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

By Roland G. Fryer, Jr., Steven D. Levitt, John List, and Sally Sadoff*

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 up-front 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 the effects on teacher value added persist post-treatment.* (JEL C93, D91, I21, J33, M52)

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 0.24σ in reading (Rockoff 2004; Rivkin, Hanushek and Kain 2005; Aaronson, Barrow and Sander 2007; Kane and Staiger 2008). These effects are comparable to reducing class size by about one-third (Krueger 1999). Similarly, Chetty, Friedman and Rockoff (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 (Rivkin, Hanushek and Kain 2005; Aaronson, Barrow and Sander 2007; 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

* Fryer: Harvard University, 1805 Cambridge Street, Cambridge, MA 02138 (email: rfryer@fas.harvard.edu). Levitt: University of Chicago, 5807 S Woodlawn Avenue, Chicago, IL 60637 (email: slevitt@uchicago.edu). List: University of Chicago, 5757 S. University Avenue, Chicago, IL 60637 (email: jlist@uchicago.edu). Sadoff: UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (email: ssadoff@ucsd.edu). We are grateful to Tom Amadio and the Chicago Heights teachers' union for their support in conducting our experiment. Charlie Brobst, Eszter Czibor, Jonathan Davis, Matt Davis, Sean Golden, Wooju Lee, William Murdock III, Lina Ramirez, Steven Shi, Phuong Ta, Haruka Uchida and Atom Vayalinkal 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.

[1] More recently, higher cost more intensive screening mechanisms show promise (Jacob et al. 2018; Goldhaber, Grout and Huntington-Klein 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, Blazar and Hogan 2018).

[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. (2020) evaluate the effectiveness of teacher performance pay programs implemented through the Teacher Incentive Fund.

districts have implemented teacher incentive programs in an effort to increase student achievement (Fryer 2013, 2017). As we discuss in detail in the next section, the empirical evidence on the effectiveness of teacher incentive programs is mixed (Neal 2011; Fryer 2017; Pham, Nguyen and Springer 2020, provide reviews). Focusing on experimental studies, in developing countries where the degree of teacher 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 (Glewwe, Ilias and Kremer 2010; Muralidharan and Sundararaman 2011; Duflo, Hanna and Ryan 2012; Loyalka et al. 2019), 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, McKie and Carey 2009; Springer et al. 2011, 2012; Fryer 2013; Chiang et al. 2020).[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 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

[4]In subsequent work using an incentive design similar to ours, Brownback and Sadoff (2020) find large impacts of incentives for instructors at a U.S. community college.

[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, Stremitzer and Tontrup 2012; Hossain and List 2012, team data; Hong, Hossain and List 2015; Armantier and Boly 2015; Imas, Sadoff and Samek 2016; Levitt et al. 2016; Bulte, List and van Soest 2020) while others find little impact of framing (Hossain and List 2012, individual data; List and Samek 2015; DellaVigna and Pope 2017; De Quidt et al. 2017; Englmaier et al. 2018) More recently, Pierce, Rees-Jones and Blank (2020) find a negative impact of loss-framing.

[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]As we discuss in detail in Section 3, our agreement with the Chicago Heights teachers' union required us to offer every teacher the opportunity to participate in the incentive program. This requirement led to an experimental structure that likely contaminated our Reading results but 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.

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 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). The post-treatment effects of the loss- and gain-framed incentives are significantly different at the $p = 0 < 0.01$ level ($p = 0.012$ including the treatment year). 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 ($0.15 - 0.49\sigma$) 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. Our study also demonstrates that loss-framed incentives can be implemented in schools: the teachers' union agreed to the structure of the contracts, about 90% of the teachers opted to participate, and participation rates increased in the second year of the program after teachers had experienced the incentives in year one.

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 and the implementation of loss-framed incentives as a policy. The final section concludes. There are two online appendices. Online Appendix B 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 C is a detailed implementation guide that describes how the experiment was implemented and milestones reached.

I. 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 com-

petition between teachers in what some believe to be a collaborative environment Johnson (1984) and Firestone and Pennell (1993).

Despite the controversy, there is a growing literature on the role of teacher incentives on student performance, including an emerging literature on the optimal design of such incentives (Neal 2018; Pham, Nguyen and Springer 2020, provide recent discussion). Figure 1 displays the standardized treatment effects on student achievement from 14 experimental studies and two meta-analyses: Pham, Nguyen and Springer (2020) and Fryer (2017). Pham, Nguyen and Springer (2020) includes 37 studies from both U.S. and non-U.S. contexts. Across the 11 experimental and 26 non-experimental studies, they estimate an average impact of 0.053σ (0.012). Consistent with the experimental studies discussed in more detail below, the average estimated impact of teacher incentives is larger in studies outside the U.S. compared to U.S. studies, 0.113σ (0.039) and 0.043σ (0.010) respectively. Fryer (2017) finds similarly small effects of teacher incentives in the U.S. in a meta-analysis limited to experimental studies (including ours): 0.022σ (0.022) in math and -0.006σ (0.012) in reading.[8]

The prior and concurrent studies that provide experimental estimates of the causal impact of teacher performance pay incentives on student achievement include, outside the U.S.: Glewwe, Ilias and Kremer (2010); Muralidharan and Sundararaman (2011); Duflo, Hanna and Ryan (2012); Barrera-Osorio and Raju (2017); and, in the U.S.: Glazerman, McKie and Carey (2009); Springer et al. (2011, 2012); Fryer (2013)—along with Marsh et al. (2011) and Goodman and Turner (2013) using the same data—and Chiang et al. (2020).[9] Subsequent to our work, four additional experimental studies provide evidence related to our design: in a developing country context, Mbiti et al. (2019); Loyalka et al. (2019); Gilligan et al. (2019); and in a U.S. post-secondary context, Brownback and Sadoff (2020).[10] For comparability across studies, Figure 1 displays the results for mathematics performance in the first year of the experiment when available. Overall treatment effects pooling across subjects and years are reported below. Details of the individual experiments and overall treatment effects pooling across subjects and years are reported in the Appendix.

A. Evidence on Incentive Design

In an important observation, Neal (2011) discusses how the incentive pay schemes tested thus far in the U.S. are either group incentives (e.g., Fryer 2013) or are sufficiently obtuse (e.g., Springer et al. 2011) that teachers may not respond to them. 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. Combining experimental and non-experimental studies in the U.S., Pham, Nguyen and Springer (2020) similarly highlight that the effectiveness of incentives varies by study context and design. They find larger effect sizes for incentive programs that are: based on a rank order tournament, include professional development, use multiple measures of performance, and have higher incentive awards (above 7.5% per capita). Our incentives include some of these features—they are based on a rank order tournament and have high average rewards—but do not include professional development and are not based on multiple outcome measures.

In line with Neal (2011), Pham, Nguyen and Springer (2020) find smaller effect sizes for group incentives. Similarly, Goodman and Turner (2013)—using experimental data from the New York city study discussed above—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

[8] By subject, Pham, Nguyen and Springer (2020) estimate impacts for math of: 0.050σ (0.012) in the U.S., 0.215σ (0.024) outside the U.S. and 0.067σ (0.014) in all studies. For language, they estimate effects of: 0.029σ (0.009) in the U.S., 0.173σ (0.051) outside the U.S. and 0.040σ (0.011) in all studies.

[9] We do not include Goodman and Turner (2013) in Figure 1 because they do not report standardized effect sizes.

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

are responsible for a greater share of the students that determine their reward). Ours is the first study to base rewards on teacher pairs. Related to our design, Muralidharan and Sundararaman (2011) compare individual and school-wide incentives in small schools averaging approximately three teachers each and, as discussed above, find evidence that individual rewards are more effective in the second year of the experiment.

Prior merit pay programs do not explicitly vary the framing of incentives and have paid out rewards after student performance is measured, as in our gain-framed incentives. In work related to our loss-framed incentives, Dee and Wyckoff (2015) use a regression discontinuity design to examine Washington D.C.'s IMPACT program and find evidence of improved performance among teachers at both the lower threshold for dismissal and the upper threshold for performance. Teachers at the threshold for dismissal could be considered as facing (high powered) loss-framed incentives. There is no comparable group facing equivalent gain-framed incentives in their setting.

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 (2020) 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 (2020) 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).[11] Finally, we contribute to a small but growing literature that uses randomized field experiments to test incentive pay in organizations (Shearer 2004; Bandiera, Barankay and Rasul 2007, 2013; Hossain and List 2012; Pierce, Rees-Jones and Blank 2020).

II. Program Details and Randomization

A. 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,[12] 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 C 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

[11] Brownback and Sadoff (2020) also test whether instructor incentives are more effective in combination with student incentives and find no evidence of complementarities.

[12] 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.

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 the high-stakes Illinois Standards Achievement Test (ISAT) taken by 3rd-8th graders in March.[13] In the "Team" treatments, rewards were based on the 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.

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.[14] We then ranked each student within his bin according to improvement between his baseline and end of the year test score.[15] 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 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.[16]

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. 2011; Glewwe, Ilias and Kremer 2010; Muralidharan and Sundararaman 2011; Fryer 2013). 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.[17] 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. In the team treatment groups, 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.

[13]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.

[14]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.

[15]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.

[16]Authors' calculations based on the school district's 2010 and 2011 Illinois State Report Card (2010-2011). 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.

[17]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.

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 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.

B. Random Assignment

We conducted the randomization after baseline testing was completed, which ensures that teachers cannot affect their students' pre-program test scores in response to their treatment 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 assigned to 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.[18] 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).[19] In year two, we randomly assigned all of a teacher's classes to either a treatment group or to control.

As noted above, 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

[18] 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.

[19] 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.

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.[20]

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 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.[21]

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.[22] 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.[23] 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.

III. Data and Analysis

A. 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 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

[20] 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.

[21] 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.

[22] 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 Wooldrige (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.

[23] 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.

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.[24] 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.[25] 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 year prior to the start of the experiment. This serves as baseline teacher value added for both year 1 and year 2 of the experiment. We measure teacher value as the average of students' raw percentile change on the statewide exam in math. The value-added measure is missing for teachers who were not in the district the year prior to the experiment.

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 with standard errors clustered at the teacher and student level. The table reinforces that our sample contains almost exclusively poor and minority students: 98 percent are eligible for free or reduced-price lunch, and 96 percent are members of a minority group. The only statistically significant difference across groups is in eligibility for free or reduced-price lunch. As shown in Appendix Table A.3 (column 5), excluding students from the analysis who are ineligible for free or reduced price lunch does not affect the results.

Columns (5) through (7) report descriptive statistics for year two students of the experiment. Column (8) reports the p-value from the test that the statistics in columns (5), (6), and (7) are equal. 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 in footnote 23, we did not re-randomize to achieve balance on these characteristics in year two). Column (9) reports the p-value for the test of differences across treatments pooling treatments across year 1 and year 2. The pooled means are well balanced with no statistically significant differences across treatments.

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 high-lights that there are no Year 1 Control teachers who also receive Control in Year 2. 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

[24]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.

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

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.

B. 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 scores, we exclude students who are missing baseline test scores (Appendix Table A.3, columns 1 and 2 respectively). To evaluate whether particular schools in our sample are driving the results we estimate treatment effects leaving out one school at a time and find similar estimates across specifications (Appendix Table A.4).

We present results 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).

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 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).

IV. Results

Table 4 presents estimates of the overall impact of our treatments on math ThinkLink scores and the standardized state test (ITBS for grades K-2 and ISAT for grades 3-8). Scores are normalized to have a within-grade standard deviation of one.[26] Standard errors clustered at the student and

[26]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.5. The table follows the same estimation structure as Table 4 except that we include

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 the ThinkLink in years one and two respectively. Column 3 presents estimates pooled across the two years. Columns 4 through 6 present the analogous results for the statewide test. The top row of the table pools all treatments relative to control. Subsequent rows show results disaggregated by treatment group.

We first discuss the impact of the incentive treatments on ThinkLink scores (columns 1-3). As shown in the top row of the table, overall our treatments increased test scores by 0.175σ (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 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.[27] 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.[28]

We next turn to the estimated treatment effects on the statewide test (columns 4-6).[29] While the estimates for the unincentivized tests are less precise, they largely mirror the results on the incentivized tests. The loss treatment increases statewide test scores by an estimated 0.151σ (0.084) in year one and 0.017σ (0.092) in year two. Pooling across years, we estimate that loss-framed incentives increase scores on the unincentivized test by 0.100σ (0.062), similar to the estimated impact of 0.124σ (0.056) on the incentivized test. There is little evidence that the gain treatment improves performance on the statewide test (with evidence of a negative impact in year two). Pooling across years, the estimated effect of the gain treatment is significantly different from the loss treatment at the $p = 0.02$ level. Our finding that the impact of our loss-framed incentives carries over to the unincentivized state test suggests that our main results are not driven by teachers "gaming" the incentives in ways that increase student scores on the incentivized test but do not improve student learning. In this spirit, making the measure the target did not corrupt our overall gains.

To investigate the heterogeneity of the program's effects, Table 5 presents results split by grade

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.

[27] 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.5, 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.

[28] Pooling across treatment arms and years raises concerns about incorrect inference due to multiple comparisons. We correct for multiple hypothesis testing within each regression using the method described in Anderson (2008). The estimated effects remain statistically significant for loss-framed incentives in year one ($p < 0.01$ without correction, $p = 0.011$ with correction), gain-framed incentives in year one ($p = 0.010, p = 0.083$), the difference between loss and gain treatments in year one ($p = 0.051, p = 0.054$), loss-framed incentives pooling year one and year two ($p = 0.026, p = 0.085$), individual loss-framed incentives in year one ($p = 0.051, p = 0.054$) and team loss-framed incentives in year one ($p = 0.060, p = 0.099$).

[29] 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.

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. Adjusting for multiple hypothesis testing across all of our subgroups, only the estimated impacts of the loss treatment in year one remain statistically significant.[30] In the remaining subgroups, there is little systematic heterogeneity that 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 1). Column 2 presents the year one treatment impact on the unincentivized statewide standardized tests, repeating the ITT estimates from Table 4 (column 4). In columns 3-8, we estimate the impact of a teacher's year one treatment on her students' test scores in the post-treatment year reported for each column. 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. 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).

We find suggestive evidence that the effects of the Loss treatment persist. In 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

[30] We adjust within year (i.e., column) using the methodology of Anderson (2008). For year one Loss treatments, the unadjusted and adjusted p-values respectively are: $p = 0.013$ and $p = 0.027$ for grades K-2; and $p < 0.01$ and $p = 0.025$ for grades 3-5.

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.

To address attrition concerns, we examine differential attrition in the post-treatment period (Appendix Table A.6) and find that teachers assigned to loss or gain treatments in year one are more likely to be in the post-treatment sample, but the differences are not statistically significant. We find suggestive evidence that a teacher's baseline value added (in the year prior to the intervention) is positively correlated with remaining in the long-run sample for control teachers, but baseline value added is not correlated with attrition for loss teachers. We argue this pattern of attrition should work against us finding positive long-run impacts. We then use inverse probability weighting to adjust the long-run sample to reflect attrition by teachers' year one treatment assignment and baseline value added (Appendix Table A.7). The results are similar to those uncorrected for attrition.

V. 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. And end with a discussion of implementing loss-framed incentives as a policy.

A. Mechanisms

Our findings are consistent with over 30 years of psychological and economic research on the power of loss aversion to motivate individual behavior.[31] We argue that the loss-framed incentives motivated teachers to increase effort in ways that improved student learning. We discuss possible alternative interpretations of our results below. First, rather than increased motivation, the improved performance of teachers could operate through alleviating credits constraints. Second, teachers may be more motivated, but the effectiveness of loss-famed incentives could be due to greater understanding, salience or trust in the incentives, rather than teachers having loss averse preferences. Finally, loss aversion could be driving teacher motivation, but teachers may direct their increased effort towards gaming the incentives in ways that do not improve student learning. We discuss each of these potential mechanisms in turn.

Credit Constraints. — 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. 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,

[31] 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.

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, 77% of teachers in the first year loss treatment and 50% 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).

Understanding, Trust and Salience. — Prior work has argued that in some cases when incentives have failed it has been because they are too confusing. If teachers do not understand the structure of the incentives, this could diminish the perceived return of effort and thus the responsiveness to incentives. If anything, we might think that loss-framed incentives are more confusing to teachers than gain-framed incentives which are more familiar to them. We also find little evidence of confusion among teachers. We average about 82% correct response rates on a series of knowledge questions about the maximum reward teachers can receive, how team rewards are determined and what occurs in the loss group if the total reward is less than the upfront reward.

A related concern is that teachers may not fully trust the experimenters to fulfill the agreement set out at the beginning of the school year and make the promised payments. It is plausible that paying some portion of the reward up front builds trust among the teachers that rewards will be paid out. Increasing the perceived probability of receiving the reward increases the perceived return of effort and thus the responsiveness to the incentives. 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.[32] 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.

Gaming/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. Rather than outright cheating, teachers may respond to performance pay in ways that increase student performance on the incentivized measure but do not improve student learning. For example, treated teachers might direct their efforts towards getting students to focus more or try harder on the incentivized test (relative to teachers in the control group).

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 4, during the intervention the treatment effects on the incentivized ThinkLink tests carry over to the unincentivized state tests. In addition, the results in Table 6 suggest positive impacts of the loss treatment on performance on the tests for which teachers did not receive incentives through five years of post-program follow up (2011/12 to 2015/16). It seems unlikely that the treatment effects would repeat

[32]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.

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.

Another form of cheating could be if teachers fail to return upfront payments they owe back at the end of the year, which we discuss in more detail below. We note here that if teachers anticipate they will keep their upfront payments even if they do not meet performance targets, this should diminish the perceived return to effort, which would work against finding treatment effects for loss-famed incentives.

B. 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. Prior evidence on desensitization to loss-framing is mixed. Hossain and List (2012) find that the effects of loss-framed incentives offered to Chinese factory workers are sustained over time; whereas List (2003, 2004, 2011) find that experience limits the impact of loss framing in trading markets. If experience with loss framing has a similar impact in our context, 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).[33]

Another possibility is that the change in sample composition between the first and second years of the experiment could be responsible for the differential impact. When we limit the sample to those teachers who participated in both years, the estimated treatment effects directionally increase for year 1 and directionally decrease for year 2, widening the gap between them (Appendix Table A.8). This suggests that, if anything, changes in sample composition are making the effects across years more similar, not less so.

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. There is mixed evidence from the prior literature on the effectiveness of teacher incentives over time. Pham, Nguyen and Springer (2020) find that shorter programs (lasting less than three years) have larger average effect sizes, but this may be driven by a few outliers. Studies conducted over multiple years variously find that estimated impacts over time are similar (Springer et al. 2011, 2012; Chiang et al. 2020), increase (Glewwe, Ilias and Kremer 2010; Muralidharan and Sundararaman 2011), or decrease (Marsh et al. 2011; Glazerman and Seifullah 2012; Goodman and Turner 2013). In a subsequent study to ours, Brownback and Sadoff (2020) 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

[33]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.

effects in the second semester. These findings suggest that the effects of loss framed incentives may indeed replicate both across contexts and over time.

C. Loss-framed Incentives as Policy

Finally, we discuss the extent to which our results might generalize in a policy context. First, there may be concern that our results are particular to our experimental setting. For example, teachers may be more responsive to incentives that they know are being tested for a short period (and that they will receive at some point during the experiment) compared to incentives being implemented as policy; teachers in our partner school district may be more responsive to incentives than those in other school districts; or, there may be something particular about how we structured the incentives or implemented the program. While we cannot rule out any of these concerns entirely, we do believe that our results can inform policy more generally. The estimated effect of gain-framed incentives in our study aligns well with the average estimated impact of incentives in both experimental and non-experimental studies in the U.S. (Pham, Nguyen and Springer 2020). This suggests that our setting is not an outlier. Further, our population, incentive structure, and program implementation are the same in both the loss and gain-framed treatments. Thus, the larger impact of loss-framed incentives suggests that, holding these factors constant, framing incentives as losses can improve their effectiveness.

This raises a second set of concerns about implementing loss-framed incentives as a policy. A general objection to loss-framed incentives is that they will not be attractive to teachers. Indeed, behavioral models predict that the same mechanism that makes loss-framed incentives more powerful also makes them welfare reducing—because people work harder in order to minimize the negative utility imposed by (the threat of) losses (see Imas, Sadoff and Samek (2016) for discussion and Bulte, List and van Soest (2020)). There is limited evidence on whether teachers—or employees more generally—are averse to working under loss-framed contracts.[34] In our setting, the teachers' union agreed to the structure of the contracts and about 90% of teachers opted into the program with higher participation rates in the second year. In a subsequent study to ours, Brownback and Sadoff (2020) also demonstrate high opt-in rates to a loss-framed incentive program among community college instructors. They additionally elicit instructors' contract preferences and find that at baseline instructors do prefer gain-framed contracts; but, after experiencing loss-framed incentives for a semester, instructors significantly increase their preference for them and are (close to) indifferent between contract types. Together with our study, these results suggest that loss-framed incentives can be acceptable to teachers and that, if anything, interest in them grows once they have been implemented.

A final set of concerns relates to the logistics of loss-framed rewards—in particular upfront bonuses that may need to be crawled back. In our study, we received 99% of the payments owed at the end of the year. The high repayment rate is particularly notable given that teachers had to actively write personal checks back to the University of Chicago with no direct connection or reporting to their employer, and no legal repercussions for failure to repay (though teachers did sign a contract at the beginning of the year pledging to make repayments). If implemented as a policy, repayment could be accomplished by schools more seamlessly through automatic paycheck deductions at the end of the year. This also points to a reason why loss-framing can work well in a school setting: there is a clear beginning and end of the year with a break in between for summer, so that upfront payments and end of year payments are clearly delineated.

While our results need to be replicated at a larger scale, our study shows that loss-framed incentives can be implemented as a policy and have the potential to increase the effectiveness of

[34]Prior work in laboratory and online studies find that, contrary to the theoretical predications, workers prefer loss-framed contracts Imas, Sadoff and Samek (2016); De Quidt (2018); Bulte, List and van Soest (2020).

performance pay programs. For those school districts loathe to physically taking back money from teachers, empirical results in Hossain and List (2012) suggest that merely framing the performance bonus as a loss rather than a gain can yield behavioral change in the correct direction. Indeed, their results are especially strong for workers in teams, as they find a significant improvement compared to workers in teams under gain framing.

VI. 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 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.

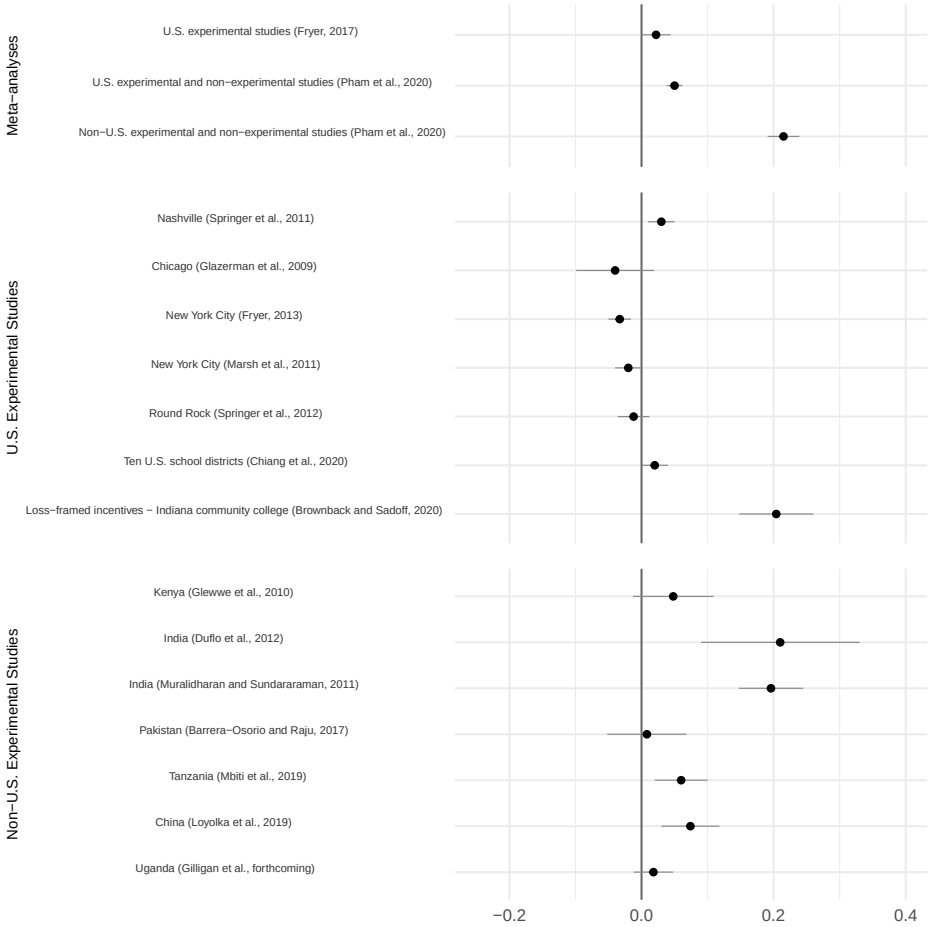
A general finding in the literature is that educational interventions tend to have larger impacts on test scores among younger students than older students (see Cascio and Staiger (2012) for discussion). This finding has also been documented for teachers, though prior work has generally not included both age groups. In their meta-analyses, Pham, Nguyen and Springer (2020) estimate larger effects of performance pay programs among elementary school students and smaller effects among middle school students. A potential reason for this finding is that, at baseline, teachers in older grades—particularly starting at third grade—already face accountability pressures for their students to perform well on tests, and so there may be less room for improvement compared to teachers in younger grades. Understanding how teacher performance pay interacts with broader accountability pressures could help policy makers design better targeted and more cost effective incentives.

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, Brownback and Sadoff (2020) provides an additional data point for 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.

VII. Figures

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



Notes: The figure presents the estimated effects of (pooled) teacher incentives on standardized math performance in the first year of the experiment, when available. Fryer (2017), Pham, Nguyen and Springer (2020), and Fryer (2013) report estimates pooled across multiple years. Brownback and Sadoff (2020) estimate effects pooling across final exams in multiple post-secondary departments. Glewwe, Ilias and Kremer (2010) and Barrera-Osorio and Raju (2017) estimate effects pooling subjects on a government exam. We report estimates pooling teacher incentive treatments for Barrera-Osorio and Raju (2017) and Loyalka et al. (2019). We do not include estimates for treatment groups that include non-teacher incentive interventions for Brownback and Sadoff (2020) and Mbiti et al. (2019). Within section, studies are listed in chronological order of implementation. Bars indicate standard errors.

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

Table 2: Summary Statistics by Treatment Arm

	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 Ed	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. Columns (1)-(3) report means for students in control, loss, and gain homerooms at the beginning of the first year. Column (4) displays a p-value from a test of equal means in the three previous columns, with standard errors clustered both at the teacher and the student level. Columns (5)-(7) report means for students in control, loss, and gain homerooms at the beginning of the second year. Column (8) displays a p-value from a test of equal means in the three previous columns, with standard errors clustered both at the teacher and the student level. 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". The number of observations are reported at the bottom of the table. 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

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
	(0.016)	(0.018)	(0.015)	(0.016)	(0.011)	(0.013)
	N = 2953		N = 2941		N = 5894	
Missing ITBS/ISAT Math Score	0.015	0.006	-0.010	-0.008	0.006	0.001
	(0.019)	(0.020)	(0.025)	(0.020)	(0.015)	(0.014)
	N = 2953		N = 2022		N = 4975	

Notes: This table presents the increase in the probability of two measures of attrition associated with our gain and loss treatments: missing a ThinkLink or a state test score. The results shown are estimated from a linear probability model where we regress the relevant dependent variable on treatment indicators, 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 are reported in parentheses and are clustered at the teacher and the student level. The number of observations is reported below the standard errors.

Table 4—The Effect of Treatment on Math Scores

	ThinkLink			State Tests		
	Year 1	Year 2	Pooled	Year 1 ITBS/ ISAT Grades K-8	Year 2 ISAT Grades 3-8	Pooled ITBS/ ISAT Grades K-8
	(1)	(2)	(3)	(4)	(5)	(6)
Any Treatment	0.175	0.017	0.099	0.107	-0.054	0.047
	(0.070)	(0.078)	(0.051)	(0.075)	(0.093)	(0.056)
Pooled Loss	0.234	0.021	0.124	0.151	0.017	0.100
	(0.080)	(0.079)	(0.056)	(0.084)	(0.092)	(0.062)
Individual Loss	0.271	0.000	0.126	0.136	0.009	0.090
	(0.087)	(0.089)	(0.060)	(0.093)	(0.110)	(0.069)
Team Loss	0.197	0.044	0.122*	0.179	0.027	0.121
	(0.106)	(0.086)	(0.067)	(0.107)	(0.101)	(0.077)
Pooled Gain	0.100	0.006	0.051	0.048	-0.179	-0.032
	(0.079)	(0.106)	(0.062)	(0.084)	(0.107)	(0.065)
Individual Gain	0.086	0.007	0.053	-0.009	-0.180	-0.079
	(0.096)	(0.106)	(0.072)	(0.093)	(0.108)	(0.070)
Team Gain	0.115		0.046	0.108		0.065
	(0.085)		(0.073)	(0.102)		(0.089)
p-value(Loss=Gain)	0.051	0.862	0.178	0.168	0.006	0.017
Observations	2630	2697	5327	2552	1896	4448
Students	2460	2543	3279	2367	1758	2747
Classrooms	135	153	288	135	106	241
Teachers	105	113	131	105	69	122

Notes: The results we report are from regressions with ThinkLink test scores (columns 1-3) or state test scores (ITBS for grades K-2 and ISAT for grades 3-8 in columns 4-6) 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). We present results both estimating years of the experiment separately and pooling across years of data. We obtain our Year 2 estimates controlling for Year 1 treatment status. 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 (we also report p-values from tests of equal coefficients between the loss and gain treatments). The coefficients we report are Intent-to-Treat estimates, i.e. students are classified based on their initial classroom assignment. Standard errors are reported in parentheses and are clustered on the classroom and the student level.

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
	(0.080)	(0.079)	(0.079)	(0.106)	(0.056)	(0.062)
Panel B: Grade Level						
K-2	0.490	0.397	0.154	0.223	0.253	0.250
	(0.185)	(0.155)	(0.129)	(0.194)	(0.117)	(0.107)
3-8	0.165	0.067	-0.043	-0.109	0.071	-0.011
	(0.061)	(0.074)	(0.101)	(0.139)	(0.059)	(0.075)
p-value	0.346	0.523	0.125	0.088	0.166	0.256
Panel C: Gender						
Male	0.200	0.163	0.113	0.066	0.155	0.119
	(0.088)	(0.084)	(0.084)	(0.105)	(0.065)	(0.063)
Female	0.294	0.071	-0.062	-0.033	0.106	0.011
	(0.090)	(0.093)	(0.092)	(0.124)	(0.063)	(0.077)
p-value	0.148	0.658	0.027	0.316	0.456	0.218
Panel D: Race						
Black	0.225	-0.100	0.033	-0.015	0.107	-0.060
	(0.104)	(0.119)	(0.126)	(0.151)	(0.086)	(0.095)
Hispanic	0.317	0.093	-0.018	0.021	0.111	0.043
	(0.102)	(0.093)	(0.083)	(0.109)	(0.063)	(0.074)
p-value	0.832	0.272	0.987	0.284	0.934	0.102
Panel E: Baseline Scores						
Above Median	0.182	0.153	0.020	0.020	0.105	0.096
	(0.072)	(0.072)	(0.091)	(0.122)	(0.060)	(0.066)
Below Median	0.132	-0.024	0.008	-0.074	0.063	-0.055
	(0.078)	(0.092)	(0.087)	(0.110)	(0.058)	(0.066)
p-value	0.171	0.686	0.429	0.433	0.989	0.335

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 testscore groups. Standard errors are reported in parentheses and are clustered on the teacher and the student level.

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

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

Notes: The results we report are from 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. The coefficients we report are Intent-to-Treat estimates, i.e. students are classified based on their initial classroom assignment. We show results pooling the impact of any Year 1 treatment, as well as gain and loss 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 type, and the grades for which test scores are available. Standard errors are reported in parentheses and are clustered on the classroom and the student level.

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.

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.

Anderson, Michael. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects." *Journal of the American statistical Association*, 103(484): 1481–1495.

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. 2007. "Incentives for Managers and Inequality among Workers: Evidence from a Firm Level Experiment." *Quarterly Journal of Economics*, 122: 729–773.

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.

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. 2020. "Improving College Instruction through Incentives." *Journal of Political Economy*, 128(8): 2925–2972.

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.

Bulte, Erwin H., John A. List, and Daan van Soest. 2020. "Toward an Understanding of the Welfare Effects of Nudges: Evidence from a Field Experiment in the Workplace." *Economic Journal*, 130(632): 2329–2353.

Cascio, Elizabeth U., and Douglas O. Staiger. 2012. "Knowledge, Tests, and Fadeout in Educational Interventions." National Bureau of Economic Research (NBER) Working Paper 18038.

Chetty, Ray, 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. 2020. "Do Educator Performance Incentives Help Students? Evidence from the Teacher Incentive Fund National Evaluation." *Journal of Labor Economics*, 38(3): 843–872.

Chicago Heights School District. 2006-2016. "Education Statistics: ThinkLink and Illinois Standard Achievement Test (ISAT)."

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. 2017. "What Motivates Effort? Evidence and Expert Forecasts." *The Review of Economic Studies*, 85(2): 1029–1069.

De Quidt, Jonathan. 2018. "Your loss is my gain: a recruitment experiment with framed incentives." *Journal of the European Economic Association*, 16(2): 522–559.

De Quidt, Jonathan, Francesco Fallucchi, Felix Kölle, Daniele Nosenzo, 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.

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, 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." *Handbook of Economic Field Experiments*, 2: 95–322.

Fryer, Roland G., Steve Levitt, John List, and Sally Sadoff. 2021. "Enhancing the Efficacy of Teacher Incentives through Framing: A Field Experiment." *AEA RCT Registry.*

Gilligan, Daniel O., Ibrahim Kasirye Naureen Karachiwalla, Adrienne M. Lucas, and Derek Neal. 2019. "Educator Incentives and Educational Triage in Rural Primary Schools." *Journal of Human Resources.*

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.*

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.*

Glazerman, Steven, and Allison Seifullah. 2012. "An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years (Final Rep.)." *Mathematica Policy Research.*

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economic*, 2(3): 205–207.

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.

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.

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.

Illinois Report Card. 2010-2011. "Chicago Heights SD 170 District Snapshot."

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

Imbens, Guido, and Jeffrey Wooldrige. 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 A., 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.

Jacob, Brian A., Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2018. "Teacher applicant hiring and teacher performance: Evidence from DC public schools." *Journal of Public Economics*, 166: 81–97.

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." National Bureau of Economic Research (NBER) Working Paper 14607.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. "The Effect of Teaching Coaching on Instruction and Achievement: A Meta-analysis of the Causal Evidence." *Review of Educational Research*, 88(4): 547–588.

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

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. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics*, 118(1): 41–71.

List, John A. 2004. "Neoclassical Theory versus Prospect Theory: Evidence from the Marketplace." *Econometrica*, 72(2): 615–625.

List, John A. 2011. "Does Market Experience Eliminate Market Anomalies? The Case of Exogenous Market Experience." *American Economic Review*, 101(3): 313–317.

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, Sean Sylvia, Changfang Liu, James Chu, and Yaojing Shi. 2019. "Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement." *Journal of Labor Economics*, 37(3): 621–662.

Marsh, Julie A., Matthew G. Springer, Daniel F. McCaffrey, Kun Yuan, Scott Epstein, Julia Koppich, Nidhi Kalra, Catherine DiMartino, and Art (Xiao) Peng. 2011. "A Big Apple for Educators: New York City's Experiment with Schoolwide Performance Bonuses." *RAND*.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania." *Quarterly Journal of Economics*, 134(3): 1627–1673.

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

Neal, Derek A. 2011. "The Design of Performance Pay in Education." *Handbook of the Economics of Education*, 4: 495–550.

Neal, Derek A. 2018. "Information, incentives, and education policy." Harvard University Press.

Pham, Lam D., Tuan D. Nguyen, and Matthew G. Springer. 2020. "Teacher Merit Pay: A Meta-Analysis." *American Educational Research Journal*.

Pierce, Lamar, Alex Rees-Jones, and Charlotte Blank. 2020. "The Negative Consequences of Loss-Famed Performance Incentives." National Bureau of Economic Research (NBER) Working Paper 26619.

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.

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 G., 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(7155): 362–363.