meet minimum achievement standards based on statewide standardized tests, the federal Race to the Top competition which rewarded states for implementing value-added systems of teacher evaluation based on standardized tests, and pushes for teacher merit pay systems which measure teacher effectiveness using such tests. If students only exhibit improvement on tests for which they are incentivized, it calls into question the appropriateness of using standardized tests that have no impact on a student's welfare as an evaluation of a student's academic progress. Students may fail to show improvement merely

because they have no incentive to show what they have learned, not because they are missing the requisite skills. Accordingly, the test would not accurately measure such students' achievement.

This potential problem has gone largely unrecognized by academics and policy makers alike. Only a handful of studies that we are aware of have explored whether standardized tests accurately measure academic progress. Levitt et al. (2011) show that student test scores increase dramatically when they are given a substantial monetary incentive to improve on the test but are not notified about the incentive until the day of the test. Because the students were unaware of the incentive beforehand, any observed test improvement can only be due to increased effort on the exam, not improved learning beforehand. Hence, students do not perform at their effort frontier in the absence of additional rewards. Corcoran et al. (2011), meanwhile, find substantial variation between teacher effects on outcomes of two standardized tests that were administered at approximately the same time, one of which is used to reward or punish teachers and schools based on the students' progress and the other of which is used only as a diagnostic assessment. As they note, "one would hope that high-stakes decisions about individual teachers are not highly test-dependent." Likewise, one would also hope that such decisions are not made on the basis of tests that do not observe the true extent of a student's improvement.

The results are largely consistent with the conclusion that students in this population do not fully exert themselves on tests which are high stakes for the schools but for which they have no personal stake. The observed gains relative to control on the incentivized exam are absent on the school-administered standardized tests.

The remainder of this study is organized as follows. Section II describes the experimental design and reviews the nature of school district where the experiment was conducted. Section III presents the empirical methodology and discusses the results. Section IV concludes.

II. Experimental Design

Our experiment was conducted in the nine elementary and middle schools in Chicago Heights, IL, a suburb thirty miles south of Chicago. 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 passed both the reading and math portions of the Illinois Standards Achievement Test (ISAT) in 2010, the results of which are applied to the No Child Left Behind Act to identify failing schools.

The district classifies students into three tiers. Tier one students are those who are on track to meet state ISAT standards. Tier two students are judged to be at risk of failing to meet state standards, while tier three students are judged to be severely at risk and in need of intervention. The tutors were hired to work with tier two students. 32 tutors were hired for 100 days at a wage of $100 per day. 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.

Our 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. 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. A total of 581 students, grades Kindergarten through eighth, worked with our 23 tutors. These students were organized into 157 groups.

Our design consists of five treatment groups and one control group. We randomized students into these treatments 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. While conducting the randomization, we blocked on school, tutor, homeroom teacher, subject (reading or math), grade level, gender, race/ethnicity, number of meetings per week the group met with the tutor, and baseline test score when available.[5]

The five treatment groups include an incentive for the tutor only, an incentive for the student only, an incentive for the student's parents only, an incentive for both the student and the parents, and an incentive for all three inputs – the student, the parents, and the tutor. A total of $90 is paid to the incentivized parties if the achievement standards are all met. In the treatments where only one input is incentivized, that input receives the entire $90. In order to judge whether potential complementarities should impact policy, the $90 is split equally among the incentivized parties. So, for example, when all three parties are given the incentive, each earns $30 if the standards are met. This allows us to judge how a given budget can be allocated most efficiently. If complementarities are strong enough, student improvement should be strongest when the money is divided between multiple inputs. However, if they do not exist or are not strong enough to overcome the (potentially) smaller effect on effort that the smaller amount may have, using the budget to incentivize multiple inputs will at best provide no advantage over incentivizing only one input, and at worst will have a smaller impact on student improvement.

The standards students are required to meet are based on those employed by Levitt, List and Sadoff (2011), who examine the impact of monthly financial incentives on the performance of high school students in Chicago Heights. These standards were provided by the school leadership, and are based on what they considered to be the minimum requirements necessary to

[5] 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 each of their student groups were assigned. Including the students of this tutor, 620 students were part of the randomization. Baseline test scores were available for 452 of these students.

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.

We modify these standards to our context. The experiment began on January 10th, 2011 and consisted of two roughly bi-monthly, rather than monthly, assessments.[6] Accordingly, we modify the absence standard to allow two unexcused absences during each assessment period rather than one. Also, the grade and testing standards of Levitt, List and Sadoff (2011) 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. Consequently, our standards were: no more than two unexcused absences and no all day suspensions during each assessment period, the student's grade in the relevant subject had to be above a failing grade of F and at least maintained at its previous level, and the student had to improve by at least one point (out of 20) on the standardized test that we created. The two assessments were independent, so those who failed to earn a reward in the first assessment period were able to do so in the second assessment period, and vice versa.

In addition to these standards, we wanted to provide incentivized 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 group of students who were part of one of the parent incentive groups (*parent only*, *student and parent*, and *student and parent and tutor*) that was designed to be a review of what they had covered that week. The tutors instructed the students to bring these

[6] 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.

assignments home to work on with their parents. Parents then faced the additional requirement of completing these assignments with their child each week, and having their child return it to the tutor.

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,[7] 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.

Our two bi-monthly assessments each occurred at the end of a trimester to coincide with the release of grade cards as closely as possible, so that the grade standard could be assessed and enforced. The first assessment coincided with the release of the second trimester grades on March 17th, 2011. The second assessment concluded with the issuance of the final trimester grade card on June 6th, 2011.

[7] 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.

Conveniently, the beginning of the experiment and each of the assessments occurred at roughly the same time as when the schools administered a standardized test. Chicago Heights students in grades three through eight take the Discovery Education ThinkLink Learning exams four times during the course of the school year. The schools administered the third exam at the beginning of the experiment in January, and the fourth exam near the end of the experiment in May. Students also took the ISAT approximately at the time of the first assessment in March. Each of these exams has a reading, math and science component. Discovery Education designs the ThinkLink exams to test the same skills as the ISAT, and the schools use them as predictors of a student's ISAT scores. The third ThinkLink exam of the year is used as a baseline score to assess improvement on the later school-administered exams that are not incentivized. We judge student improvement at our first and second assessment points by comparing the baseline results to scores on the ISAT exam and the final ThinkLink exam, respectively.

The incentives in our experiment are not based on performance on these official exams, however. Rather, we design our own exams using resources provided by Discovery Education, which make it possible to create ThinkLink "probes" to measure a student's progress at any time. These exams randomly draw questions from a test bank of questions that again are based on the same skills and knowledge that is tested on the official ThinkLink tests as well as the ISAT. Therefore, each of the exams for which we have data – the ThinkLink exams, the ISAT, and our ThinkLink probes – theoretically measure the same thing. A separate probe was created for each grade level (K through 8) and subject (reading and math). The probes consist of 20 questions, are

administered by the tutors,[8] and are taken on a computer. Each probe was administered beginning the week following the official standardized test with which it is paired, so they measure the students' knowledge at roughly the same time. The baseline ThinkLink exam was taken during the week of January 10th 2011; our first probe was taken during the week of January 17th. The ISAT was taken during the week of March 14th 2011; our second probe, used for our first assessment, was taken during the week of March 21st. Finally, the final ThinkLink exam was taken between May 9th and May 23rd 2011; our final probe, used for our second assessment, was taken by most students beginning on May 23rd.[9] Performance on these probes was critical for receiving the rewards, while performance on ThinkLink exams and the ISAT was not.

For the first assessment, grades and information about absences and suspensions were available at the time each student took the probe, and 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,

[8] 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.

[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.

and made sure the parents were aware that the incentive program was continuing and that each student started with a clean slate. We 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.

For the second assessment, immediate payment for the students was not possible because the 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. Results

III.1 Balance on covariates

Table 1 reports the sample means by treatment group for pre-treatment characteristics and for baseline achievement in our sample.[10] The tables indicate significant differences between treatment and control group means, with standard errors clustered by tutor-groups. As expected, there are no statistically significant differences in baseline achievement and very few statistically significant differences in demographic characteristics. The only significant differences are the proportion of females in the *student* treatment and the proportion of Hispanics in *tutor* and *student-parent*. As shown below, including controls for pre-treatment characteristics as well as baseline performance does not alter the results.

III.2 Empirical strategy

[10] The first panel reports probe outcomes for the baseline assessment at the start of the experiment. The second panel reports performance in the standardized tests previous to the ones that coincide with end of assessment periods one and two: ISAT and ThinkLink, as well as grades at the start of the program. The third 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.

While our incentive program is based on a vector of outcomes, for several reasons, the focus of the analysis is on improvement on the ThinkLink probes and the companion official standardized tests that were not incentivized. The district's goal for the tutor program was improvement on ISAT scores, and in general standardized test scores are the most widely relied upon measure of student achievement. Another goal of the study is to compare performance on incentivized standardized tests to performance on tests that lack incentives.

Discovery Education classifies each question on the 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 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. Finally, we compare student performance on the ThinkLink probes to performance on the official standardized test that coincides with the end of the respective assessment period (ISAT for assessment period one and the final ThinkLink exam for assessment period two).

For each of these outcome measures, we estimate a standard value-added model of student achievement. Variants of the following equation are estimated by Ordinary Least Squares:

$$A_{igjrvt} = \alpha A_{igjrvt-1} + \beta_1 T_{jr} + \beta_2 X_i + \beta_3 \gamma_g + \beta_4 \theta_j + \mu_v + \varepsilon_{igjrvt},$$

where A_{igjst} is the achievement of student i in grade g, assigned to tutor j and group r, who sees homeroom teacher v, in assessment period t; $A_{igjrst\text{-}1}$ is the baseline assessment from the previous

period,[11] T_{jr} 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;[12] γ_g, θ_j and μ_v are grade, tutor and teacher fixed effects, respectively; and ε_{igjrvt} measures white noise. Standard errors are clustered by tutor-group, which is the level of randomization.

III.3 Results

Table 2 presents the results for the full set of outcome variables, including all control variables, in the first assessment period.[13] Column 1 reports the effects of the treatments on the overall ThinkLink probe score, standardized by grade and subject (reading or math).[14] The individual input incentives *Student* as well as *Parent* and *Tutor* each have a statistically significant and sizeable positive effect on probe scores, our primary measure of student progress. The individual reward conditions have substantial impact on performance: an increase in test scores ranging from roughly 0.3 to 0.5 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. However, the estimated coefficients on the *Student and*

[11] Exceptions are suspensions and absences, as these data are not currently available for the period before the start of the program.

[12] 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 lunch, a dummy variable indicating whether the student appears in the data more than once because she sees both the reading and math tutor in her school, the percentage of homework returned to the tutor (recorded as zero for students in treatments with no parent incentives), a dummy variable indicating whether the initial mailing was received by the parents, and a dummy variable indicating if the student's parents did not speak English.

[13] Columns 1 through 3 in Table 3 show specifications where we alter the set of control variables. They show that the results reported below are highly robust to changes in the characteristics and types of fixed effects that are included as regressors.

[14] The number of observations falls short of our full sample of 581 students because a handful were absent at the time when either the initial assessment probe or the second assessment probe was administered. These missing test scores leave us with 547 observations.

Parent and *All* treatment indicators are each statistically insignificant, indicating that the gains are much weaker when multiple inputs are incentivized but the reward is smaller.

Interestingly, the incentives have the biggest impact on student performance on the easiest exam questions. Columns 2 through 4 of Table 2 report the results of regressions where the outcome variable is the percentage of easy, moderate and difficult questions that the students answer correctly, respectively.[15] Each individual incentive increases the percentage of easy questions answered correctly by about six to seven percentage points, which again represents a roughly 0.3 standard deviation increase, and again, these effects are not observed in treatments with more than one incentivized input. No such gains are evident on the more difficult questions, with one exception: solely incentivized students see a similar-sized gain in the percentage of the most difficult questions answered correctly. However, it is clear that the majority of the improvement observed comes on the easier questions. One intuitive interpretation of this result is that incentivized students simply exert more effort on easier questions where it takes less additional effort to deduce the correct response. However, it is also possible that the observed improvement represents true gains in ability, as tutors may be able to provide knowledge about the easiest material more effectively.

The latter interpretation is cast into doubt, however, when we examine the impact of the treatment groups on ISAT scores. Column 5 of Table 2 reports these results where the dependent variable is the student's ISAT score in the subject area in which the student receives tutoring (reading or math), standardized by grade level. Table 3 examines the robustness of these results to the inclusion or exclusion of controls. The probe results are reported in columns 1 through 3

[15] Only 505 observations can be used in these regressions because the 8th grade math exam was deleted from the system before information about the difficulty of each question could be recorded.

and the ISAT scores are reported in columns 4 through 6. The qualitative pattern of results is the same regardless of the set of student characteristics and fixed effects that are included in the specification. Despite the fact that the incentivized Thinklink probes and the non-incentivized ISAT measure the same set of skills, parallel treatment effects on ISAT scores are not observed. It is therefore likely that the observed probe improvements are due not to improvements in these skills, but rather to increased effort and concentration on the test in response to the financial incentives.

Improvements are observed not only in response to incentives for the students, but also in response to potential rewards solely for parents and tutors. Accordingly, in order for this interpretation to be correct, students' optimal effort levels on an exam must increase in response to rewards for the other inputs, suggesting that parent and tutor welfare enter the students' utility functions. In this scenario, the lack of parallel ISAT improvement implies that students fall short of their effort frontier when not properly incentivized. This calls into question the ability of such tests to accurately measure student knowledge and the usefulness of these tests as an instrument of policy.

One potential problem with this conclusion is that different sets of students take each test. Only third through eighth graders take the ISAT and Thinklink exams, and scores are not available for many students even in these grades. Data are only available on the relevant subject exam for 230 of the 411 students who take the ISAT. Also, for various reasons such as transfers into or out of the school district or prolonged absences during the testing intervals, many students did not take all four tests - the baseline and assessment probes, and the baseline Thinklink and the ISAT. Among our sample of 547 students who took both probe exams, only 226 also took the first Thinklink and ISAT exams in the relevant subject area. It is possible that the absence of

treatment effects on the ISAT is due merely to selection and not because of real differences in how the treatments impact incentivized and non-incentivized tests. The strong treatment effects observed on the probe results might be driven by the students for whom ISAT scores are not available.

To guard against this possibility, Table 4 presents estimates with and without controls for both the probes (columns 1 through 3) and for the ISAT (columns 4 through 6) using only the subsample of students who took all four exams. The pattern of results using this restricted sample is very similar to what is found using the full sample. There are substantial treatment effects of the individual incentives on the incentivized exam, but no effects on the exam which is not incentivized. In fact, as reported in column 3 which includes all controls, the treatment effects among this subsample are stronger than observed before, and the *Student and Parent* treatment now has a similar impact on probe results as the individual input treatments. However, there is still no parallel effect on the ISAT exam.

Returning to Table 2, the treatments also do not have similarly strong effects on the other incentivized outcomes. Columns 6, 7 and 8 report the results of regressions where the dependent variable is class grade, number of unexcused absences and number of suspensions, respectively. While the effects of the individual party rewards (*Student*, *Tutor*, and *Parent*) are positive for grades, they are not statistically significant. However, this lack of improvement on grades is 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 absences and suspensions, although the point estimates are largely consistent with the hypothesis that the incentives should reduce both of these indicators of poor behavior.

Finally, while not quite statistically significant at traditional levels for any treatment other than the individual tutor incentive, the individual incentives result in increased probabilities of the student satisfying all of the achievement standards. Column 9 of Table 2 reports estimated marginal effects from a probit where the dependent variable indicates whether the student met the achievement threshold to qualify for a reward. Although not statistically significant, the point estimates suggest that the individual student and parent incentives, as well as the treatment where both student and parent are incentivized, each result a sizeable increase in the probability that all of the four standards required to receive a reward are met. The tutor incentive, meanwhile, results in a statistically significant 31 percentage point increase in the chance that the threshold is achieved. However, the point estimate suggests that students in the treatment where all three inputs are targeted actually are less likely to have met all standards than students in control.

Table 5 presents some sensitivity analyses on these results for the ThinkLink probe scores. Columns 1 and 2 report regressions where math and reading students are examined separately, and columns 3 and 4 report regressions where females and males are considered separately. The same qualitative pattern of results observed for the entire sample is present for each of these subsamples.

Table 6 presents the same sensitivity analyses for the ISAT results. The pattern of results is again qualitatively similar when we divide the sample by subject or gender – no significant treatment effects on ISAT scores are observed. We also restrict attention only to students who improved their probe score in column 5. Again, there is no significant impact of the treatments on ISAT result when limiting the sample in this way.

We can conclude that rewards for an individual input 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* and *Parent* 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 reduces 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 suggest that any improvements resulting from complementarities are overwhelmed by the impact of reduced effort from the individual inputs. Improvements are significantly smaller and indeed appear to be completely eliminated when multiple parties are incentivized. Hence, from a policy perspective, we can conclude that given a certain budget, it is far better to incentivize individual parties than to split the money between multiple parties.

Table 7 displays the results for the second assessment. The results are quite different from what we observe in the first assessment – indeed, we see hardly any treatment effects.[16] None of the treatment groups exhibit differences from control on either the incentivized exam or the probability of meeting the overall achievement threshold, as reported in columns 1 and 9, respectively. There are some significant coefficients for some treatments on more difficult questions and perhaps harmful effects of some treatments on grades and unexcused absences, but no systematic picture emerges.

We are cautious to interpret these results as several factors may have impacted student behavior towards the end of the school year when this assessment was conducted.[17] However, there was one crucial difference between the two assessments. For the first assessment, the test

[16] 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 2, 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.

[17] Aside from the sample attrition mentioned above, these factors include the following. First, 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. Second, students may not have taken the exam seriously because it was so close to the end of the school year. Indeed, 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* treatment and could have earned $90. Finally, 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.

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.

Our results support the findings of Levitt et al. (2011). In this experiment, they announce incentives the day of a standardized test, but provide payment immediately in some treatments while delaying payment by one month in other treatments. They find substantial treatment effects resulting from immediate rewards but no effects resulting from delayed rewards.[18] While we cannot rule out that the other factors mentioned above also contributed to our finding of insignificant effects in the second assessment, consistent with Levitt et al. (2011), we also find strong treatment effects when rewards are provided immediately but no treatment effects when rewards are delayed. As they note, this finding has important policy implications. When immediate rewards are not present, as is frequently the case in the educational setting, measures

[18] Fryer (2010) also finds no treatment effects from performance incentives on output measures in experiments conducted in New York City and Chicago where the provision of the rewards was delayed by as long as several weeks after the assessments took place.

of teacher effectiveness, school quality and the achievement gap which rely on standardized tests will be biased.

IV. Conclusion

This paper presents one of the few studies that examine how potential complementarities between inputs in the education production function can be exploited. The conclusion that emerges from our identification strategy is clear. Should strong complementarities exist, incentivizing only one input with a certain amount of money could potentially have a smaller impact on student achievement than spreading that money across multiple inputs. Instead, we find the opposite. While incentives for individual inputs have a large impact on student achievement as measured by standardized tests, an equivalent budget spent on two or more inputs has no such impact. This result has implications for both theory and policy. Nearly all recent studies that estimate an education production function have assumed it to be linear and additively separable in its inputs. This strict functional form assumption has never been justified, but the evidence presented here gives some hope that the assumption is innocuous. The results also suggest that policy makers with a limited budget can expect larger gains when targeting only one input with available funds, rather than spending portions of their budget on more than one input.

The results also should give policy makers strong pause when using standardized tests in which the students have no personal stake as a tool to evaluate the ability of schools or teachers to improve students' academic achievement. Students in our treatment groups show improvement when incentives are in place for themselves, their parents or their tutor on standardized tests. However, they show no such improvement on standardized tests that measure the same knowledge and skills when no incentives are in place. Students apparently improved

their scores because they exerted increased effort in response to the incentives, suggesting that they fall short of their effort frontier when incentives are not in place and they have no personal stake in the test results. Standardized tests in which the students have no stake therefore cannot be expected to accurately measure the true extent of their academic achievement.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25 (1) (January): 95-135.

Angrist, JD, and V Lavy. 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *Quarterly Journal of Economics* 114 (2): 533-575.

Corcoran, Sean P, Jennifer L Jennings, and Andrew A Beveridge. 2011. "Teacher effectiveness on high- and low-stakes tests." Unpublished mimeo, New York University.

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

Fryer, Roland G. Jr. 2010. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." NBER working paper 15898. Available at http://www.nber.org/papers/w15898.

Fryer, Roland G. Jr. 2012. "Aligning student, parent and teacher incentives: evidence from Houston public schools." NBER working paper 17752. Available at http://www.nber.org/papers/w17752.

Hanushek, EA. 2002. Publicly provided education. In *Handbook of Public Finance, vol. 4*, ed. Alan Auerbach and Martin Feldstein, 2045-2141. 4th ed. Amsterdam: North-Holland Press.

Hanushek, Eric a., John F. Kain, Jacob M. Markman, and Steven G. Rivkin. 2003. "Does peer ability affect student achievement?" *Journal of Applied Econometrics* 18 (5): 527-544.

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

Hoxby, C.M. 2000. "The Effects of Class Size on Student Achievement: New Evidence from Population Variation." *Quarterly Journal of economics* 115 (4): 1239–1285.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics* 26 (1): 101-136.

Krueger, A.B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114 (2): 497–532.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff. 2011. "The Impact of Short-term Incentives on Student Performance." Unpublished mimeo, University of Chicago.

Levitt, Steven D., John A. List, and Sally Sadoff. 2011. "The Effect of Performance-Based Incentives on Educational Achievement: Evidence from a Randomized Experiment." Unpublished mimeo, University of Chicago.

Rivkin, S.G., E.A. Hanushek, and J.F. Kain. 2005. "Teachers, schools, and academic achievement." *Econometrica* 73 (2): 417–458.

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

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *The Economic Journal* 113 (485): F3-F33.

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.
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.
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.
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

Table 1: Summary Statistics by Treatment Group: Baseline Assessment

	Control (1)	Parent (2)	Student (3)	Tutor (4)	Student and Parent (5)	All (6)
Standardized Baseline Probe	-0.017	0.026	0.154	-0.008	-0.102	-0.026
	(0.94)	(1.04)	(1.04)	(0.96)	(0.98)	(0.96)
Percent of Easy Questions Correct	44.715	47.421	47.043	47.245	41.798	45.323
	(21.36)	(23.82)	(23.42)	(24.50)	(20.81)	(24.34)
Percent of Moderate Questions Correct	35.576	39.981	42.310	41.377	38.772	38.762
	(20.16)	(22.94)	(22.32)	(21.07)	(21.25)	(19.97)
Percent of Difficult Questions Correct	38.454	41.519	36.110	37.524	41.748	35.073
	(24.33)	(24.27)	(22.37)	(24.02)	(23.79)	(20.57)
Standardized 2010 ISAT Score	209.640	215.724	209.417	211.010	205.613	210.091
	(20.24)	(22.57)	(24.85)	(26.09)	(19.98)	(20.97)
Standardized Thinklink 3 Score	1491.757	1504.267	1478.961	1477.955	1483.403	1476.795
	(79.77)	(69.39)	(83.00)	(102.07)	(69.50)	(86.66)
Standardized Baseline Grades	-0.005	0.485	-0.266	0.266	0.310	-0.124
	(1.06)	(1.06)	(1.12)	(0.78)	(0.89)	(1.10)
Gender, 1 = Female	0.549	0.527	0.415*	0.557	0.451	0.489
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Reduced or Free Lunch, 1 = Yes	0.896	0.848	0.813	0.884	0.875	0.936
	(0.31)	(0.36)	(0.39)	(0.32)	(0.33)	(0.25)
African American, 1 = Yes	0.313	0.212	0.297	0.316	0.375	0.234
	(0.47)	(0.41)	(0.46)	(0.47)	(0.49)	(0.43)
Hispanic, 1 = Yes	0.458	0.404	0.374	0.305**	0.284**	0.543
	(0.50)	(0.49)	(0.49)	(0.46)	(0.45)	(0.50)
Number of Meetings with Tutor per Week	3.263	3.505	3.281	3.537	3.379	3.484
	(1.21)	(1.17)	(1.23)	(1.17)	(1.47)	(1.23)
Parents Received Mail, 1 = Yes	0.905	0.889	0.944	0.937	0.943	0.892
	(0.29)	(0.32)	(0.23)	(0.25)	(0.23)	(0.31)
First Assessment Attrition	1	1	3	5	0	2
First Assessment Attrition (Percent)	1.042	1.031	3.297	5.263	0.000	2.128
Second Assessment Attrition	14	12	12	12	13	10
Second Assessment Attrition (Percent)	14.737	12.245	13.636	13.333	14.773	10.870

Note: The table reports means and robust standard errors clustered by tutor group. The asterisks indicate statistical significance from the control group at 10/5/1 percent level. Every treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *Parent* treatment, students in the *Student* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* 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 Probe and Grade are both standardized using our sample and the 2010 ISAT is standardized using the population of students who took the test.

Table 2: First Assessment

	Probe	Easy	Moderate	Difficult	ISAT Score	Grade	Unexcused	Suspension	Threshold
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Parent	0.458**	6.987**	3.562	0.458	1.689	0.184	-0.203	-0.035	0.175
	(0.190)	(2.855)	(3.666)	(3.734)	(3.813)	(0.171)	(0.391)	(0.086)	(0.141)
Student	0.315**	6.207**	0.644	6.527*	0.715	0.031	-0.293	-0.092	0.166
	(0.140)	(2.551)	(3.258)	(3.596)	(2.938)	(0.147)	(0.248)	(0.067)	(0.144)
Tutor	0.319**	6.693**	-1.740	1.894	-1.418	0.283	-0.198	-0.053	0.310**
	(0.152)	(2.667)	(3.539)	(3.852)	(3.048)	(0.179)	(0.250)	(0.058)	(0.136)
Student and Parent	0.265	1.894	3.091	-2.626	3.595	-0.110	0.186	0.026	0.134
	(0.174)	(2.928)	(3.560)	(3.877)	(3.412)	(0.192)	(0.335)	(0.114)	(0.136)
All	0.093	-0.256	-3.990	-1.621	-4.841	-0.283	-0.166	-0.036	-0.175
	(0.211)	(3.360)	(3.651)	(3.993)	(4.271)	(0.175)	(0.342)	(0.073)	(0.106)
Constant	-0.809*	55.100***	38.231***	37.963***	102.138**	-1.575***	2.823***	0.027	
	(0.419)	(7.507)	(9.273)	(12.810)	(38.840)	(0.479)	(1.035)	(0.120)	
Basline	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tutor FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade Level FE	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Teacher FE	No	No	No	No	No	Yes	No	No	Yes
N	547	505	505	505	230	561	561	551	551
Adj. R-sq	0.154	0.152	0.177	0.065	0.713	0.343	0.109	0.378	0.178

Note: The table reports coefficient estimates and robust standard errors clustered by tutor group. The asterisks indicate statistical significance at 10/5/1 percent level. Every treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *Parent* treatment, students in the *Student* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. We standardized ISAT scores using the population of students who took the 2011 ISAT. Probes and grades are standardized using our 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 dependant variable, respectively. Unexcused and Suspension columns use the number of unexcused absences and the number of all-day suspensions as the outcome. Threshold is a probit where the outcome is 1 if students met the threshold. The Coefficient estimates are the marginal effects and the Adj. R-sq reports the psuedo R-sq for this regression. Student characteristics include race, gender, reduced-lunch status, the subject in which the student was tutored, whether the student was tutored in both subjects, parent's native language, whether the parent received mail, home many extra homework assignments were turned into tutors, 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, while Grade uses the students baseline grades. ISAT Score uses Thinklink 3 as its baseline.

Table 3: Control Variables

	Probe (1)	Probe (2)	Probe (3)	2011 ISAT (4)	2011 ISAT (5)	2011 ISAT (6)
Parent	0.398**	0.404***	0.458**	-3.605	2.403	1.689
	(0.161)	(0.146)	(0.190)	(5.532)	(3.567)	(3.813)
Student	0.269*	0.300**	0.315**	-3.398	-0.442	0.715
	(0.148)	(0.137)	(0.140)	(4.250)	(3.310)	(2.938)
Tutor	0.207	0.286*	0.319**	-10.650**	-2.132	-1.418
	(0.163)	(0.149)	(0.152)	(4.818)	(3.604)	(3.048)
Student and Parent	0.211	0.226	0.265	-3.307	4.813	3.595
	(0.152)	(0.141)	(0.174)	(5.083)	(4.067)	(3.412)
All	0.043	0.04	0.093	-6.124	-4.165	-4.841
	(0.170)	(0.167)	(0.211)	(4.524)	(4.388)	(4.271)
Constant	-0.185*	-0.950***	-0.809*	-79.590***	93.283**	102.138**
	(0.107)	(0.325)	(0.419)	(25.350)	(37.720)	(38.840)
Baseline	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	No	No	Yes	No	No	Yes
Tutor FE	No	Yes	Yes	No	Yes	Yes
Grade Level FE	No	Yes	Yes	No	No	Yes
N	547	547	547	230	230	230
Adj. R-sq	0.102	0.154	0.154	0.562	0.706	0.713

*Note:*The table reports coefficient estimates and robust standard errors clustered by tutor group. The asterisks indicate statistical significance at 10/5/1 percent level. Every treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *Parent* treatment, students in the *Student* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. Probe Scores were standardized using our sample and ISAT scores using the population of students who took that test. Probes use the first probe as the baseline and ISAT uses the third Thinklink as the baseline. Columns (1) and (4) control only for treatment and outcome baseline. Columns (2) and (5) control for tutor and grade level fixed effects in addition to the outcome baseline. Columns (3) and (6) control for the outcome baseline, tutor fixed effects, grade level fixed effects, and student characteristics. These characteristics include race, gender, reduced-lunch status, the subject in which the student was tutored, whether the student was tutored in both subjects, parent's native language, whether the parent received mail, home many extra homework assignments were turned into tutors, and the number of meetings with the tutor per week.

Table 4: Selection

	Probe	Probe	Probe	2011 ISAT	2011 ISAT	2011 ISAT
	(1)	(2)	(3)	(4)	(5)	(6)
Parent	0.686***	0.808***	0.677***	-3.968	1.866	4.244
	(0.204)	(0.223)	(0.250)	(5.559)	(3.587)	(3.490)
Student	0.278	0.502**	0.508**	-3.763	-0.962	0.587
	(0.225)	(0.210)	(0.193)	(4.283)	(3.273)	(2.621)
Tutor	0.422*	0.600**	0.591**	-11.01**	-2.642	-0.306
	(0.216)	(0.228)	(0.229)	(4.848)	(3.564)	(22.568)
Student and Parent	0.619***	0.787***	0.630**	-3.673	4.416	2.890
	(0.207)	(0.242)	(0.254)	(5.110)	(4.045)	(3.478)
All	0.054	0.160	0.050	-6.491	-4.738	-2.112
	(0.244)	(0.264)	(0.286)	(4.554)	(4.354)	(3.747)
Constant	-0.287**	-1.472***	-1.864**	-74.594***	90.289**	103.134**
	(0.142)	(0.547)	(0.710)	(25.350)	(37.727)	(32.481)
Baseline	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	No	No	Yes	No	No	Yes
Tutor FE	No	Yes	Yes	No	Yes	Yes
Grade Level FE	No	Yes	Yes	No	No	Yes
N	226	226	226	226	226	226
Adj. R-sq	0.145	0.18	0.171	0.563	0.709	0.737

Note: The table reports coefficient estimates and robust standard errors clustered by tutor group. Only subjects who took the 2011 ISAT, the second probe, and both baselines are included in the regression. The asterisks indicate statistical significance at 10/5/1 percent level. Every treatment had bi-monthly monetary incentives for student performance. Parents received a $ 90 incentive in the *Parent* treatment, students in the *Student* treatment, and tutors in the *Tutor* treatment. Both students and parents received a $ 45 incentive in the *Student and Parent* treatment while everyone received a $ 30 incentive in the *All* treatment. Probe Scores were standardized using our sample. Probes use the first probe as the baseline. Only students who took the 2011 ISAT and Thinklink 3 were included in these regressions. Columns (1) and (4) control only for treatment and outcome baseline. Columns (2) and (5) control for tutor and grade level fixed effects in addition to the outcome baseline. Columns (3) and (6) control for the outcome baseline, tutor fixed effects, grade level fixed effects, and student characteristics. These characteristics include race, gender, reduced-lunch status, the subject in which the student was tutored, whether the student was tutored in both subjects, parent's native language, whether the parent received mail, home many extra homework assignments were turned into tutors, and the number of meetings with the tutor per week.

Table 5: First Assessment Probe Sensitivity

	Math	Reading	Female	Male
	(1)	(2)	(3)	(4)
Parent	0.671**	0.299	0.430*	0.563**
	(0.324)	(0.235)	(0.246)	(0.250)
Student	0.391	0.361**	0.252	0.310
	(0.275)	(0.175)	(0.212)	(0.196)
Tutor	0.409	0.274	0.336*	0.454**
	(0.279)	(0.186)	(0.185)	(0.222)
Student and Parent	0.587	0.100	0.497*	0.179
	(0.358)	(0.188)	(0.259)	(0.211)
All	-0.020	0.203	0.186	0.068
	(0.305)	(0.291)	(0.257)	(0.280)
Constant	-1.308**	-1.261*	-1.116	-0.805
	(0.545)	(0.648)	(0.716)	(0.620)
Basline	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes
Tutor FE	Yes	Yes	Yes	Yes
Grade Level FE	Yes	Yes	Yes	Yes
N	206	341	280	267
Adj. R-sq	0.148	0.147	0.137	0.166

Note: The table reports coefficient estimates and robust standard errors clustered by tutor group. The asterisks indicate statistical significance at 10/5/1 percent level. Every treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *Parent* treatment, students in the *Student* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. Probe scores were standardized using our sample. Columns (1) and (2) divide the sample by subject while columns (3) and (4) divide the sample by gender. All outcomes use the first probe as their baseline. Student characteristics include race, gender, reduced-lunch status, the subject in which the student was tutored, whether the student was tutored in both subjects, parent's native language, whether the parent received mail, home many extra homework assignments were turned into tutors, and the number of meetings with the tutor per week.

Table 6: 2011 ISAT Sensitivity

	Math	Reading	Female	Male	Improved
	(1)	(2)	(3)	(4)	(5)
Parent	2.974	-4.133	10.090	-4.994	-0.143
	(5.233)	(6.613)	(6.471)	(6.027)	(4.453)
Student	2.207	1.744	-1.107	1.237	-1.946
	(4.365)	(3.483)	(4.083)	(5.259)	(3.822)
Tutor	-1.081	-2.090	-4.825	4.669	-2.699
	(3.022)	(4.452)	(3.837)	(6.132)	(3.869)
Student and Parent	-4.475	7.814	2.469	7.732	2.137
	(4.669)	(4.857)	(4.642)	(5.899)	(4.242)
All	-7.092	-4.276	0.242	-7.631	-3.963
	(5.488)	(5.926)	(4.628)	(7.459)	(4.821)
Constant	-34.990	89.530	42.420	146.7***	171.7***
	(39.580)	(58.550)	(47.940)	(52.850)	(33.990)
Basline	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes
Tutor FE	Yes	Yes	Yes	Yes	Yes
Grade Level FE	Yes	Yes	Yes	Yes	Yes
N	91	139	122	108	116
Adj. R-sq	0.838	0.638	0.75	0.712	0.782

Note: The table reports coefficient estimates and robust standard errors clustered by tutor group. The asterisks indicate statistical significance at 10/5/1 percent level. Every treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *Parent* treatment, students in the *Student* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. ISAT scores were standardized using the population of students who took the ISAT in 2011. Columns (1) and (2) divide the sample by subject while the columns (3) and (4) divide the sample by gender. Column (5) restricts our sample to students who improved their probe scores from the Baseline to first assessment. All outcomes use the third Thinklink as their baseline. Student characteristics include race, gender, reduced-lunch status, the subject in which the student was tutored, whether the student was tutored in both subjects, parent's native language, whether the parent received mail, home many extra homework assignments were turned into tutors, and the number of meetings with the tutor per week.

Table 7: Second Assessment

	Probe	Easy	Moderate	Difficult	Thinklink	Grade	Unexcused	Suspension	Threshold
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Parent	-0.061	-3.421	3.859	-2.434	-6.867**	0.012	0.296	0.009	-0.146
	(0.219)	(7.326)	(5.072)	(7.007)	(3.354)	(0.234)	(0.324)	(0.070)	(0.065)
Student	0.053	-7.804	6.865	2.338	3.126	0.032	0.608**	-0.053	-0.882
	(0.151)	(5.183)	(4.490)	(6.943)	(2.526)	(0.182)	(0.306)	(0.046)	(0.080)
Tutor	0.212	-0.238	7.823*	12.310**	1.520	-0.342*	0.743**	0.001	-0.051
	(0.157)	(5.390)	(4.355)	(5.916)	(2.928)	(0.203)	(0.376)	(0.063)	(0.090)
Student and Parent	-0.029	-8.080	9.580**	2.896	-2.023	-0.188	0.124	0.029	-0.072
	(0.198)	(5.499)	(4.658)	(7.290)	(3.357)	(0.260)	(0.325)	(0.060)	(0.077)
All	0.044	-3.786	10.790**	-0.717	-2.237	-0.089	0.252	-0.017	-0.078
	(0.194)	(6.621)	(4.842)	(6.124)	(3.759)	(0.222)	(0.301)	(0.051)	(0.085)
Constant	-1.735***	37.071***	27.110	51.663***	18.462***	0.670	1.754*	0.003	
	(0.502)	(13.056)	(16.933)	(14.624)	(6.610)	(0.502)	(1.024)	(0.079)	
Basline	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Tutor FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Grade Level FE	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Teacher FE	No	No	No	No	No	Yes	No	No	Yes
N	474	424	424	424	547	556	556	514	393
Adj. R-sq	0.289	0.134	0.158	0.221	0.378	0.356	-0.003	0.077	0.290

Note: The table reports coefficient estimates and robust standard errors clustered by tutor group. The asterisks indicate statistical significance at 10/5/1 percent level. Every treatment had bi-monthly monetary incentives for student performance. Parents received incentives in the *Parent* treatment, students in the *Student* treatment, and tutors in the *Tutor* treatment. Both students and parents received incentives in the *Student and Parent* treatment while everyone received incentives in the *All* treatment. Probes and grades are standardized using our 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 dependant variable, respectively. Unexcused and Suspension columns use the number of unexcused absences and the number of all-day suspensions as the outcome. Threshold is a probit where the outcome is 1 if students met the threshold. The Coefficient estimates are the marginal effects and the Adj. R-sq reports the psuedo R-sq for this regression. Student characteristics include race, gender, reduced-lunch status, the subject in which the student was tutored, whether the student was tutored in both subjects, parent's native language, whether the parent received mail, home many extra homework assignments were turned into tutors, and the number of meetings with the tutor per week. Probe, Easy, Moderate, and Difficult use the respective score on the first assessment as their baseline, while Grade uses the students first assessment grades. Thinklink uses the previous Thinklink as its baseline.