Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment*

Roland G. Fryer, Jr.
Harvard University

Steven D. Levitt
The University of Chicago

John List
The University of Chicago

Sally Sadoff
UC San Diego

April, 2018

Abstract

In a field experiment, we provide financial incentives to teachers framed either as *gains,* received at the end of the year, or as *losses,* in which teachers receive upfront bonuses that must be paid back if their students do not improve sufficiently. Pooling two waves of the experiment, loss-framed incentives improve math achievement by an estimated 0.124 standard deviations (σ) with large effects in the first wave and no effects in the second wave. Effects for gain framed incentives are smaller and not statistically significant, approximately 0.051σ. We find suggestive evidence that effects on teacher value added persist post-treatment.

* Fryer: Harvard University, 1805 Cambridge Street, Cambridge, MA, 02138; rolandfryer@edlabs.harvard.edu. Levitt: University of Chicago, 5807 S Woodlawn Avenue, Chicago, IL 60637; slevitt@uchicago.edu. List: 5757 S. University Avenue, Chicago, IL 60637; jlist@uchicago.edu. Sadoff: UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093; ssadoff@ucsd.edu. We are grateful to Tom Amadio and the Chicago Heights teachers' union for their support in conducting our experiment. Eszter Czibor, Jonathan Davis, Matt Davis, Sean Golden, William Murdock III, Phuong Ta and Wooju Lee provided exceptional research assistance. Financial support from the Kenneth and Anne Griffin Foundation is gratefully acknowledged. The research was conducted with approval from the University of Chicago Institutional Review Board. Please direct correspondence to Sally Sadoff: ssadoff@ucsd.edu.

I. Introduction

Good teachers matter. A one-standard deviation improvement in teacher quality translates into annual student achievement gains of 0.15 to 0.24 standard deviations (hereafter σ) in math and 0.15σ to 0.20σ in reading (Rockoff, 2004; Rivkin et al, 2005; Aaronson et al., 2007; Kane and Staiger, 2008). These effects are comparable to reducing class size by about one-third (Krueger, 1999). Similarly, Chetty et al. (2014) estimate that a one-standard deviation increase in teacher quality in a single grade increases earnings by about 1% per year; students assigned to these teachers are also more likely to attend college and save for retirement, and less likely to have children when teenagers.

Despite great interest, it has proven difficult to identify public policies that materially improve teacher quality. One strategy is to hire better teachers, but attempts to identify *ex ante* the most productive teachers have been mostly unsuccessful (Aaronson et al., 2007; Rivkin et al., 2005; Kane and Staiger, 2008; Rockoff et al., 2011).[1] A second approach is to provide training to existing teachers to make them more effective. Such programs, unfortunately, have had little impact on teacher quality (see e.g., Boyd et al., 2007 for a review).[2]

A third public policy approach has been to tie teacher incentives to the achievement of their students. Since 2006, the U.S. Department of Education has provided over $1 billion to incentive programs through the Teacher Incentive Fund (now the Teacher and School Leader Incentive Program); a program designed specifically to support efforts for developing and implementing performance-based compensation systems in schools.[3] At least seven states and many more school districts have implemented teacher incentive programs in an effort to increase student achievement (Fryer, 2013; Fryer, 2017). The empirical evidence on the effectiveness of teacher incentive programs is mixed (Neal, 2011 and Fryer, 2017 provide reviews). In developing countries where the degree of teacher

[1] More recently, higher cost more intensive screening mechanisms show promise (Jacob et al., 2016; Goldhaber et al., 2017).

[2] An alternative approach to traditional professional development, teacher coaching, has demonstrated positive impacts in smaller scale studies but has been less successful in effectiveness trials (Kraft et al., in press).

[3] When states apply for the funding through the $4.4 billion Race to the Top initiative, one of the criteria they are evaluated on is their program's use of student achievement in decisions of raises, tenure, and promotions. As discussed below, Chiang et al. (2017) evaluate the effectiveness of teacher performance pay programs implemented through the Teacher Incentive Fund.

professionalism is extremely low and absenteeism is rampant, field experiments that link pay to teacher performance have been associated with substantial improvements in student test scores (Duflo et al., 2012; Glewwe et al. 2010; Muralidharan and Sundararaman, 2011; Loyalka et al., 2016), though implementation by policymakers rather than researchers has been less successful (Barrera-Osorio and Raju, 2017). Conversely, the few other field experiments conducted in the United States have shown small, if not negative, treatment effects (Glazerman et al., 2009; Springer et al., 2011; Springer et al., 2012; Fryer, 2013; Chiang et al., 2017).[4]

This paper reports the results of a field experiment examining the impact of teacher incentives on math performance. The experiment was conducted during the 2010-2011 and the 2011-2012 school years in nine schools in Chicago Heights, IL. In the design of the incentives, we exploit loss aversion by framing the teacher rewards as losses rather than gains in some of our treatments.[5] One set of teachers – whom we label the "Gain" treatment – received "traditional" financial incentives in the form of bonuses at the end of the year linked to student achievement. Other teachers – the "Loss" treatment – were given a lump sum payment at the beginning of the school year and informed that they would have to return some or all of it if their students did not meet performance targets. Teachers in the "Gain" and "Loss" groups with the same performance received the same final bonus. Within the "Loss" and "Gain" groups we additionally test whether there are heterogeneous effects for individual rewards compared to team rewards.

In all groups, we incentivized performance according to the "pay for percentile" method developed by Barlevy and Neal (2012). Teachers are rewarded according to how

[4] Non-experimental analyses of teacher incentive programs in the United States have also shown little measurable success (Vigdor, 2008) with larger impacts among subgroups of teachers who should arguably be most responsive to the incentives as they are designed (Dee and Wycoff, 2015; Imberman and Lovenheim, 2015), though one should interpret these data with caution due to the lack of credible causal estimates. In an important observation, Neal (2011) discusses how the incentive pay schemes tested thus far in the US are either team incentives (e.g. Fryer, 2013) or are sufficiently obtuse (e.g. Springer et al., 2011). This leads to problems when trying to calculate the incentive effect at the individual teacher level and could be the reason past experiments observed little to no incentive effects. In subsequent work using an incentive design similar to ours, Brownback and Sadoff (2018) find large impacts of incentives for instructors at community colleges.

[5] There is mixed evidence from online, laboratory and field studies on the impact of framing on effort and productivity. Some studies find evidence suggesting that behavior is more responsive to incentives framed as losses (Brooks et al., 2012; Hossain and List, 2012; Hong et al., 2015; Armantier and Boly, 2015; Imas et al., 2016; Levitt et al., 2016), while others find little impact of framing (List and Samek, 2015; Della Vigna and Pope, 2017; de Quidt et al., 2017; Englmaier et al., 2018).

highly their students' test score improvement ranks among peers from other schools with similar baseline achievement and demographic characteristics.[6]

A number of results emerge from our study. First, our intervention was more successful than previous field experiments in the United States using teacher incentives. The estimated pooled treatment effect across all incentives and years of the program is a 0.099σ (standard error = 0.051) improvement in math test scores.[7]

Second, the effects are concentrated on loss-framed incentives and the first year of the experiment. In the first year the incentives are offered, loss-framed incentives improve math performance by an estimated 0.234σ (0.080). Teacher incentives that are framed as gains demonstrate smaller effects that are economically meaningful but not statistically significant, improving math performance by an estimated 0.1σ (0.079). The effects of the loss- and gain-framed incentives are significantly different at the $p = 0.051$ level. There is no impact of incentives in the second wave of the experiment. As we discuss in more detail in Section 6, this may be due in part to the constraints of our experimental design in which both teachers and students moved between incentive treatments across years. However, we cannot rule out that the effects of our incentives may not replicate. The pooled treatment effect for loss-framed incentives across both waves of the experiment is 0.124σ (0.056). For gain-framed incentives, the pooled treatment effects are 0.051σ (0.062). The results are similar whether incentives are provided to individual teachers or teams of two teachers.

Third, we find suggestive evidence that the impact of loss-framed incentives on teacher value added persists after treatment. For teachers who received loss-framed

[6] As Neal (2011) describes, pay for percentile schemes separate incentives and performance measurements for teachers since this method only uses information on relative ranks of the students. Thus, motivation for teachers to engage in behaviors (e.g. coaching or cheating) that would contaminate performance measures of the students is minimized. Pay for percentile may also help uphold a collaborative atmosphere among teachers within the same school by only comparing a teacher's students to students from a different school.

[7] Our agreement with the Chicago Heights teachers' union required us to offer every teacher the opportunity to participate in the incentive program, including Social Studies teachers, Language Arts teachers, and interventionists. Since the district only administers Math, Reading, and Science tests (the last only in 4th and 7th grades), we allowed Social Studies teachers, Language Arts teachers, and reading interventionists to earn rewards based on their students' performance on the Reading examination. In other words, a student's reading performance often determined the rewards of multiple teachers, who were potentially assigned to different treatment groups. While this structure created some confusion among teachers and likely contaminated our Reading results, it allowed us to preserve a rigorous experimental design for our math treatments. In the interest of full disclosure, we present results for reading tests in the Appendix, but our discussion will focus primarily on the impacts of the various treatments on math performance. We discuss these issues in more detail in Section 3.

incentives in the first year, the effects on teacher value added are 0.167σ (0.112) pooling five years of follow up (and 0.177σ (0.065) including the treatment year). There is no impact of gain-framed incentives on post-treatment value added, -0.007σ (0.116). We also find suggestive evidence that the impact of incentives, whether framed as losses or gains, is largest among younger students in Kindergarten through second grade. For these grades, the estimated effects of incentives are economically meaningful (0.15-0.49 σ) in both years of the experiment with effects of approximately 0.25σ (0.12) pooling across years**.**

Together, our findings suggest that, in contrast to previous experimental results, incentives can improve the performance of U.S. teachers and that the addition of framing can improve their effectiveness. The results of our experiment are consistent with over three decades of psychological and economic research on the power of framing to motivate individual behavior (Kahneman and Tversky, 1979), though other models may also be consistent with the data.

The remainder of the paper is organized as follows. Section 2 provides a brief literature review. Section 3 details the experiment and its implementation, including the randomization. Section 4 describes the data and analysis. Section 5 presents estimates of the impact of teacher incentives on student achievement. Section 6 discusses alternative interpretations of our results. The final section concludes. There are two online appendices. Online Appendix A provides details on how we construct our covariates and our sample from the school district administrative files and survey data used in our analysis. Online Appendix B is a detailed implementation guide that describes how the experiment was implemented and milestones reached.

II. A Brief Review of the Literature

The theory underlying teacher incentives programs is straightforward: if teachers lack motivation to put effort into important inputs of the education production function (e.g. lesson planning, parental engagement), financial incentives tied to student achievement may have a positive impact by motivating teachers to increase their effort.

There are a number of reasons, however, why teacher incentives may fail to operate in the desired manner. For instance, teachers may not know how to increase student achievement, the production function may have important complementarities outside their

control, or the incentives may be either too confusing or too weak to induce extra effort. Moreover, if teacher incentives have unintended consequences such as explicit cheating, teaching to the test, or focusing on specific, tested objectives at the expense of more general learning, teacher incentives could have a negative impact on student performance (Holmstrom and Milgrom, 1991; Jacob and Levitt, 2003). Others argue that teacher incentives can decrease a teacher's intrinsic motivation or lead to harmful competition between teachers in what some believe to be a collaborative environment (Johnson, 1984; Firestone and Pennell, 1993).

Despite the controversy, there is a growing literature on the role of teacher incentives on student performance (Glazerman et al., 2009; Glewwe et al., 2010; Lavy, 2002; Lavy, 2009; Muralidharan and Sundararaman, 2011; Fryer 2013, Springer et al., 2011; Vigdor, 2008), including an emerging literature on the optimal design of such incentives (Neal, 2011). There are nine prior and concurrent studies, four of them outside the US, which provide experimental estimates of the causal impact of teacher performance pay incentives on student achievement: Glewwe et al. (2010), Muralidharan and Sundararaman (2011), Duflo et al. (2012), Barrera-Osorio and Raju (2017), Glazerman et al. (2009), Springer et al. (2011), Springer et al. (2012), Fryer (2013) and Chiang et al. (2017).[8] Subsequent to our work, two additional experimental studies provide evidence related to our design: in a developing country context, Loyalka et al. (2016); and in a post-secondary context, Brownback and Sadoff (2018). Figure 1 displays the treatment effects from these eleven experiments. For comparability across studies, we display the results for mathematics performance (when available) in the first year of the experiment.[9] Overall treatment effects pooling across subjects and years are reported below.

Evidence from Developing Countries

Duflo et al. (2012) randomly sampled 60 schools in rural India and provided them with financial incentives to reduce absenteeism. The incentive scheme was simple:

[8] In related work, Glazerman et al. (2013) examine the impact of incentives for high performing teachers to transfer to low performing schools.

[9] Brownback and Sadoff (2018) estimate effects pooling across final exams in multiple post-secondary departments. Glewwe et al. (2010) and Barrera-Osorio and Basu (2017) estimate effects pooling subjects on a government exam.

teachers' pay was linear in their attendance, at the rate of Rs 50 per day, after the first 10 days of each month. They found that teacher absence rates were significantly lower in treatment schools (22 percent) compared to control schools (42 percent) and that student achievement in treatment schools was 0.17σ (0.09) higher than in control schools.

Glewwe et al. (2010) report results from a randomized evaluation that provided 4th through 8th grade teachers in Kenya with group incentives based on test scores. They find that while test scores increased in program schools in the short run, students did not retain the gains after the incentive program ended. Glewwe et al. (2010) interpret these results as being consistent with teachers expending effort towards short-term increases in test scores but not towards long-term learning.

Muralidharan and Sundararaman (2011) investigate the effect of individual and group incentives in 300 schools in Andhra Pradesh, India and find that group and individual incentives increased student achievement by 0.165σ (.042) after one year. While the effects of the group incentive and the individual incentive treatments are very similar in year one, they diverge in the second year. Two-year effects are 0.283σ (.058) and 0.154σ (0.057) for the individual and group treatments, respectively.

Barrera-Osorio and Raju (2017) present results from a government-administered teacher performance pay program in Punjab, Pakistan that offered both individual and group incentives based on school-wide improvements in enrollment, exam participation and exam scores. They find an increase in exam participation rates in the third year of the program but no impact on student performance. They argue that the limited impact may be due to administrative constraints on the incentive structure and available data.

Loyalka et al. (2016) test alternative incentive performance pay structures in primary schools in Western China. They find an overall impact of incentives on math performance of 0.074σ (0.044) with the largest effects among teachers rewarded using the same "pay-for-percentile" scheme as in our study, 0.148σ (0.064).

Evidence from Experiments in America

Glazerman et al. (2009) evaluate the first year of a randomized rollout of the Teacher Advancement Program (TAP) in Chicago Public Schools. Of the sixteen K-8 schools that volunteered to participate, eight were randomly assigned to receive the

program in the first year (the other eight schools began the program the following year). Teachers in the program received an expected annual bonus of $2,000 based on their value-added and classroom observations. Teachers could also earn extra pay by being promoted to being a Mentor ($7,000) or Lead Teacher ($15,000). As Mentors, teachers were expected to provide ongoing classroom support to other teachers in the school. Lead Teachers served on the leadership team responsible for implementing TAP, analyzing student data, and developing achievement plans. In addition, Mentors and Lead Teachers conducted weekly group meetings to foster collaboration. Glazerman et al. (2009) find that the first year of the program increases teacher retention but has no impact on teacher satisfaction or student achievement.

Springer et al. (2011) evaluate a three-year pilot initiative on teacher incentives conducted in the Metropolitan Nashville School System between the 2006-2007 school year and the 2008-2009 school year. Approximately 300 middle school mathematics teachers who volunteered to participate in the program were randomly assigned to the treatment or the control group, and those assigned to the treatment group could earn up to $15,000 as a bonus if their students made gains in state mathematics test scores equivalent to the 95th percentile in the district. They were awarded $5,000 and $10,000 if their students made gains equivalent to the 80th and the 90th percentiles, respectively. Springer et al. (2010) found there was no significant treatment effect either on student achievement or on measures of teachers' behavior such as teaching practices.

Fryer (2013) conducted an experiment on teacher incentives in over 200 New York City public schools. Each participating school could earn $3,000 for every union-represented staff member, which the school could distribute at its own discretion, if the school met the annual performance target set by the Department of Education based on school report card scores. Each participating school was given $1,500 per union staff member if it met at least 75% of the target, but not the full target. Each school had the power to decide whether all of the rewards would be given to a small subset of teachers with the highest value-added, whether the winners of the rewards would be decided by lottery, or virtually anything in-between. The only restriction was that schools were not allowed to distribute rewards based on job seniority. Yet, despite this apparent flexibility, the vast majority of schools chose to distribute the rewards evenly, and there was no effect

on student achievement or teacher behavior. If anything, there was a negative impact, especially in larger schools where free-riding may have been problematic.[10]

Springer et al. (2012) evaluate another group incentive experiment that took place in the Round Rock Independent School District in Texas. The program awarded teams of middle school teachers bonuses based on their collective contribution to students' test score gains. Two years after the initial randomization, Springer et al. (2012) found no significant impact on the attitudes and practices of teachers or on the academic achievement of students.

Chiang et el. (2017) report the results from four years of program implementation of the Teacher Incentives Fund (TIF) in ten school districts with the structure of the performance-based bonuses varying across districts. In all four years of the program, the estimated impact of incentives on student test scores is small ranging from 0.02 to 0.06σ in math and 0.03-0.04σ in reading.[11]

Our specific contribution is straightforward: this is the first experimental study to test whether teacher incentives framed as a "Loss" are more effective than traditional incentives that are framed as "Gains." Subsequent to our study, Brownback and Sadoff (2018) test the effect of loss-framed bonuses among community college instructors in Indiana. Similar to our design, instructors received upfront bonuses at the start of the semester that had to be paid back if students did not meet performance targets. Brownback and Sadoff (2018) estimate that incentives improved student exam performance by 0.2σ (0.056) compared to a no incentive control group (they do not test gain-framed incentives).[12] Finally, we contribute to a small but growing literature that uses randomized

[10] Goodman and Turner (2013) -- using the same data as Fryer (2013) – and Imberman and Lovenheim (2015) using non-experimental data from the ASPIRE program in Houston find evidence that when teachers are offered group incentives, effects on student performance are larger when there are lower incentives to free-ride (e.g., teachers are responsible for a greater share of the students that determine their reward).

[11] There is mixed evidence from non-experimental evaluations of teacher incentive programs in the U.S. Vigdor (2008) reports a non-significant effect of the ABC School-wide Bonus Program in North Carolina. Sojourner et al. (2014) find evidence of small effects (0.03σ) in reading and no effect in math of the Quality Compensation program in Minnesota. Using a regression discontinuity design, Dee and Wycoff (2015) find evidence of improved performance among teachers at both the lower threshold for dismissal and the upper threshold for performance bonuses in Washington D.C.'s IMPACT program. Outside the US, Lavy (2002, 2009) reports significant results for teacher incentive programs in Israel.

[12] Brownback and Sadoff (2018) also test whether instructor incentives are more effective in combination with student incentives and find no evidence of complementarities.

field experiments to test incentive pay in organizations (Shearer, 2004; Bandiera et al., 2007, 2013; Hossain and List, 2009).

III. Program Details and Randomization

Incentive Design and Implementation

The city of Chicago Heights is located thirty miles south of Chicago, IL. The district contains nine Kindergarten through eighth grade schools with a total of approximately 3,200 students. Like larger urban school districts, Chicago Heights is made up primarily of low-income minority students with achievement rates well below the state average. In the pre-treatment year, 64% of students met the minimum standard on the Illinois State Achievement Test (ISAT) compared to 81% of students statewide. Roughly 98% of the elementary and middle school students in our sample are eligible for free or reduced-price lunch.

As part of our agreement with the teachers' union to conduct an experiment with teacher incentives, (1) program participation had to be made available to every K-8 classroom teacher in subjects tested on the statewide exam, as well as reading and math interventionists,[13] and (2) teachers who participated in the experiment both years were required to be placed in a treatment group at least once (more on this, and the challenges for inference, below). For ease of exposition, we will describe the details of the year one experiment below and note any important departures that took place in year two. Online Appendix B provides a detailed implementation guide for both years.

Table 1 provides a brief summary of the treatments. Participating teachers were randomly assigned to the control group or to one of four treatment arms: "Individual Loss", "Individual Gain", "Team Loss", or "Team Gain". In the second year, the "Team Gain" treatment group was dropped to increase power in the other treatment arms. In the "Individual" treatments, teachers received rewards based on their students' end of the year performance on the ThinkLink Predictive Assessment (ThinkLink). ThinkLink is an otherwise low stakes standardized diagnostic assessment that is designed to be aligned with

[13] Interventionists pull students from class for 30-60 minutes of instruction in order to meet the requirements of Individualized Education Plans (IEPs) developed for students who perform significantly below grade level. All but one of the interventionists in Chicago Heights taught reading. The remaining interventionist taught math.

the high-stakes Illinois Standards Achievement Test (ISAT) taken by 3rd-8th graders in March.[14] In the "Team" treatments, rewards were based on an average performance of the teacher's own students and students in a paired classroom in the school that was matched by grade, subject, and students taught. Classrooms were assigned to teams before the randomization and teachers knew who their team teacher was. Teachers in the control group administered an identical set of assessments at the same time, but did not receive incentives based on their students' performance. In our agreement with the teachers' union, teachers who were in the control group in year one had to receive incentives in year two so that everyone who signed up to participate in both years of the program eventually received treatment.

We calculated rewards using the "pay for percentile" methodology developed by Barlevy and Neal (2012). At baseline, we placed each student in a bin with his nine nearest neighbors in terms of pre-treatment test performance.[15] We then ranked each student within his bin according to improvement between his baseline and end of the year test score.[16] Each teacher received an "overall percentile," which was the average of all her incentivized students' percentile ranks within their respective bins. Teachers received $80 per percentile

[14] The results of ISAT were used to determine whether schools were meeting yearly targets under the No Child Left Behind law in place during the experiment. The ThinkLink was administered to 3rd-8th grade students four times a year in September, November, January and May. K-2 students took the test in May only. Each subject test lasted 30-60 minutes and was either taken on the computer (3rd-8th grade students in all schools and 2nd grade students in some schools) or on paper (all K-1 students and some 2nd grade students). All students were tested in math and reading. In addition, 4th and 7th grade students took a science test as they do on ISAT. We proctored all end of the year testing in order to ensure consistency and discourage cheating. In the first year, we used the prepackaged test for all grades. In the second year, we used the prepackaged ThinkLink Test C (the final test) for grades K-2 and we used ThinkLink probes that we created from a bank of questions for grades 3-8 because the district did not purchase Test C that year.

[15] For each student, the nine nearest neighbors are the nine students in the same grade with the closest baseline predicted score to that student. In both years, we administered a baseline test to Kindergarteners in the fall before the program began. The test was a practice version of the Iowa Test of Basic Skills (ITBS). For students without prior year test scores, we use their actual beginning of year score as their baseline score (fall testing was completed before the program began). Students are placed in separate bins for each subject. In order to avoid competition among teachers (or students) in the same school, students are never placed in a bin with students from the same school. Note that it is not a restriction that Student A be in Student B's neighborhood just because Student B is in Student A's neighborhood.

[16] When there is a tie for students to be included in the neighborhood that would lead to there being more than nine comparison students, we use the average final test score of the tied students when calculating the percentile rank.

for a maximum possible reward of $8,000. The expected value of the reward ($4,000) was equivalent to approximately 8% of the average teacher salary in Chicago Heights.[17]

Teachers assigned to the "Gain" treatment received their rewards at the end of the year, much like most previous programs have done (Springer et al. 2010; Fryer, 2013; Glewwe et al., 2010; Muralidharan and Sundararaman, 2011). In the "Loss" treatment, however, the timing changes significantly. Teachers in these treatment arms received $4,000 (i.e., the expected value of the reward) *at the beginning of the year*.[18] Teachers in the "Loss" treatment signed a contract stating that if their students' end of the year performance was below average, they would return the difference between $4,000 and their final reward. If their students' performance was above average, we issued the teacher an additional payment of up to $4,000 for a total of up to $8,000. Thus, "Gain" and "Loss" teachers received identical net payments for a given level of performance. The only difference is the timing and framing of the rewards.

Within the "Gain" and "Loss" groups, teachers were also randomly assigned to receive either individual or team rewards in the first year. Teachers in the individual treatment groups received rewards based on the performance of their own students. The team treatment paired teachers in a school who were closely matched by grade and subject(s) taught. These teachers received rewards based on their average team performance. For example, if teacher A's overall percentile was 60% and teacher B's overall percentile was 40%, then their team average was 50% and each teacher received $4,000.[19]

We introduced the program at the district-wide Teacher Institute Day at the start of the 2010-2011 school year. Teachers had until the end of September (approximately one

[17] Authors' calculations based on the school district's 2010 and 2011 Illinois State Report Card available at http://webprod.isbe.net/ereportcard/publicsite/getReport.aspx?year=2010&code=140161700_e.pdf and http://webprod.isbe.net/ereportcard/publicsite/getReport.aspx?year=2011&code=070161700_e.pdf At the end of the year we rounded up students' percentiles to 100%, 90%, 80% . . . .20%, 10%, so that the average percentile was 55% (rather than 50%) and the average reward was $4,400 (rather than $4,000). Teachers were not informed of this rounding up in advance.

[18] For tax reasons, some teachers requested that we issue the upfront payment in January. Pooling the first and second years, about thirty five percent of teachers in the loss treatment received the upfront reward at the beginning of January.

[19] Ours is the first study to base rewards on teacher pairs. Previous studies have either tested individual or school-wide rewards. Muralidharan and Sundararaman (2011) compare individual and school-wide incentives in small schools averaging approximately three teachers each.

month) to opt-in to the program. In the first year, 105 of the 121 math teachers who were eligible to participate (87%) did so. In the second year, 113 of the eligible 121 (93%) elected to participate. The experiment formally commenced at the end of September after baseline testing was completed. Informational meetings for each of the incentive groups were held in October at which time the incentivized compensation was explained in detail to the teachers. Midway through the school year we provided teachers with an interim report summarizing their students' performance on a midyear assessment test (the results were for informational use only and did not affect teachers' final reward). We also surveyed all participating teachers about their time use, collaboration with fellow teachers and knowledge about the rewards program. See Appendix Table A.1 for details on the project timeline and implementation milestones.

Random Assignment

Before any randomization occurred, we paired all teachers in each school with their closest match by grade, subject(s), and students taught. In the first year, teachers were randomly assigned to one of the four treatments, or the control group, subject to the restriction that teachers in the "team treatments" must be in the same treatment group as his/her teammate. In the second year, teachers were similarly assigned to one of three treatments, or the control group, with the additional constraint that control teachers from the first year could not be control again in the second year.

In year one of the experiment, teachers who taught multiple homerooms were subject to a slightly different procedure. We randomly assigned a subgroup of their classes to one of the treatment groups with the remaining classes assigned to control.[20] As a result, a teacher in the first year of the experiment received incentives based on the performance of some of her classes taught throughout the day and no incentives for others (unless otherwise noted these classes are included in the control group in the analysis).[21] In year

[20] We rewarded teachers of contained classrooms (who teach a single classroom throughout the day) based on the performance of their homeroom on both reading and math (and science in 4th and 7th grades only). We rewarded teachers of rotating classrooms on all incentivized homeroom-subjects they taught. Rotating teachers taught an average of 4.36 classrooms with an average of 3.00 classrooms subject to incentives. Only 1 of the 67 rotating teachers had all of her classes assigned to control.

[21] Excluding these students from the control group increases our estimated treatment effects, though they are qualitatively unchanged. See Appendix Table A.3, panel A, column 4.

two, we randomly assigned all of a teacher's classes to either a treatment group or to control.

Our agreement with the Chicago Heights teachers' union required us to offer the incentive program to all classroom teachers and interventionists who signed up to participate. This presents two complications.

First, teachers in non-tested subjects (i.e. social studies) were required to have the opportunity to earn incentives in the first year. This presents few complications in math; students typically have only one math teacher, so there is nearly a one-to-one mapping between teachers and students. However, since the district does not administer exams in Social Studies, we offered incentives to these teachers based on their students' performance on the Reading exam. At the request of the district, we also based incentives for Language Arts and Writing teachers on student performance on the Reading exam. Moreover, students receiving special education services through an Individualized Education Plan—roughly 11% of the sample—also received additional reading instruction from a reading specialist. Thus, more than one-third of the students in the year one sample have reading teachers in different treatments. Because of the confusion this likely induced among reading teachers and the difficulties that arise in the statistical analysis, due to contamination and lack of power, we focus our discussion on the math results in what follows.[22]

The exposure to multiple teachers in reading is less of a concern in year two of the experiment because we limited eligibility to classroom reading teachers. However, there is a second complication in year two of the experiment. Per our agreement with the teachers' union, teachers could only be assigned to the control group one of two years because we committed to all interested teachers that they would receive treatment one or both years. To address this potential issue, we re-randomized teachers into treatment and control groups at the beginning of year two (Fall 2011) with the constraint that all control teachers in year one must be treated in year two. The complication for the second wave of the experiment is that from year one to year two, teachers moved between treatments and students moved between teachers, so that both teachers and students could be exposed to

[22] We also incentivized 4th and 7th grade science, which is tested on the statewide exam. However, the sample sizes were too small to conduct a meaningful statistical analysis.

one treatment in the first year and a different treatment in the second year. There is also no ``pure control'' group of teachers who never received incentives: the control group in year two is made up of teachers who received incentives in year one or were new to the study in year two, which is a selected sample.[23]

With the above caveats in mind, our randomization procedure for the first year is straightforward. To improve balance among the control group and the treatment arms, over a pure random draw, we re-randomized teachers after the initial draw.[24] First, we calculated a balance statistic for the initial assignments, defined as the sum of the inverse *p*-values from tests of balance across all five groups.[25] Our algorithm then searches for teachers to "swap" until it finds a switch that does not violate any of the rules outlined above. If switching these teachers' treatments would improve the balance statistic, the switch is made; otherwise it is ignored. The algorithm continues until it has tested forty potential swaps. The randomization procedure for the second year is similar except that there is a constraint that does not allow first year control teachers to be control again.

IV. Data and Analysis

Data

Our primary data source is student-level administrative data provided by the Chicago Heights School District (CHSD). These data include information on student gender, race, attendance, eligibility for free or reduced-price lunch, eligibility for Special Education services, Limited English Proficiency (LEP) status, and teacher assignments. Three types of test scores are available. The first set of test scores is ThinkLink, which is

[23] Among teachers assigned to control in year two, teachers new to the study in year two perform approximately 0.3σ worse than teachers who were in the study in year one, significant at the $p < 0.01$ level.

[24] There is an active discussion on which randomization procedures have the best properties. Treasure and MacRae (1998) prefer a method similar to the one described above. Imbens and Wooldridge (2009) and Greevy et al. (2004) recommend matched pairs. Results from simulation evidence presented in Bruhn and McKenzie (2009) suggest that for large samples there is little gain from different methods of randomization over a pure single draw. For small samples, however, matched-pairs, re-randomization (the method employed here), and stratification all perform better than a pure random draw. Following the recommendation of Bruhn and McKenzie (2009), we have estimated our treatment effects including all individual student baseline characteristics used to check balance.

[25] We use chi-squared tests to test for balance at the class level across categorical variables (school, grade, and subject) and rank-sum tests for continuous variables (baseline ThinkLink math score, baseline ThinkLink reading score, percent female, percent black, percent Hispanic, and contact minutes with teacher). In year two, we only balanced on the categorical variables not the continuous variables.

administered to students in all grades, and is the basis of our teacher incentives. Ninety percent of students have a valid end of year ThinkLink math score. Students in third through eighth grades also take the Illinois Standard Achievement Tests (ISAT), a statewide high-stakes exam conducted each spring that determined whether schools were meeting yearly targets under the No Child Left Behind law in place during our experiment. All public-school students were required to take the math and reading tests unless they were medically excused or had a severe disability. Ninety-two percent of students in third to eighth grades have a valid math and reading state test score.[26] Finally, in the first year of our intervention only, students in Kindergarten through second grade took the Iowa Test of Basic Skill (ITBS). This exam was not a high-stakes exam and only 72 percent of eligible students have a valid math and reading ITBS test score. We have administrative data spanning the 2006-2007 to 2015-2016 school years, ThinkLink data for the 2010-11 and 2011-12 school years, ITBS data through the 2010-11 school year, ISAT data through the 2013-14 school year and the statewide exam that replaced ISAT, the Partnership for Assessment of Readiness for College and Career (PARCC) data for the 2014-2015 and 2015-2016 school years. In all analyses, the test scores are normalized (across the school district) to have a mean of zero and a standard deviation of one for each test, grade, and year.

Table 2 presents summary statistics for students in the "Gain" treatment, "Loss" treatment and control group by year.[27] We report group means for the following baseline student characteristics: gender, race/ethnicity, eligibility for free or reduced-price lunch, whether a student receives accommodations for limited English proficiency (LEP), whether a student receives special education services, and baseline student test scores. At the teacher level, we report mean teacher value added in the prior year as measured by students' percentile change on the statewide exam. The value added measure is missing for teachers who were not in the district the year prior to the experiment.

[26] Students with moderate disabilities or limited English proficiency must take both math and reading tests, but may be granted special accommodations (additional time, translation services, alternative assessments, and so on) at the discretion of school or state administrators. In order to ensure that as many students take the test as possible, the state provides a make-up testing window and the principal/district is required to provide the state with a written explanation of why a student registered at a specific school was not tested.

[27] See Appendix Table A.2 for a similar table that partitions the data into the four treatment arms.

Accounting for within-homeroom correlation, the groups are very well balanced within year. Columns (1) through (3) of Table 2 display descriptive statistics on individual student characteristics and baseline teacher value added for our experimental sample. Column (4) provides the *p*-value from the test that the statistics in columns (1), (2), and (3) are equal. The table reinforces that our sample contains almost exclusively poor minority students: 98 percent are eligible for free or reduced-price lunch, and 96 percent are members of a minority group. Of the eight variables, only one is statistically different across the groups.

Columns (5) through (7) report descriptive statistics for year two students of the experiment. As in year one, we are well-balanced on baseline student characteristics. There are marginally significant differences in LEP status and baseline math scores (as noted above, we did not re-randomize to achieve balance on these characteristics in year two). Panel B summarizes the assignments teachers received in the prior year of the experiment: control, loss, gain or new to the study in year two. This panel highlights that there are no Year 1 Control teachers who also receive Control in Year 2. As discussed above, every teacher who participated in both years of the experiment was required to receive incentives at least once. For all other Year 1 assignments, there are no significant differences in Year 2 assignment.

We also administered a survey to teachers towards the end of both school years. The survey included questions about program knowledge, collaboration with fellow teachers and time use. We received a 53% overall response rate (49% in the "Gain" group, 62% in "Loss" group and 36% in Control) in the first year and a 55% response rate (55% in the "Gain" group, 64% in "Loss" group and 31% in Control) in the second year. Finally, we worked with principals and teachers to confirm the accuracy of class rosters.

Experimental Specifications

The results we report are from linear regressions with a variety of test scores as the outcome variable. Included on the right-hand side of the regression is the student's treatment assignment, school and grade fixed effects, demographic and socio-economic characteristics of the student (gender, race/ethnicity, free/reduced lunch status, limited English proficiency status, special education status), baseline test score in the relevant

subject interacted with grade, and teacher value added in the year prior to the experiment. For year two outcomes, we also include controls for the teacher's year one treatment status. We replace missing covariates with zero for dummy variables and the sample mean for continuous variables and include an indicator variable for missing values. The results are qualitatively unchanged if we limit the set of covariates to only include school and grade fixed effects, and baseline test scores; or if, rather than imputing baseline test score, we exclude students who are missing baseline test scores (Appendix Table A.3, columns 1 and 2 respectively).

We present results both estimating years of the experiment separately and pooling across years of data. When pooling the data across years, the control variables are fully interacted with year dummies. We show results for each of our treatment arms separately, as well as pooling the team and individual treatments and pooling the gain and loss treatments. The coefficients we report are Intent-to-Treat estimates, i.e. students are classified based on their initial classroom assignment.

Recall, given our design, it is possible that a student has two or more teachers who face different incentive treatments within the same subject area. Because we focus on math teachers, this inconvenience is easily overcome: 94% of the students in our sample see a single math teacher and only 1.9% are exposed to teachers in different treatments. We include each student-teacher observation (i.e., a student with two teachers is observed twice) and two-way cluster standard errors by student and teacher. Dropping all students exposed to multiple teachers yields qualitatively identical results (Appendix Table A.3, column 3).[28]

One concern in any experiment is missing outcome variables and, in particular, differences in missing data across treatment and control. For instance, if students of incentivized teachers are more (or less) likely to take the incentivized ThinkLink test than those in the control group, then our estimates may be biased even with random assignment. Fortunately, in our setting attrition rates are relatively low and there is little evidence of differential attrition. Table 3 shows results from a linear probability model with an indicator for missing the ThinkLink exam as the dependent variable and the full set of

[28] The situation is significantly more complex for reading, where one-third of the year one sample is exposed to teachers in different treatment arms.

covariates on the right-hand side. Treatment status carries substantively small and statistically insignificant coefficients in both years of our data. There is also no evidence of differential attrition on the statewide ITBS/ISAT standardized tests (which are not incentivized by our study, but for which we report results).

V. Results

Table 4 presents estimates of the overall impact of our treatments on math ThinkLink scores normalized to have a within-grade standard deviation of one.[29] Standard errors clustered at the student and teacher level are in parentheses below each estimate. The number of observations, the number of students and the number of teachers is displayed in the bottom two rows. The rows specify the treatments estimated, and the *p*-value on the difference between the "Pooled Loss" and "Pooled Gain" coefficients is reported at the bottom of the table. Columns 1 and 2 report results for years one and two respectively. Column 3 presents estimates pooled across the two years. The top row of the table pools all treatments relative to control. Subsequent rows show results disaggregated by treatment group.

As shown in the top row of the table, overall our treatments increased test scores by 0.175σ (se = 0.070) in the first year with no impact in the second year. Pooling across years, the overall impact is 0.099σ (0.051).

Rows 2-4 of the table show estimates for the loss treatments, both pooling individual and team treatments (row 2) and showing those separately (rows 3 and 4). The remaining rows in the table have a parallel structure, but report results for the gain treatments. The loss treatments outperform the gain treatments substantially in year one. The estimated impact of the pooled loss treatments is 0.234σ (0.080) compared to an estimated impact of 0.1σ (0.079) of the pooled gain treatments. The difference between the treatment effects is statistically significant at the $p = 0.051$ level as reported in the

[29] Subject to a number of important caveats related to implementation described in Section 3, the estimated effects on reading scores are presented in Appendix Table A.4**.** The table follows the same structure as Table 4 except that we include only one observation per student and students exposed to multiple treatments across classes receive weights for each treatment (e.g., a student exposed to Individual Gain incentives in one class and Team Loss incentives in another class receives a 0.5 weight for Individual Gain and a 0.5 weight for Team Loss). We then cluster the standard errors at the class level. We include the same covariates as in Table 4 except for baseline teacher value added, which is missing for a substantial proportion of teachers.

bottom panel of the table. In year two, however, neither the loss or gain treatments are effective with estimated impacts of 0.021σ (0.079) and 0.006σ (0.106) respectively. Combining the estimates across years, the loss treatment yields bigger estimates than the gain treatment -- 0.124σ (0.056) versus 0.051σ (0.062) -- but the differences are not statistically significant. Within the loss and gain treatments, the estimated impacts of the individual and team treatments are nearly identical in the pooled estimates.[30]

To investigate the heterogeneity of the program's effects, Table 5 presents results split by grade level, gender, race, and baseline test performance. We present the year one estimates in columns 1-2, year two estimates in columns 3-4 and the pooled estimates in columns 5-6. Odd-numbered columns present estimates for the "Loss" treatment; even-numbered columns present the estimates for "Gain." Panel A presents the results for the full sample, repeating the ITT estimates shown in Table 4. Panel B breaks down the results by grade level, Panel C divides the sample by gender, Panel D by race/ethnicity and Panel E according to whether a student's baseline test score was above or below the median baseline score in his grade.

We find suggestive evidence of substantial heterogeneity in treatment effects by grade level in Panel B. For younger students in grades K-2, the estimated impacts of both the loss and gain treatments are economically meaningful in both year one and year two, ranging from 0.154σ (0.133) to 0.490σ (0.185). Pooling across years, the estimated effects of the loss and gain treatments are almost identical, 0.253σ (0.116) and 0.250σ (0.115) respectively. Among 3rd-8th graders, the effects are more muted and only the loss treatment effect in year one is differentiable from zero, 0.165σ (0.059). Whether these findings will prove robust is, of course, an open question. We did not design our experiment expecting to observe such strong heterogeneity across age groups, and the treatment effects are not statistically distinguishable, raising the specter of incorrect inference due to multiple hypothesis testing. In the remaining subgroups, there is little systematic heterogeneity that

[30] Interestingly, we estimate positive impacts of the loss framed incentives on reading scores in year two that are economically meaningful, 0.1σ, but not statistically significant (Appendix Table A.4, column 2). As noted in Section 3, there were fewer complications in reading in year two when we limited enrollment to classroom reading teachers compared to in year one when students were exposed to multiple treatments through multiple teachers.

we are able to detect by gender, race/ethnicity or baseline achievement (Panels C, D and E respectively).

Finally, we examine the long run impact of treatment on teacher performance. We focus on year one treatment because treatments were not effective in year two. Table 6 presents estimates for the treatment year (year one) and five post-treatment years (we treat year two of the experiment as the first post-treatment year). The first row estimates the impact of overall treatment. The second and third rows present estimates for "Loss" and "Gain", respectively. In column 1, we present the year one treatment impact on ThinkLink, repeating the ITT estimates from Table 4. Column 2 presents the year one treatment impact on the unincentivized statewide standardized tests: the Iowa Tests of Basic Skills (ITBS) for grades K-2 and the Illinois Standard Achievement Tests (ISAT) for grades 3-8. In columns 3-8, we estimate the impact of a teacher's year one treatment on her value added in the relevant year. In all regressions we control for students' test scores in the prior year along with the full set of baseline characteristics.

We have ThinkLink scores for grades K-8 in the treatment year (2010/11) and the first post-treatment year (2011/12). We have statewide test scores for K-8 students in the treatment year (ITBS for grades K-2 and ISAT for grades 3-8). For the 2011/12- 2013/14 school years, we have ISAT scores for grades 3-8 (the district stopped administering the ITBS to K-2 students after year one). Starting in the 2014/15 school year, the district administered the Partnership for Assessment of Readiness for College and Careers (PARCC) rather than the ISAT. The PARCC was administered to grades 3-8 in 2014/15 and to grades 2-8 in 2015/16. In each year, we include all teachers who participated in year one of the experiment and whose students appear in the testing data.[31] The final two columns pool estimates using the statewide exams for all years including the treatment year (column 9) and all post-treatment years –i.e., excluding the treatment year (column 10).

The results of Table 6 suggest that the loss treatment had a lasting impact on teacher value added. In year one, the pattern of effects on the statewide tests is similar to the impacts on ThinkLink with the magnitude of the estimates slightly smaller. The estimated impact of the overall treatments is 0.107σ (0.075). As in our main results, the estimated

[31] We find no evidence of differential attrition from the test score data across years between control and treatment teachers (Appendix Table A.5).

effects of the loss treatments are larger than the gain treatments (though the effects are not statistically distinguishable): the estimated impact of the pooled loss treatments is 0.151 (0.084) compared to an estimated impact of 0.048 (0.084) for the gain treatments.[32]

Turning to the first post-treatment year (2011/12) -- which is also year two of the experiment -- the estimated impact on teacher value added of receiving loss incentives the prior year is similar in magnitude to the year one treatment effects, though not statistically significant: 0.156σ (0.098) on ThinkLink and 0.211σ (0.137) on ISAT. Taken together, the five years of post-treatment estimates for the loss treatment are all positive and economically meaningful except for small negative estimates in 2012/13. Pooling across years, the estimated impact of the loss treatment on teacher value added is 0.177σ (0.065) including the treatment year and 0.167σ (0.112) excluding the treatment year. In contrast, we find no impact of the gain treatment when pooling across years. The difference between the pooled effects of the gain and loss treatments is significant at the p=0.003 level including the treatment year and at the p=0.012 level excluding the treatment year.

VI. Discussion

In this section, we first discuss potential mechanisms for the year one treatment effects, in particular the larger impact of the loss treatment compared to the gain treatment. We then turn to a discussion of the null results in year two.

Year one results

We begin with a simple model that incorporates several possible explanations for the year one treatment effects. Let θ denote student performance and r denote the piece rate incentive pay for that performance. Suppose that the teacher production function, $g(e)$, can be written as $g(e) = e + u$, where e represents teacher effort and u is a classic error term. We assume that teacher utility is separable in the cost of effort, $C(e)$, which is twice continuously differentiable and convex, and the utility of money, $v(\cdot)$, which is an

[32] Appendix Table A.6 reports estimated treatment effects on the statewide tests for year one treatment, year two treatment and pooling across year one and year two. The table has the same structure as Table 4. However, we note that in year two we only observe test scores for students in grades 3-8 because, as discussed above, the school district did not administer the ITBS to K-2 students in year two.

increasing function of payments. Without loss of generality, we normalize the piece rate *r* to 1 for ease of exposition.

With these assumptions in hand, a teacher's utility maximization problem can be written as:

$$\max_{e} \int v(e+u-T)f(u)du + v(T) - C(e),$$

where $f(u)$ is the decision weight attached to event *u* and *T* captures the extent to which incentives are front-loaded. In the "Gain" treatment, each teacher is rewarded at the end of the year with performance-based pay equal to $e+u$. This is represented above as $T=0$. In the "Loss" treatment, each teacher initially receives a fixed amount $\Omega > 0$ and then receives $e+u-\Omega$ at the end of the year. This is represented above as $T=\Omega$. Note that if $e+u-\Omega < 0$ then a teacher makes an end of year payment back to the principal. Since the overall reward to the teacher is constant across the two treatments – $(e+u-\Omega)+\Omega = e+u$ – any differences in effort can be attributed to the timing and framing of the payments.

Assuming first order conditions identify the optimum, the solution to the teacher's maximization problem can be shown as: $\int v'(e+u-T)f(u)du = C'(e)$. In words, for a given incentive scheme, teachers choose effort such that the marginal benefit of effort equals the marginal cost of that effort.

This simple model illustrates the difficulties of interpreting our main results. At the most general level, the experiment suggests that front-loading teachers' incentives can lead to increased teacher effort and increased student achievement. In what follows, we use the framework above to explore alternative interpretations of the data.

Credit Constraints

Let $v(\cdot)$ represent a concave neoclassical utility function and suppose teachers are constrained in that they cannot borrow against their end of year consumption, even though increasing beginning of year consumption is welfare enhancing ($v'(T) > \int v'(e+u-T)f(u)du$.) By concavity of $v(\cdot)$, the marginal benefit of effort increases in front-loading *T*. Intuitively, teachers who are not able to save their up-front payment are relatively poorer down the road. Hence, they demand more income and put forth more effort.

It is also possible that front-loading enables teachers to make productivity-enhancing investments in their classroom (say, new workbooks or dry-erase markers). Notice, under perfect credit markets, we would not expect any difference in effects between our gain and loss treatments. Teachers in the loss group could use their upfront check if necessary, and cash-strapped teachers in the gain treatments could borrow money to finance their purchases. If teachers are liquidity-constrained, however, the loss treatment effectively gives them access to a form of financing unavailable to teachers in the gain treatment. It is possible that this mechanism could create the effects that we observe.

Survey evidence, however, does not support this explanation. Table 7 reports treatment effect estimates on several survey outcomes. In both years of the experiment, the amount of personal money spent on classroom materials reported by loss group teachers was statistically indistinguishable from that reported by gain and control teachers. What's more, 80% of teachers in the first year loss treatment and 53% in the second year loss treatment report that they had not spent any money from their checks when they were surveyed in March (three quarters of the way through the given year of the experiment).

Trust

Suppose once more that $v(\cdot)$ is a neoclassical utility function, and that teachers are not credit constrained. However, teachers may not fully trust the experimenters to fulfill the agreement set out at the beginning of the school year. Under these conditions, we can write the maximization problem as:

$$\max_{e} \int P(T)v(e+u)f(u)du + C(e),$$

where *P(T)* denotes teachers' perceived probability that rewards will be paid at the end of the year.

It is plausible that paying some portion of the reward up front builds trust among the teachers and, that therefore $P(\cdot)$ is increasing in *T*. This increases the perceived return to effort, leading to increased production in equilibrium. If this holds, then our results could be explained purely by the role of up-front payments in establishing credibility with teachers.

It is difficult to test this theory without a measure of trust.[33] A similar argument could be made that the effect of the upfront payment operates through salience. Since these interpretations are both consistent with our findings, we present them alongside other explanations and leave the reader to judge the appropriateness of each.

Prospect Theory Preferences

Alternatively, the response could be purely due to framing. Suppose that agents' utility can be expressed using a prospect theory value function, and that teachers therefore value payments as they arrive. In the canonical prospect theory value specification, $v(\cdot)$ is concave in gains but convex in losses. Therefore, as in the credit-constrained example, the concavity in the gains domain increases the effectiveness of the front-loaded incentive. However, the effect is dampened somewhat by the convexity of $v(\cdot)$ in the loss domain. The net change in effort therefore depends on the size of T and the specific functional form specified for $v(\cdot)$.[34]

If $v(\cdot)$ is specified to allow for loss aversion – i.e. there is a kink at $v(0)$ such that teachers prefer avoiding losses to procuring gains – then the role of framing in explaining our results becomes more important. While strict loss aversion is not necessary to explain the results of our experiment, our findings are consistent with over 30 years of psychological and economic research on the power of loss aversion to motivate individual behavior.

Cheating

Finally, one might naturally worry that tying bonuses to test scores might induce certain teachers to cheat. Indeed, Jacob and Levitt (2003) find an uptick in estimated cheating after Chicago Public Schools instituted a performance-bonus system.

[33] Baseline trust may have been fairly high among the teachers. We had worked with the district for several years prior to the experiment, including running a pilot study of the teacher program in which we distributed incentives to all participants. Several of these participants described their experience in the pilot when we introduced the program at the district-wide Teacher Institute Day at the start of the 2010-11 school year.

[34] Armantier and Boly (2015) discuss the predicted effort response to incentives framed as pure gains, pure losses, or a mix of losses and gains (below and above a threshold respectively) under a prospect theory model with both loss aversion and diminishing sensitivity (i.e., utility is convex in losses and concave in gains). They demonstrate both theoretically and empirically that mixed loss/gain incentives, like the "loss" incentives in our experiment, can increase worker effort compared to pure gain or pure loss incentives.

We find this explanation uncompelling, however, primarily because the results on the state tests – for which teachers received no rewards under our scheme and for which the entire school was under pressure to perform – mirror the ThinkLink results. As shown in Table 6, we find positive impacts of the loss treatment on performance on the tests for which teachers did not receive incentives for year one of the experiment, both in the treatment year (2010/11) and through five years of post-program follow up (2011/12 to 2015/16). It seems unlikely that the treatment effects would repeat themselves on the concurrent statewide test and also persist into post-treatment years if differential cheating practices across treatment and control groups were driving our primary results.

Year two results

We now turn to potential explanations for the null results in year two of the experiment. As discussed above, the assignment to treatment in year two was complicated by our agreement with the teachers' union that any teacher who participated in both years of the experiment had to receive incentives in at least one year of the program. As a result, there are no teachers who were in the control group in both years. In addition, both teachers and students were exposed to different treatments across years. As shown in Table 6, we find evidence that the impact of year one treatment persists into year two. Such persistent treatment effects could confound the impact of a new treatment in year two. An alternative explanation is that the motivational power of incentives – and loss-framing in particular – diminishes over time. If this were the case, we might expect that teachers who received the loss treatment in year one would be especially desensitized in year two.

We do not find strong support for either explanation in our results. If either persistence of treatment effects or desensitization to loss framing were driving the null results in year two, we would expect this to be largely due to teachers who were in the loss treatment in year one. However, excluding year one loss teachers from the analysis does not affect the results (Appendix Table A.3, panel B, column 4).[35]

Using our data, we are therefore not able to rule out that the effects of loss framed incentives that we find in the first year may not replicate. In a subsequent study to ours,

[35] Another test of this hypothesis would be to estimate treatment effects among teachers new to the study in year two, but we do not have sufficient sample size to conduct this analysis.

Brownback and Sadoff (2018) test loss framed incentives among community college instructors over two semesters. Instructors remain in the same treatment -- incentives or control -- over both semesters (the study does not include gain framed incentives). They find impacts on student performance similar to ours, 0.2σ, and that incentives are effective in both the first and second semester with *larger* estimated effects in the second semester. These findings suggest that the effects of loss framed incentives may indeed replicate both across contexts and over time.

VII. Conclusion

In this study, we present the results of a two-year field experiment that provides financial incentives to teachers. In contrast to previous experimental studies in the developed world, we find a substantial impact on test scores. We also present evidence that framing a teacher incentive program in terms of losses rather than gains leads to improved student outcomes. The impacts we observe are large – roughly the same order of magnitude as increasing average teacher quality by a standard deviation. The impacts are apparent not only on the tests that determine the teacher payouts, but also on unincentivized state tests. These test score gains also show substantial persistence after the intervention ends. Whether the incentives were tied to individual teachers or to teams of two teachers does not affect student outcomes.

One striking result to emerge from our study is that the impact of incentives – whether framed as losses or gains -- is largest among younger students in second grade and below. If this result proves robust, namely that financial incentives to teachers in early grades increase test scores, then such incentives would be an extremely cost effective educational intervention. The cost per student of the intervention is roughly $200, with a test score increase of one quarter of a standard deviation. These impacts are larger than those of the highly acclaimed Tennessee STAR class size experiment at less than one-fifth the cost per student. Whether these findings will prove robust is, of course, an open question. We did not design our experiment expecting to observe such strong heterogeneity across age groups, raising the specter of incorrect inference due to multiple hypothesis testing.

An open question is why we find very large effects in the first year of the experiment and no impact in the second year. The study design did not lend itself to understanding the precise mechanisms that might underlie the source of improvements (or lack thereof) in teacher performance. As discussed above, Sadoff and Brownback (2018) provides an additional data point on the impact of loss-framed incentives on instructor performance in another context. Given the potential public policy relevance, there would be value in replicating (or disproving) these results, as well as further exploring the underlying mechanisms.

Our findings have implications not only within education, but more broadly. While there is overwhelming laboratory evidence that rewards framed as losses are more effective than rewards framed as gains, there have been few prior field experimental demonstrations of this phenomenon. Our results, along with those of Hossain and List (2012) suggest that there may be significant potential for exploiting loss framing in the pursuit of both optimal public policy and the pursuit of profits.

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): 95-135.

Armantier, Olivier and Amadou Boly. 2015. "Framing of Incentives and Effort Provision." *International Economic Review*, 56(3): 917-938.

Bandiera, Oriana, Iwan Barankay and Imran Rasul. 2013. "Team Incentives: Evidence from a Firm-Level Experiment," *Journal of the European Economic Association*, 11(5): 1079-1114.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2007. "Incentives for Managers and Inequality among Workers: Evidence from a Firm Level Experiment," *Quarterly Journal of Economics*, 122: 729-773.

Barlevy, Gadi and Derek Neal. 2012. "Pay for Percentile." *American Economic Review*, 102(5): 1805-1831.

Barrera-Osorio, Felipe, and Dhushyanth Raju. 2017. "Teacher Performance Pay: Experimental Evidence from Pakistan." *Journal of Public Economics*, 148: 75-91.

Boyd, Donald, Daniel Goldhaber, Hamilton Lanjford, and James Wyckoff. 2007. "The Effect of Certification and Preparation on Teacher Quality." *The Future of Children* 17(1): 45–68.

Brooks, Richard R., Alexander Stremitzer, and Stehpen Tontrup. 2012."Framing Contracts: Why Loss Framing Increases Effort*." Journal of Institutional and Theoretical Economics*, 168(1): 62-82.

Brownback, Andy and Sally Sadoff. 2018. "Improving College Instruction through Incentives." Working paper. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3152028

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1: 200-232.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review* 104(9): 2633-2679.

Chiang, Hanley, Cecilia Speroni, Mariesa Herrmann, Kristin Hallgren, Paul Burkander, and Alison Wellington. 2017. "Evaluation of the Teacher Incentive Fund: Final Report on

Implementation and Impacts of Pay-for-Performance across Four Years. NCEE 2018-4004." *National Center for Education Evaluation and Regional Assistance*.

De Quidt, Jonathan, Francesco Fallucchi, Felix Kolle, Daniele Nosenzo, D., and Simone Quercia. 2017. "Bonus Versus Penalty: How Robust are the Effects of Contract Framing? *Journal of the Economic Science Association*, 3(2):174-182.

Dee, Thomas S., and James Wyckoff. 2015. "Incentives, selection, and teacher performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34(2): 267-297.

DellaVigna, Stefano and Devin Pope. 2016. "What Motivates Effort? Evidence and Expert Forecasts." *The Review of Economic Studies*, 85(2): 1029-1069.

Duflo, Esther, Rema Hanna, and Stephen Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review*, 102(4): 1241-1278.

Englmaier, Florian, Stefan Grimm, David Schindler, D., and Simeon Schudy. 2018. "Effect of Incentives in Non-routine Analytical Team Tasks - Evidence from a Field Experiment." Rationality and Competition Discussion Paper Series 71, CRC TRR 190 Rationality and Competition.

Firestone, William A. and James R. Pennell. 1993. "Teacher Commitment, Working Conditions, and Differential Incentive Policies." *Review of Educational Research*, 63(4): 489-525.

Fryer, Roland G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics*, 31(2): 373-427.

Fryer, Roland G. 2017. "The Production of Human Capital in Developed Countries: Evidence From 196 Randomized Field Experiments." In *Handbook of Economic Field Experiments* 2: 95-322.

Glazerman, Steven, Allison McKie, and Nancy Carey. 2009. "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year One Impact Report." Mathematica Policy Research, Inc.

Glazerman, Steven, Ali Protik, Bing-ru Teh, Julie Bruch, and Jeffrey Max. 2013. "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. NCEE 2014-4004." *National Center for Education Evaluation and Regional Assistance*.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics*, 2(3): 205-227.

Goldhaber, Dan, Cyrus Grout, and Nick Huntington-Klein. 2017. "Screen Twice, Cut Once: Assessing the Predictive Validity of Applicant Selection Tools." *Education Finance and Policy*, 12(2): 197-223.

Goodman, Sarena F., and Lesley J. Turner. 2013. "The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." *Journal of Labor Economics* 31.2: 409-420.

Greevy, Robert, Bo Lu, Jeffrey Silber, and Paul Rosenbaum. 2004. "Optimal Multivariate Matching before Randomization," *Biostatistics,* 5: 263-275.

Hong, Fuhai, Tanjim Hossain and John A. List. 2015. "Framing Manipulations in Contests: A Natural Field Experiment." *Journal of Economic Behavior & Organization*, 118:372-382.

Hossain, Tanjim and John A. List. 2012. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." *Management Science*, 58(12): 2151-2167.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7: 24-52.

Imas, Alex, Sally Sadoff, and Anya Samek. 2016. "Do People Anticipate Loss Aversion?" *Management Science*, 63(5):1271-1284.

Imbens, Guido, and Jeffrey Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, 47: 5-86.

Imberman, Scott A., and Michael F. Lovenheim. 2015. "Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system." *Review of Economics and Statistics* 97(2): 364-386.

Jacob, Brian, Jonah E. Rockoff, J. E., Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2016. "Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools." NBER Working Paper No. 22054.

Jacob, Brian and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 118(3): 843-877.

Johnson, Susan M. 1984. "Merit Pay for Teachers: A Poor Prescription for Reform." *Harvard Education Review*, 54(2): 175-186.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, 47(2): 263-292.

Kane, Thomas J., and Douglas O. Staiger. 2008. “Estimating Teacher Impacts on Student Achievement: An Experimental Validation.” NBER Working Paper No. 14607.

Kraft, Matthew A., David Blazar, and Dylan Hogan. In press. “The Effect of Teaching Coaching on Instruction and Achievement: A Meta-analysis of the Causal Evidence.” *Review of Educational Research.*

Lavy, Victor. 2002. “Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement.” *The Journal of Political Economy*, 110(6): 1286-1317.

Lavy, Victor. 2009. “Performance Pay and Teachers’ Effort, Productivity, and Grading Ethics.” *American Economic Review*, 99(5): 1979-2021.

Levitt, Steven D., John A. List, Susanne Neckermann and Sally Sadoff. 2016. “The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance.” *American Economic Journal: Economic Policy*, 8(4): 183-219.

List, John A. and Anya Samek. 2015. “The Behavioralist as Nutritionist: Leveraging Behavioral Economics to Improve Child Food Choice and Consumption.” *Journal of Health Economics*, 39: 133-146.

Loyalka, Prashant Kumar, Sean Sylvia, Changfang Liu, James Chu and Yaojing Shi. 2016. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." Working Paper. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2775461

Muralidharan, Karthik and Venkatesh Sundararaman. 2011. “Teacher Performance Pay: Experimental Evidence from India.” *Journal of Political Economy*, 119 (1): 39-77.

Neal, Derek. 2011. “The Design of Performance Pay in Education.” In *Handbook of the Economics of Education*, 4: 495-550.

Krueger, Alan B. 1999. “Experimental Estimates of Education Production Functions.” *The Quarterly Journal of Economics*, 114(2): 497-532.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, Schools and Academic Achievement." *Econometrica*, 73(2): 417-458.

Rockoff, Jonah E. 2004. “The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.” *American Economic Review*, 94(2): 247-252

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane and Douglas O. Staiger, 2011. “Can You Recognize an Effective Teacher when You Recruit One?” *Education Finance and Policy*, 6(1): 43-71.

Shearer, Bruce. 2004. "Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment." *The Review of Economic Studies* 71(2): 513-534.

Sojourner, Aaron J., Elton Mykerezi, and Kristine L. West. 2014. "Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-comp in Minnesota." *Journal of Human Resources* 49(4): 945-981.

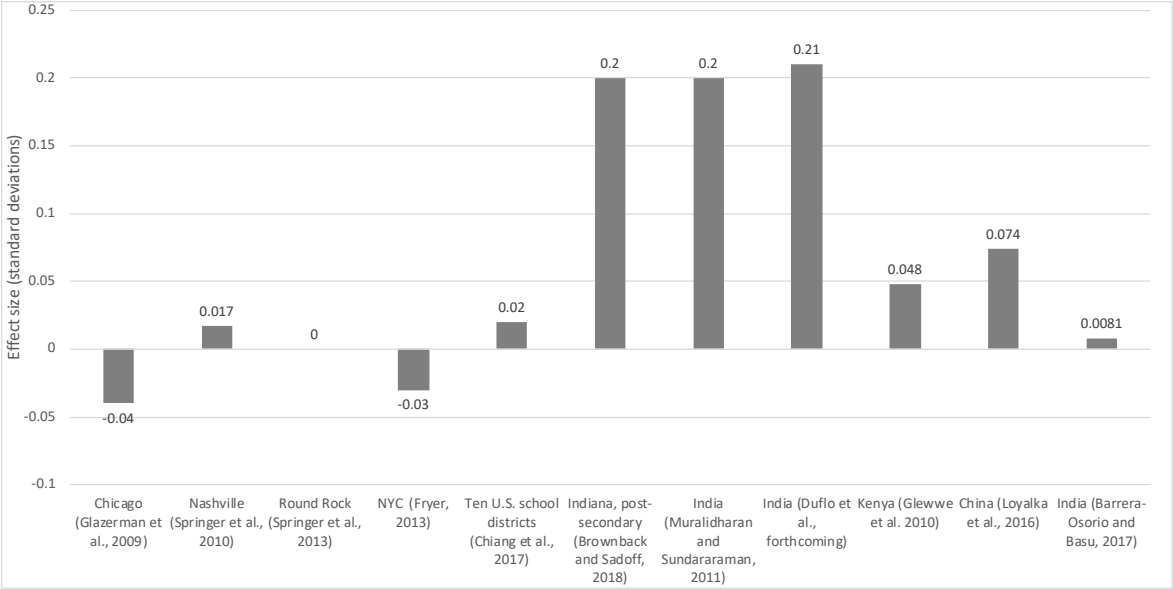
Springer, Matthew G., Dale Ballou, Laura S. Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2011. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT)." *Society for Research on Educational Effectiveness*.

Springer, Matthew, John Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan Burns, Laura Hamilton, and Brian M. Stecher. 2012. "Team Pay for Performance." *Educational Evaluation and Policy Analysis*, 34(4): 367-390.

Treasure, Tom, and Kenneth MacRae. 1998. "Minimisation: The Platinum Standard for Trials?" *British Medical Journal*, 317: 317-362.

Vigdor, Jacob L. 2008. "Teacher Salary Bonuses in North Carolina." Conference paper, National Center on Performance Incentives.

Figure 1: Effects of Teacher Performance Pay Programs on Student Achievement



Notes: The figure presents the estimated effects of (pooled) incentive treatments on standardized math performance (when available) in the first year of the experiment. Brownback and Sadoff (2018) estimate effects pooling across final exams in multiple post-secondary departments. Glewwe et al. (2010) and Barrera-Osorio and Basu (2017) estimate effects pooling subjects on a government exam.

Table 1: Summary of Teacher Incentive Program

Panel A: Overview		
Schools	Nine K-8 schools in Chicago Heights, IL	
First Year Operations	$632,960 distributed in incentive payments, 90% opt-in rate.	
Second Year Operations	$474,720 distributed in incentive payments, 94% opt-in rate.	
Panel B: Outcomes of Interest		
	Subjects and Grades	Date of Assessment
Thinklink Learning Diagnostic Assessment (ThinkLink)	Math (K-8), Reading (K-8), and Science (4 and 7)	May 2011 and May 2012
Illinois Standards Achievement Test (ISAT)	Math (3-8), Reading (3-8), and Science (4 and 7)	March 2011 and March 2012
Iowa Test of Basic Skills (ITBS)	Math (K-2) and Reading (K-2)	March 2011
Panel C: Treatment Details		
	Timing of Rewards	Basis For Rewards
Individual Loss	Teachers receive $4,000 check in October; must pay back difference in June	Teacher's own students
Individual Gain	Teachers paid in full in June	Teacher's own students
Team Loss	Teachers receive $4,000 check in October; must pay back difference in June	Teacher's and teammate's students
Team Gain	Teachers paid in full in June	Teacher's and teammate's students
All Treatments	Treated teachers earned between $0 and $8,000 in bonus payment based on students' performance relative to nine statistically similar students in one of the other eight schools. Rewards are linear in a student's rank, so the expected value of the reward is $4,000.	

Notes: This table presents a summary of the two-year teacher incentive experiment conducted in Chicago Hieghts, IL.

Table 2: Summary Statistics by Treatment

	Year 1				Year 2			
	Control	Loss	Gain	*p-val*	Control	Loss	Gain	*p-val*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Pre-Randomization Characteristics								
Female	0.493	0.491	0.486	0.963	0.494	0.474	0.496	0.563
	(0.500)	(0.500)	(0.500)		(0.500)	(0.499)	(0.500)	
Black	0.396	0.421	0.358	0.559	0.413	0.365	0.418	0.698
	(0.490)	(0.494)	(0.480)		(0.493)	(0.482)	(0.494)	
Hispanic	0.553	0.521	0.577	0.561	0.535	0.570	0.533	0.798
	(0.498)	(0.500)	(0.494)		(0.499)	(0.495)	(0.499)	
Free or Reduced Lunch	0.981	0.954	0.951	0.005***	0.948	0.944	0.960	0.715
	(0.138)	(0.209)	(0.216)		(0.221)	(0.231)	(0.196)	
Limited English Proficiency	0.137	0.091	0.131	0.616	0.167	0.196	0.091	0.062*
	(0.344)	(0.287)	(0.337)		(0.373)	(0.397)	(0.288)	
Special Education services	0.138	0.131	0.099	0.238	0.117	0.117	0.100	0.727
	(0.345)	(0.338)	(0.299)		(0.322)	(0.321)	(0.301)	
Standardized Baseline Math Score	0.027	-0.033	0.093	0.571	-0.201	0.002	0.104	0.097*
	(1.022)	(0.993)	(1.041)		(0.931)	(1.020)	(0.873)	
Standardized Baseline Reading Score	0.017	-0.109	-0.001	0.523	-0.038	0.044	0.137	0.512
	(1.065)	(0.938)	(0.978)		(0.944)	(1.065)	(0.921)	
Teacher Value Added	10.530	10.497	13.899	0.659	15.356	12.167	18.080	0.316
	(15.174)	(18.757)	(18.229)		(14.463)	(23.729)	(5.743)	
Value Added Measure Missing	0.200	0.261	0.128	0.365	0.394	0.367	0.467	0.843
	(0.400)	(0.440)	(0.335)		(0.489)	(0.482)	(0.499)	
Panel B: First-Year Teacher Assignments								
Control	—	—	—		0.000	0.119	0.071	0.000***
Loss	—	—	—		0.450	0.324	0.476	0.000***
Gain	—	—	—		0.316	0.349	0.222	0.000***
New	—	—	—		0.235	0.208	0.231	0.250
Observations	700	1198	1059		703	1685	553	
Joint F-Test from Panel A				0.400				0.273

Notes: This table presents summary statistics and balance tests for baseline observables and pretreatment Thinklink scores. Panel A reports means for demographic variables controlled for in our main regression specification. Panel B reports means of the first year assignments for teachers included in the second year of our experiment. If a teacher was not in the first year of the experiment, they are labeled as "New". Columns (4) and (8) display p-values from a test of equal means in the three previous columns, with standard errors clustered both at the teacher and the student level. We also report the p-value from a joint F-test of the null hypothesis that there are no differences between treatment and control groups across all reported demographics in Panel A, estimated via seemingly unrelated regressions.

Table 3: Attrition

	Year 1		Year 2		Pooled	
	Loss	Gain	Loss	Gain	Loss	Gain
	(1)	(2)	(3)	(4)	(5)	(6)
Missing Thinklink Math Score	0.004	-0.019	0.007	0.007	0.005	-0.009
	(0.018)	(0.019)	(0.011)	(0.013)	(0.010)	(0.012)
	N = 2953		N = 2941		N = 5894	
Missing ITBS/ISAT Math Score	0.015	0.006	-0.010	-0.008	0.006	0.001
	(0.017)	(0.021)	(0.018)	(0.016)	(0.013)	(0.015)
	N = 2953		N = 2022		N = 4975	

Notes: The results shown are estimated from a linear probability model where we regress the relevant dependent variable reported for each row on treatment indicators (pooled loss and pooled gain reported in each column), our full list of control variables summarized in Panel A of Table 2, and dummy variables for each student's school and grade. Columns (1) and (2) present estimates for the first year of the experiment, columns (3) and (4) present estimates for the second year of the experiment, and columns (5) and (6) present estimates obtained by pooling data from both years together. Standard errors reported in parentheses are clustered at the teacher and student level. The number of observations is reported below the standard errors. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 4: The Effect of Treatment on ThinkLink Math Scores

	Year 1	Year 2	Pooled
	(1)	(2)	(3)
Any Treatment	0.175**	0.017	0.099*
	(0.070)	(0.078)	(0.051)
Pooled Loss	0.234***	0.021	0.124**
	(0.080)	(0.079)	(0.056)
Individual Loss	0.271***	0.000	0.126**
	(0.087)	(0.089)	(0.060)
Team Loss	0.197*	0.044	0.122*
	(0.106)	(0.086)	(0.067)
Pooled Gain	0.100	0.006	0.051
	(0.079)	(0.106)	(0.062)
Individual Gain	0.086	0.007	0.053
	(0.096)	(0.106)	(0.072)
Team Gain	0.115		0.046
	(0.085)		(0.073)
Pr(Gain=Loss)	0.051	0.862	0.178
Observations	2630	2697	5327
Students	2460	2543	3279
Classrooms	135	153	288
Teachers	105	113	131

Notes: The results we report are from linear regressions with ThinkLink test scores as the outcome variable. Included on the right-hand side of the regression is the student's treatment assignment, demographic and socio-economic characteristics of the student (gender, race, eligibility for free or reduced price lunch, Limited English Proficiency status, eligibility for Special Education services), school and grade fixed effects, once-lagged test scores interacted with grade, and once-lagged teacher value added. We impute missing data with zeros, adding indicator variables for missing values (missing value added measures are replaced with the sample mean). Year 2 estimates additionally control for the teacher's Year 1 treatment status. In pooled estimates, the control variables are fully interacted with year dummies. Standard errors reported in parentheses are clustered on the teacher and student level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 5: The Effect of Treatment on ThinkLink Scores within Demographic Subgroups

	Year 1		Year 2		Pooled	
	Loss	Gain	Loss	Gain	Loss	Gain
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Full Sample	0.234***	0.100	0.021	0.006	0.124**	0.051
	(0.072)	(0.078)	(0.079)	(0.100)	(0.055)	(0.063)
Panel B: Grade Level						
K-2	0.490**	0.397**	0.154	0.223	0.253**	0.250**
	(0.185)	(0.155)	(0.133)	(0.190)	(0.116)	(0.115)
3-8	0.165***	0.067	-0.043	-0.109	0.071	-0.011
	(0.059)	(0.068)	(0.098)	(0.128)	(0.056)	(0.070)
p-value	0.330	0.521	0.114	0.078	0.162	0.280
Panel C: Gender						
Male	0.200**	0.163*	0.113	0.066	0.155**	0.119*
	(0.081)	(0.084)	(0.089)	(0.111)	(0.061)	(0.068)
Female	0.294***	0.071	-0.062	-0.033	0.106*	0.011
	(0.080)	(0.089)	(0.088)	(0.111)	(0.062)	(0.072)
p-value	0.102	0.645	0.028	0.326	0.415	0.184
Panel D: Race						
Black	0.225**	-0.100	0.033	-0.015	0.107	-0.060
	(0.103)	(0.118)	(0.127)	(0.148)	(0.086)	(0.097)
Hispanic	0.317***	0.093	-0.018	0.021	0.111*	0.043
	(0.099)	(0.091)	(0.084)	(0.104)	(0.066)	(0.072)
p-value	0.845	0.291	0.988	0.320	0.941	0.173
Panel E: Baseline Scores						
Above Median	0.182**	0.153*	0.020	0.020	0.105*	0.096
	(0.078)	(0.079)	(0.091)	(0.114)	(0.061)	(0.067)
Below Median	0.132*	-0.024	0.008	-0.074	0.063	-0.055
	(0.078)	(0.090)	(0.089)	(0.109)	(0.060)	(0.071)
p-value	0.214	0.693	0.441	0.459	0.989	0.333

Notes: This table reports the effect of the treatment on ThinkLink scores, estimated separately for various subgroups in the data. Included on the right-hand side of each regression are the same set of control variables as used in Table 3. We present results both estimating years of the experiment separately and pooling across years of data. The coefficients we report are Intent-to-Treat estimates, i.e. students are classified based on their initial classroom assignment. We also report p-values from tests of equal coefficients between grade level groups, genders, races and baseline test score groups. Standard errors reported in parentheses are clustered on the teacher and student level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 6: The Long-Term Impact of Treating Teachers on their Value Added

	Treatment Year		Post-Treatment Years						Pooled	
	2010/11		2011/12		2012/13	2013/14	2014/15	2015/16	2010/11-2015/16	2011/12-2015/16
	ThinkLink	ITBS/ISAT	ThinkLink	ISAT	ISAT	ISAT	PARCC	PARCC	ITBS/ISAT/PARCC	ISAT/PARCC
	K-8	K-8	K-8	3-8	3-8	3-8	3-8	2-8	K-8	2-8
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Any Treatment	0.175**	0.107	0.044	0.098	-0.191	0.026	0.553***	0.293*	0.086	0.049
	(0.070)	(0.075)	(0.097)	(0.148)	(0.187)	(0.150)	(0.193)	(0.176)	(0.065)	(0.115)
Loss	0.234***	0.151*	0.156	0.211	-0.068	0.098	0.813***	0.207	0.177***	0.167
	(0.080)	(0.084)	(0.098)	(0.137)	(0.179)	(0.163)	(0.199)	(0.198)	(0.065)	(0.112)
Gain	0.100	0.048	-0.055	0.022	-0.275	0.013	0.530***	0.296*	0.007	-0.007
	(0.079)	(0.084)	(0.101)	(0.151)	(0.201)	(0.155)	(0.177)	(0.175)	(0.074)	(0.116)
Pr(Gain=Loss)	0.051	0.168	0.013	0.032	0.147	0.418	0.041	0.498	0.003	0.012
Observations	2630	2552	2115	1498	1296	1150	1079	856	8446	5894
Students	2460	2367	1973	1368	1270	1139	1078	855	3953	3281
Teachers	105	105	87	52	41	36	36	36	105	65

Notes: The results we report are from linear regressions with various standardized test scores as the outcome variable. Included on the right-hand side of the regression is the Year 1 treatment assignment of the student's teacher, and the same set of control variables as in Table 3. We show results pooling the impact of any Year 1 treatment, as well as pooled loss and pooled gain treatments separately. We also report p-values from tests of equal coefficients between the loss and gain treatments. Column headers indicate the year the test was taken, the test name, and the grades for which test scores are available. Standard errors reported in parentheses are clustered on the teacher and student level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 7: Teacher Survey Results

	Year 1		Year 2	
	Gain	Loss	Gain	Loss
	(1)	(2)	(3)	(4)
Hours Grading	-0.328	-1.647	-0.682	-0.665
	(1.438)	(1.440)	(1.027)	(0.790)
	82		82	
Hours Calling or Meeting w/ Parents	0.138	-0.071	-0.455	-0.306
	(0.458)	(0.459)	(0.518)	(0.399)
	82		82	
Hours Tutoring Outside of Class	1.092	0.953	0.182	-0.858
	(1.742)	(1.744)	(1.086)	(0.836)
	82		82	
Hours Leading Extracurricular Activities	0.585	0.555	-0.273	-0.107
	(1.330)	(1.332)	(1.225)	(0.943)
	82		82	
Hours Completing Administrative Work	-1.587*	-1.258	-0.364	0.835
	(0.936)	(0.938)	(1.403)	(1.079)
	82		82	
Hours Completing Professional Development Coursework	0.464	0.084	0.727	1.018
	(1.392)	(1.394)	(2.026)	(1.558)
	82		82	
Personal Money Spent on Class Materials ($)	-11.026	-109.474	-34.091	47.342
	(115.917)	(116.090)	(102.521)	(78.963)
	82		81	

Notes: This table presents results gathered from surveys of teachers in our experimental group at the end of each school year. Columns (1) and (2) report the results for the first year of the experiment and columns (3) and (4) report the results for the second year of the experiment. All coefficients are derived by regressing the outcome variable in the first column on two dummy variables that indicate if the teacher participated in the Gain or Loss treatment arms for the given year. The sample size for each regression in the first year is 82 teachers. The sample size for each regression in the second year is 81 teachers. Teachers are considered to have participated in either type of treatment if they receive that type of incentive based on any of their students' performance. Standard errors are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

1 Appendix Figures and Tables

Table A.1: Project Timeline and Implementation Milestones

Month	Year 1	Year 2
August	Announcement of Program	—
September	ThinkLink 1 Teacher Opt-in Deadline — —	Invite Teachers to the Second Year ThinkLink 1 Teacher Opt-in Deadline Information Sessions
October	Information Sessions Loss Teachers Paid	Loss Teachers Paid —
November	ThinkLink 2	ThinkLink 2
December	—	—
January	Tax Deferred Loss Teachers Paid ThinkLink 3	Tax Deferred Loss Teachers Paid ThinkLink 3
February	—	—
March	Interim Report Provided State Testing Teacher Survey	Interim Report Provided State Testing —
April	—	Teacher Survey
May	Incentivized ThinkLink	Incentivized ThinkLink
June	Teachers Paid	Teachers Paid
July	—	—

Notes: This table presents the major implementation milestones for the two years of the experiment.

Table A.2: Summary Statistics by All Treatment Arms

	Control	Ind. Loss	Team Loss	Ind. Gain	Team Gain	*p-val*
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Year 1						
Female	0.493	0.496	0.487	0.471	0.502	0.854
Black	0.396	0.454	0.391	0.296	0.423	0.299
Hispanic	0.553	0.478	0.560	0.632	0.521	0.297
Free or Reduced Lunch	0.981	0.942	0.966	0.940	0.962	0.012**
Limited English Proficiency	0.137	0.084	0.096	0.110	0.153	0.804
Special Ed	0.138	0.090	0.171	0.098	0.100	0.030**
Standardized Baseline Math Score	0.027	0.019	-0.087	0.212	-0.031	0.545
Standardized Baseline Reading Score	0.017	0.069	-0.291	0.116	-0.113	0.044**
Teacher Value Added	10.530	6.621	14.827	14.316	13.479	0.631
Value Added Measure Missing	0.200	0.218	0.304	0.118	0.139	0.696
Panel B: Year 2						
Female	0.494	0.466	0.481	0.496	—	0.677
Black	0.413	0.321	0.403	0.418	—	0.443
Hispanic	0.535	0.614	0.532	0.533	—	0.482
Free or Reduced Lunch	0.948	0.941	0.946	0.960	—	0.844
Limited English Proficiency	0.167	0.169	0.219	0.091	—	0.132
Special Ed	0.117	0.095	0.136	0.100	—	0.549
Standardized Baseline Math Score	-0.201	-0.075	0.068	0.104	—	0.120
Standardized Baseline Reading Score	-0.038	0.064	0.026	0.137	—	0.703
Teacher Value Added	15.356	16.648	8.804	18.080	—	0.302
Value Added Measure Missing	0.394	0.414	0.327	0.467	—	0.879
Panel A Observations	700	597	601	525	534	
Panel A Classrooms	36	35	28	23	24	
Panel A Joint F-Test						0.002
Panel B Observations	703	780	905	553	—	
Panel B Classrooms	45	51	56	30	—	
Panel B Joint F-Test						0.024

Notes: This table presents summary statistics and balance tests for baseline observables and pretreatment Thinklink scores. Panel A reports means for year 1. Panel B reports means for year 2 Column (6) displays p-values from a test of equal means in the five previous columns, with standard errors clustered both at the teacher and the student level. We also report the p-value from a joint F-test of the null hypothesis that there are no differences between treatment and control groups across all reported demographics in the respective panel, estimated via seemingly unrelated regressions.

Table A.3: The Effect of Treatment on ThinkLink Math Scores: Sensitivity Checks

	Limited set of covariates (1)	Exclude students missing covariates (2)	Exclude students with multiple teachers (3)	Exclude spillover classes (4)
Panel A: Year 1				
Any Treatment	0.168**	0.094	0.213***	0.244**
	(0.072)	(0.060)	(0.081)	(0.113)
Pooled Loss	0.225***	0.140**	0.295***	0.345***
	(0.083)	(0.066)	(0.095)	(0.119)
Individual Loss	0.245***	0.144**	0.337***	0.386***
	(0.087)	(0.070)	(0.092)	(0.130)
Team Loss	0.205*	0.141	0.248*	0.313**
	(0.112)	(0.089)	(0.138)	(0.130)
Pooled Gain	0.098	0.037	0.128	0.178
	(0.083)	(0.071)	(0.088)	(0.114)
Individual Gain	0.091	0.015	0.113	0.156
	(0.097)	(0.079)	(0.104)	(0.125)
Team Gain	0.105	0.062	0.140	0.208*
	(0.093)	(0.088)	(0.094)	(0.119)
Pr(Gain=Loss)	0.080	0.105	0.020	0.017
Observations	2630	2408	2280	2235
Students	2460	2247	2280	2126
Classrooms	135	134	130	119
Teachers	105	104	102	104
	Limited set of covariates	Exclude students missing covariates	Exclude students with multiple teachers	Exclude year 1 Loss teachers
Panel B: Year 2				
Any Treatment	-0.006	-0.013	0.029	-0.003
	(0.081)	(0.074)	(0.080)	(0.106)
Pooled Loss	-0.015	-0.012	0.048	0.013
	(0.083)	(0.075)	(0.079)	(0.112)
Individual Loss	-0.039	-0.011	0.027	0.019
	(0.095)	(0.086)	(0.088)	(0.119)
Team Loss	0.008	-0.013	0.071	0.001
	(0.089)	(0.081)	(0.087)	(0.124)
Pooled Gain	0.032	-0.019	-0.046	-0.038
	(0.109)	(0.092)	(0.107)	(0.121)
Individual Gain	0.034	-0.019	-0.047	-0.038
	(0.109)	(0.092)	(0.107)	(0.121)
Pr(Gain=Loss)	0.595	0.919	0.250	0.611
Observations	2697	2421	2391	1661
Students	2543	2291	2391	1615
Classrooms	153	153	147	101
Teachers	113	113	110	74

Notes: The results we report correspond to our main specification in Table 4 with estimates for year 1 in Panel A and estimates for year 2 in Panel B. Column (1) includes only indicators for treatment, school and grade, and baseline math test scores interacted with grade. Column (2) includes the same set of covariates as in column (1) and excludes students with missing baseline test scores. Column (3) includes the same covariates as in Table 4 and and excludes students who were taught by multiple teachers participating in the experiment. In column (4), for year 1 we report results after excluding students in "spillover classes," which were not incentivized but were taught by teachers receiving incentives for other classes; for year 2, we exclude students taught by teachers assigned to the Loss treatment in year 1. Standard errors are reported in parentheses and are clustered on the student and teacher level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table A.4: The Effect of Treatment on ThinkLink Reading Scores

	Year 1	Year 2	Pooled
	(1)	(2)	(3)
Any Treatment	0.005	0.091	0.028
	(0.067)	(0.067)	(0.040)
Pooled Loss	0.058	0.102	0.053
	(0.082)	(0.069)	(0.045)
Individual Loss	-0.103	0.095	0.015
	(0.109)	(0.077)	(0.057)
Team Loss	0.164*	0.109	0.086
	(0.088)	(0.079)	(0.058)
Pooled Gain	-0.036	0.039	-0.024
	(0.076)	(0.088)	(0.058)
Individual Gain	0.039	0.039	0.038
	(0.092)	(0.088)	(0.066)
Team Gain	-0.175**	—	-0.165**
	(0.088)		(0.081)
Probability(Gain=Loss)	0.269	0.394	0.221
Observations	2556	2561	5117
Clusters	142	150	292

Notes: The results we report are from linear regressions with ThinkLink test scores as the outcome variable. Included on the right-hand side of the regression is the student's treatment assignment, demographic and socio-economic characteristics of the student (gender, race, eligibility for free or reduced price lunch, Limited English Proficiency status, eligibility for Special Education services), school and grade fixed effects, and once-lagged test scores interacted with grade. We impute missing data with zeros, adding indicator variables for missing values. Year 2 estimates additionally control for the teacher's Year 1 treatment status. In pooled estimates, the control variables are fully interacted with year dummies. Standard errors reported in parentheses are clustered on the classroom level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table A.5: Teacher Attrition

	2011/12	2012/13	2013/14	2014/15	2015/16
	(1)	(2)	(3)	(4)	(5)
Any Treatment	-0.111	0.011	-0.078	-0.067	0.067
	(0.106)	(0.138)	(0.140)	(0.140)	(0.138)
Gain	-0.127	0.019	-0.115	-0.068	0.048
	(0.114)	(0.150)	(0.151)	(0.152)	(0.149)
Loss	-0.096	0.004	-0.044	-0.065	0.084
	(0.113)	(0.148)	(0.150)	(0.150)	(0.148)
N	105	105	105	105	105

Notes: The table reports results from linear probability models estimating the impact of Year 1 treatment status on the probability that a teacher is missing student test score data from our sample in the subsequent years reported for each column. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table A.6: The Effect of Treatment on Statewide ITBS/ISAT Math Scores

	Year 1 ITBS/ISAT Grades K-8 (1)	Year 2 ISAT Grades 3-8 (2)	Pooled ITBS/ISAT Grades K-8 (3)
Any Treatment	0.107	-0.054	0.047
	(0.075)	(0.093)	(0.056)
Pooled Loss	0.151*	0.017	0.100
	(0.084)	(0.092)	(0.062)
Individual Loss	0.136	0.009	0.090
	(0.093)	(0.110)	(0.069)
Team Loss	0.179*	0.027	0.121
	(0.107)	(0.101)	(0.077)
Gain	0.048	-0.179*	-0.032
	(0.084)	(0.107)	(0.065)
Individual Gain	-0.009	-0.180*	-0.079
	(0.093)	(0.108)	(0.070)
Team Gain	0.108		0.065
	(0.102)		(0.089)
Pr(Gain=Loss)	0.168	0.006	0.017
Observation	2552	1896	4448
Students	2367	1758	2747
Classrooms	135	106	241
Teachers	105	69	122

Notes: The results we report are from linear regressions with state test scores (ITBS and ISAT) as the outcome variable. Included on the right-hand side of the regression is the student's treatment assignment, demographic and socio-economic characteristics of the student (gender, race, eligibility for free or reduced price lunch, Limited English Proficiency status, eligibility for Special Education services), school and grade fixed effects, once-lagged test scores interacted with grade, and once-lagged teacher value added. We impute missing data with zeros, adding indicator variables for missing values (missing value added measures are replaced with the sample mean). Year 2 estimates additionally control for the teacher's Year 1 treatment status. In pooled estimates, the control variables are fully interacted with year dummies. Standard errors reported in parentheses are clustered on the teacher and student level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.