NBER WORKING PAPER SERIES

# ENHANCING THE EFFICACY OF TEACHER INCENTIVES THROUGH LOSS AVERSION: A FIELD EXPERIMENT

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

Working Paper 18237
http://www.nber.org/papers/w18237

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2012

We are grateful to Tom Amadio and the Chicago Heights teachers union. Matt Davis, Sean Golden, Phuong Ta and Wooju Lee provided exceptional research assistance. Financial support from the Broad Foundation (Fryer) and the Kenneth and Anne Griffin Foundation is gratefully acknowledged. The usual caveat applies. Financial support from the Broad Foundation (Fryer) and the Kenneth and Anne Griffin Foundation is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment
Roland G. Fryer, Jr, Steven D. Levitt, John List, and Sally Sadoff
NBER Working Paper No. 18237
July 2012
JEL No. J24


## ABSTRACT

Domestic attempts to use financial incentives for teachers to increase student achievement have been ineffective. In this paper, we demonstrate that exploiting the power of loss aversion—teachers are paid in advance and asked to give back the money if their students do not improve sufficiently—increases math test scores between 0.201 (0.076) and 0.398 (0.129) standard deviations. This is equivalent to increasing teacher quality by more than one standard deviation. A second treatment arm, identical to the loss aversion treatment but implemented in the standard fashion, yields smaller and statistically insignificant results. This suggests it is loss aversion, rather than other features of the design or population sampled, that leads to the stark differences between our findings and past research.



Roland G. Fryer, Jr
Department of Economics
Harvard University
Littauer Center 208
Cambridge, MA 02138
and NBER
rfryer@fas.harvard.edu

John List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and NBER
jlist@uchicago.edu

Steven D. Levitt
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
and NBER
slevitt@midway.uchicago.edu

Sally Sadoff
Rady School of Management
University of California San Diego
sadoff@uchicago.edu

## I. Introduction

Good teachers matter. A one-standard deviation improvement in teacher quality translates into annual student achievement gains of 0.15 to 0.24 standard deviations (hereafter σ) in math and 0.15σ to 0.20σ in reading (Rockoff, 2004; Rivkin et al, 2005; Aaronson et al., 2007; Kane and Staiger, 2008). These effects are comparable to reducing class size by about one-third (Krueger, 1999). Similarly, Chetty et al. (2012) 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.

It has proven difficult, however, to identify *ex ante* the most productive teachers. Observable characteristics such as college-entrance test scores, grade point averages, or major choice are not highly correlated with teacher value-added on standardized test scores (Aaronson et al., 2007; Rivkin et al., 2005; Kane and Staiger, 2008; Rockoff et al., 2008). And, programs that aim to make teachers more effective have shown little impact on teacher quality (see e.g., Boyd et al 2007 for a review). Some argue that these two facts, coupled with the inherent challenges in removing low performing teachers due to collective bargaining agreements and increased job market opportunities for women, contributes to the fact that teacher quality and aptitude has declined significantly in the past 40 years (Corcoran et al., 2004; Hoxby and Leigh, 2004).[1]

To increase teacher productivity, there is growing enthusiasm among policy makers for initiatives that tie teacher incentives to the achievement of their students. At least ten states and many more school districts have implemented teacher incentive programs in an effort to increase student achievement (Fryer *forthcoming*).[2] Yet, the

[1]Corcoran et al. (2004) find that in the 1964-1971 period, 20-25 percent of new female teachers were ranked in the top 10 percent of their high school cohort, while in 2000, less than 13 percent were ranked at the top decile. Hoxby and Leigh (2004) similarly find that the share of teachers in the highest aptitude category fell from 5 percent in 1963 to 1 percent in 2000 and the share in the lowest aptitude category rose from 16 percent to 36 percent in the same period.

[2] Merit pay faces opposition from the two major unions: The American Federation of Teachers (AFT) and the National Education Association (NEA). Though in favor of reforming teacher compensation systems, the AFT and the NEA officially object to programs that reward teachers based on student test scores and principal evaluations, while favoring instead systems that reward teachers based on additional roles and responsibilities they take within the school or certifications and qualifications they accrue. The AFT's official position cites the past underfunding of such programs, the confusing metrics by which teachers

empirical evidence on the effectiveness of such programs is ambivalent. In developing countries where the degree of teacher professionalism is extremely low and absenteeism is rampant, field experiments that link pay to teacher performance are associated with substantial improvements in student test scores (Duflo et al. (forthcoming); Glewwe et al. 2010; Muralidharan and Sundararaman (forthcoming)). On the other hand, the only two field experiments conducted in the United States have shown small, if not negative, treatment effects (Springer et al. 2010, Fryer (forthcoming)).[3]

In this paper, we implement the first field experiment on teacher incentives (and one of the only field experiments in any domain) that exploits the power of framing in the presence of loss aversion. A large literature on reference dependent preferences demonstrates behavior consistent with a notion of *loss aversion* (Kahneman and Tversky 1979, Tversky and Kahneman, 1991).[4] Lab experiments have consistently demonstrated that subjects are more responsive to protocols framed as losses than to protocols framed as gains. Until recently, however, these findings have been largely confined to the lab. Only one previous field study has tested loss aversion in the context of incentive pay, finding that bonuses framed as losses improve productivity of teams in a Chinese factory (Hossain and List 2009).[5]

During the 2010-2011 school year we conducted an experiment in nine schools in Chicago Heights, IL. At the beginning of the school year teachers were randomly selected to participate in a pay-for-performance program. Among those who were

---

were evaluated, and the crude binary reward system in which there is no gradation of merit as the reasons for its objection. The NEA's official position maintains that any alterations in compensation should be bargained at the local level, and that a singular salary scale and a strong base salary should be the standard for compensation.

[3] Non-experimental analyses of teacher incentive programs in the United States have also shown no measurable success, though one should interpret these data with caution due to the lack of credible causal estimates (Glazerman et al., 2009; Vigdor, 2008)

[4] Examples in this literature include behavioral anomalies such as the endowment effect (Thaler, 1980), status quo bias (Samuelson and Zeckhauser, 1988), and observed divergences of willingness to pay and willingness to accept measures of value (Haneman, 1991).

[5] Previous field experiments have tested the effect of the loss frame in marketing messages on product demand (Ganzach and Karsahi 1995, Bertrand et al 2010). In the context of education, two studies run concurrently to ours find that framing student incentives as losses improves short-term test performance (Levitt et al., 2012), but does not affect children's food choice (List and Savikhin, 2012). All of these studies frame incentives as losses through messaging rather than actual endowments. In this paper we test a much stronger treatment, endowing teachers for a full school year with payments worth about 8% of the average teacher's salary.

selected the timing and framing of the reward payment varied significantly. 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. Importantly, 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 (2011), in which teachers are rewarded according to how highly their students' test score improvement ranks among peers with similar baseline achievement and demographic characteristics.

The results of our experiment are consistent with over 30 years of psychological and economic research on the power of loss aversion to motivate individual behavior. Students whose teachers were in the "Loss" treatment show large and statistically significant gains between 0.201σ (0.076) and 0.398σ (0.129) standard deviations in math test scores.[6] In line with previous studies in the United States, we do not find an impact of teacher incentives that are framed as gains. The difference between the "Loss" and "Gain" treatments in math improvement is statistically significant.

We conclude our statistical analysis with three additional robustness checks to our interpretation of the data. First, we investigate the extent to which attrition out of sample may explain our results. We find little evidence suggesting that students in either treatment type are more likely to attrite. Second, analyzing a survey we administered to

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

all treatment and control teachers suggests that alternative explanations such as alleviating teacher liquidity constraints are not likely to explain the patterns in our data. If anything, teachers in the loss treatment report spending less money on classroom materials. Third, we consider whether our results could be driven by cheating, but believe this hypothesis to be unlikely since we find very similar results on non-incentivized statewide exams.

Together, this suggests that the divergence in our findings from previous results is due to the addition of framing in the incentive design – rather than differences in settings, reward structure or implementation.

The remainder of the paper is organized as follows. Section 2 provides a brief literature review. Section 3 details the experiment and its implementation. Section 4 describes the data and research design used in the analysis. Section 5 presents estimates of the impact of teacher incentives on student achievement. Section 6 describes potential threats to our interpretation. The final section concludes. There is an online data appendix that provides details on how we construct our covariates and our sample from the school district administrative files used in our analysis.

## II. A Brief Review of the Literature

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

There are a number of reasons, however, why teacher incentives may fail to operate in the desired manner. For instance, teachers may not know how to increase student achievement, the production function may have important complementarities outside their control, or the incentives may be either too confusing or too weak to induce extra effort. Moreover, if teacher incentives have unintended consequences such as explicit cheating, teaching to the test, or focusing on specific, tested objectives at the expense of more general learning, teacher incentives could have a negative impact on student performance (Holmstrom and Milgrom, 1991; Jacob and Levitt, 2003). Others argue that teacher incentives can decrease a teacher's intrinsic motivation or lead to

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

Despite the controversy, there is a growing literature on the role of teacher incentives on student performance (Glazerman et al., 2009; Glewwe et al., 2010; Lavy, 2002; Lavy, 2009; Muralidharan and Sundararaman, forthcoming; Fryer forthcoming, Springer et al., 2010; Vigdor, 2008.), including an emerging literature on the optimal design of such incentives (Neal 2011). To date, there are five papers, three of them outside the US, which provide experimental estimates of the causal impact of teacher incentives on student achievement: Duflo and Hanna (2005), Glewwe et al. (2010), Muralidharan and Sundararaman (2011), Fryer (forthcoming) and Springer et al. (2010). Figure 1 displays the one-year treatment effects from these five experiments.

*Evidence from Developing Countries*

Duflo and Hanna (2005) randomly sampled 60 schools in rural India and provided them with financial incentives to reduce absenteeism. The incentive scheme was simple; teachers' pay was linear in their attendance, at the rate of Rs 50 per day, after the first 10 days of each month. They found that teacher absence rate was significantly lower in treatment schools (22 percent) compared to control schools (42 percent) and that student achievement in treatment schools was $0.17\sigma$ (0.09) higher than in control schools.

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

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

*Evidence from Experiments in America*

Springer et al. (2010) evaluate a three-year pilot initiative on teacher incentives conducted in the Metropolitan Nashville School System between the 2006-2007 school year and the 2008-2009 school year. Approximately 300 middle school mathematics teachers who volunteered to participate in the program were randomly assigned to the treatment or the control group, and those assigned to the treatment group could earn up to $15,000 as a bonus if their students made gains in state mathematics test scores equivalent to the 95$^{th}$ percentile in the district. They were awarded $5,000 and $10,000 if their students made gains equivalent to the 80$^{th}$ and the 90$^{th}$ percentiles, respectively. Springer et al. (2010) found there was no significant treatment effect on student achievement and on measures of teachers' response such as teaching practices.[7]

Fryer (forthcoming) conducted an experiment on teacher incentives in over 200 New York City public schools, which distributed a total of roughly $75 million to over 20,000 teachers. Each participating school could earn $3,000 for every union-represented staff member, which the school could distribute at its own discretion, if the school met the annual performance target set by the Department of Education based on school report card scores. Each participating school was given $1,500 per union staff member if it met at least 75% of the target, but not the full target. Each school had the power to decide whether all of the rewards would be given to a small subset of teachers with the highest value-added, whether the winners of the rewards would be decided by lottery, or virtually anything in-between. The only restriction was that schools were not allowed to distribute rewards based on job seniority. Yet, despite this apparent flexibility, the vast majority of schools chose to distribute the rewards evenly, and there was no effect on student achievement or teacher behavior. If anything, there was a negative impact, especially in larger schools where free-riding may have been problematic.

Our contribution is three-fold. First, this paper is the first experimental study to demonstrate that teacher merit pay can have a significant impact on student performance

[7] There are several non-experimental evaluations of teacher incentive programs in the US, all of which report non-significant impact of the program on student achievement. Glazerman et al. (2009) report a non-significant effect of -0.04 standard deviations on student test scores for the Teacher Advancement Program in Chicago and Vigdor (2008) reports a non-significant effect of the ABC School-wide Bonus Program in North Carolina. Outside the US, Lavy (2002, 2009) reports significant results for teacher incentive programs in Israel.

in the U.S. Second, we show that these findings differ from previous studies as a result of incorporating insights from behavioral economics into the incentive design. Third, this is the first study to test loss aversion using high stakes endowments in the field. Moreover, we contribute to a small but growing literature that uses randomized field experiments to test incentive pay in organizations (Bandiera, Barankay, and Rasul 2007, Hossain and List 2009).

## III. Program Details

### Background

The city of Chicago Heights is located thirty miles south of Chicago. The district contains nine K-8 schools with a total of approximately 3,200 students. Like larger urban school districts, Chicago Heights is made up of primarily low-income minority students who struggle with low achievement rates—in the year prior to our program, 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.

In cooperation with the Chicago Heights superintendent and Chicago Heights Teachers Union, the implementation of our field experiment followed standard protocols. As part of our agreement with the Teachers Union to conduct an experiment with teacher incentives, program participation was made available to every K-8 classroom teacher, as well as reading and math interventionists.[8] Approximately 160 teachers were eligible to participate in the experiment. After we introduced the program at the district-wide Teacher Institute Day at the start of the 2010-11 school year, teachers had until the end of September to opt-in to the program; 150 teachers (93.75%) elected to participate.

Table 1 provides a brief summary of the treatment. Participating teachers were randomly assigned to one of four treatment groups or a control group. Teachers in the incentive groups received rewards based on their students' end of the year performance on the ThinkLink Predictive Assessment, an otherwise low stakes standardized diagnostic

[8] 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 teaches reading.

assessment that is designed to be aligned with the high-stakes Illinois Standards Achievement Test (ISAT) taken by 3rd-8th graders in March (K-2 students take the Iowa Test of Basic Skills (ITBS) in March).[9] We calculate rewards using the "pay for percentile" methodology developed by Barlevy and Neal (2011). At baseline, we place each student in a bin with his nine nearest neighbors in terms of pre-program test performance.[10] We then rank each student within his bin according to improvement between his baseline and end of the year test score. Each teacher receives an "overall percentile", which is the average of all her students' percentile ranks within their bins.[11] Teachers receive $80 per percentile for a maximum possible reward of $8,000. The expected value of the reward ($4,000) is equivalent to approximately 8% of the average teacher salary in Chicago Heights.[12]

Random Assignment

Before any randomization occurs, we paired all teachers in each school with their closest match by grade, subject(s), and students taught. From there, the randomization procedure is straightforward for "contained" teachers – i.e. those who teach the same homeroom for the entire day. These teachers are 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.

[9] The results of ISAT are used to determine whether schools are meeting yearly targets under the No Child Left Behind law. The ThinkLink is 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 lasts 30-60 minutes and is either taken on the computer (3rd-8th grade students at all but one school) or on paper (all K-2 students and 3rd-8th grade students at one school). All students are tested in math and reading. In addition, 4th and 7th grade students take a science test as they do on ISAT. We proctored all end of the year testing in order to ensure consistency and discourage cheating.

[10] We construct a predicted baseline (fall 2010) test score based on up to three years of prior test score data. For students without prior year test scores, we use their actual fall 2010 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.

[11] We rewarded teachers of contained classrooms 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 a subset of the homeroom-subjects they taught.

[12] 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 in advance.

Teachers who teach multiple homerooms are subject to a slightly different procedure. Their students are grouped by class and subject and randomized into one of the five groups. As a result, a teacher can receive incentives based on the performance of some of her classes taught throughout the day and no incentives for others. In our main specifications, the students whose performance does *not* determine a teacher's reward are included in the control group.

This aspect of our design allows us to examine the role of spillovers in our incentive scheme. If teaching production has significant fixed costs, we might expect that students in the control group whose teacher is incentivized for *other* students' performance would benefit from the treatment. For instance, if a teacher teaches both a treated and a control math class, the added incentive to prepare an effective lesson for the treated class might result in a better lesson for control students.

One difficulty with the research design is that our agreement with the Chicago Heights Teachers Union required us to offer the incentive program to all elementary and middle school teachers who signed up to participate. This presents few complications in math; students typically have only one math teacher, so there is a nearly 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.[13] Moreover, students with an Individualized Education Plan (IEP)—roughly 11% of the sample—also receive additional reading instruction from a reading specialist. Thus, more than one-third of the students in the sample have reading teachers in different treatments. Because of the confusion this likely induced among reading teachers and the difficulties that arise in the statistical analysis due to contamination and lack of power, we focus our discussion on the math results in what follows. Results for reading can be found in Appendix Tables 1 and 2.[14]

Our randomization procedure is straightforward. To improve balance among the control group and the four treatment arms, over a pure random draw, we re-randomized teachers after the initial draw. First, we calculated a balance statistic for the initial

[13] At the request of the district, we also based incentives for Language Arts and Writing teachers based on student performance on the Reading exam.

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

assignments, defined as the sum of the inverse p-values from tests of balance across all five groups.[15] 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.

There is an active discussion on which randomization procedures have the best properties. Treasure and MacRae (1998) prefer a method similar to the ones described above. Imbens and Wooldridge (2009) and Greevy et al. (2004) recommend matched pairs. Results from simulation evidence presented in Bruhn and McKenzie (2009) suggest that for large samples there is little gain from different methods of randomization over a pure single draw. For small samples, however, matched-pairs, re-randomization (the method employed here), and stratification all perform better than a pure random draw. Following the recommendation of Bruhn and McKenzie (2009), we have estimated our treatment effects including all variables to check balance. Whether we include these variables or a richer set of controls does not significantly alter the results.

Implementation

Teachers assigned to one of the two "Gain" treatments received their rewards at the end of the year, much like all previous programs have done (Springer et al. 2010; Fryer (forthcoming)). 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*.[16] 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.[17] 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.

[15] We use chi-squared tests to test for balance across categorical variables (school and subject) and rank-sum tests for continuous variables (baseline ThinkLink math score, baseline ThinkLink reading score, contact minutes with teacher, percent black, percent Hispanic, percent female and grade).
[16] For tax reasons, some teachers requested that we issue the upfront payment in 2011. About half of teachers received the loss reward at the beginning of January 2011.
[17] We received back 98% of the payments owed at the end of the year.

Within the "Gain" and "Loss" groups, teachers were also randomly assigned to receive either individual or team rewards. Teachers in the individual treatment groups received rewards based on the performance of their own students. The team treatment paired teachers in a school who were closely matched by grade and subject(s) taught. These teachers received rewards based on their average team performance. For example, if teacher A's overall percentile is 60% and teacher B's overall percentile is 40%, then their team average is 50% and each teacher receives $4,000.[18]

The study formally commenced at the end of September 2010 after baseline testing was completed. Informational meetings for each of the incentive groups were held in October 2010, at which time the incentivized compensation was explained in detail to the teachers. Midway through the experiment, we provided incentivized 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.

## IV. Data and Research Design

### Data

We merged administrative data from the Chicago Heights school district with survey data from the teachers in the experiment. The administrative data include: teacher assignments, gender, race, eligibility for free or reduced price lunch, eligibility for Special Education services, Limited English Proficiency (LEP) status, and student test scores. The test score data include: 2011 ISAT score (3rd-8th grade), 2011 ITBS score (K-2nd grade), 2010-2011 ThinkLink scores (baseline, fall, winter and spring), and up to 3 prior years of ISAT, ITBS and ThinkLink test scores (depending on the student's age and when he entered the district). The survey includes questions about program knowledge, collaboration with fellow teachers and time use. We received a 53% overall response rate

[18] Ours is the first study to base rewards on teacher pairs. Previous studies have either tested individual or school-wide rewards. Muralidharan and Sundararaman (2011) compare individual and school-wide incentives in small schools averaging approximately 3 teachers each.

(49% in the "Gain" group, 62% in "Loss" group and 36% in Control). We also worked with principals and teachers to confirm the accuracy of class rosters.

We use administrative data from 2007 through fall 2010 (pre-treatment) to construct baseline controls and spring 2011 test scores (post-treatment) for outcome measures. The main outcome variables are the direct outcomes we provide incentives for: spring ThinkLink scores in math for all K-8 students We also analyze the effect of treatment on Illinois's statewide exams taken in March (ISAT and ITBS).

Table 2 presents summary statistics for students in the "Gain" treatment, "Loss" treatment and control group.[19] Accounting for within-homeroom correlation, the groups are very well balanced. The table reinforces that our sample contains almost exclusively poor minority students; 98 percent are eligible for free or reduced-price lunch, and 96 percent are members of a minority group.

Econometric Approach

To estimate the causal impact of providing teacher incentives on outcomes, we estimate intent-to-treat (ITT) effects, i.e., differences between treatment and control group means. Let $GainTreat_h$ and $LossTreat_h$ be indicators for starting the year in a homeroom assigned to the Gain or Loss treatment, respectively. Furthermore, let $X_i$ be a vector of baseline covariates measured at the individual level that includes all the variables summarized in Table 2. The individual level covariates, $X_i$, a set of school fixed-effects, $\gamma_s$, and set of grade fixed effects, $\mu_g$, comprise our parsimonious set of controls. All these variables are measured pre-treatment, and the school and grade effects reflect a students' enrollment at the beginning of the school year.

The ITT effects, $\pi_G$ and $\pi_L$, are estimated from the equation below:

$$achievement_i = \alpha + GainTreat_h\pi_G + LossTreat_h\pi_L + X_i\beta + \gamma_s + \mu_g + \varepsilon_{isgt},$$

where $h$ indexes homerooms (the unit of randomization), $i$ students, $g$ grades, and $s$ schools.

[19] Table 2 excludes students exposed to multiple treatments. These students are included in the regression results.

Recall, given our design, it is possible that a student has two or more teachers that face different incentive treatments within the same subject area. Because we focus on math teachers, this inconvenience is easily overcome; students have at most two math teachers and 99% of the students in our sample see a single math teacher. For students with one math teacher, the treatment variable is dichotomous and we estimate a traditional intent-to-treat regression. In the 1% of cases in which a student has two math teachers who are in different treatments, we assign her a value of 0.5 for *GainTreat* and *LossTreat*. Correspondingly, if one teacher is in control and one is in treatment, the relevant treatment variable takes on a value of 0.5. Dropping all students exposed to multiple teachers yields qualitatively identical results.[20]

## V. Results

Table 3 presents ITT estimates of the effect of teacher incentives on math achievement measured by the May 2011 ThinkLink exam, the test used in determining teacher incentive compensation in our study. We present estimated treatment effects in both standard deviation units and percentile ranks within groups of nine nearest neighbors – precisely the metric on which the rewards were calculated. The top three rows correspond to the "Loss" treatments, including the pooled loss results as well as the "Individual Loss" and "Team Loss" results separately. The bottom three rows present parallel results for the "Gain" treatments. Standard errors, clustered at the homeroom level, are presented in parentheses below each estimate.

Columns (1) and (4) include only a baseline test score as a control. Columns (2) and (5) add the remaining student-level variables in Table 2 as controls; their inclusion has almost no effect on our estimates. In columns (3) and (6), we include only treated and "pure control" students in the regression sample. In these specifications, we drop students whose teacher receives rewards based on other students' performance, but not these particular students. If our treatments create positive (resp. negative) spillovers onto

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

these students, we would expect these estimates to be higher (resp. lower) than those in columns (2) and (5).

We find large impacts of the Loss treatment on ThinkLink math scores. Pooling the Team and Individual treatments together produces a treatment effect (with controls) of 0.220σ (0.065), or 6.840 (2.554) percentile points. When we drop control students at risk of receiving spillovers, our estimates are even higher: 0.338σ (0.101) and 9.559 (3.585) percentile points.

In contrast, similar to Fryer (forthcoming) and Springer (2010), we find smaller, mostly insignificant coefficients in our pooled "Gain" treatments, in which teachers are rewarded at the end of the year for student achievement. The main pooled effect is 0.092σ (0.070), or 1.884 (2.843) percentile points. When we account for possible spillovers, the aggregate Gain treatment effect climbs to 0.185 (0.097) and is marginally significant. Importantly, we can reject the two-sided null hypothesis that Gain and Loss treatments have the same treatment effects with at least 95% confidence in all specifications.

Examining the individual and team treatment arms separately, the estimated effects of Team Loss are almost identical to the effects of Individual Loss. Similarly, there is little of interest in the Individual Gain and Team Gain coefficients, except to note that they mirror the pooled results.

Table 4 presents an identical set of regressions that estimate the effect of student performance on statewide tests for which teachers did not receive any incentives. The similarities are striking: the pooled loss effect is 0.201σ (0.076) or 6.680 (3.194) percentile points. Accounting for potential spillovers, the coefficients increase to 0.398σ (0.129) and 10.838 (5.470) percentile points. The pooled gain effects are once again small and insignificant [-0.003σ (0.084) and 1.012 (3.351) percentile points], and quite similar in magnitude to the ThinkLink coefficients when we restrict the sample to pure control students [0.175σ (0.128), or 4.556 (5.198) percentile points].

Interestingly, the Team Loss estimates are consistently larger than the Individual Loss effects on the state tests. In percentile points, the team effect [10.803 (3.350)] dwarfs the individual effect [3.240 (3.917)], and we can reject the two-sided null

hypothesis of equality with 99% confidence. The two effects are not statistically differentiable when we measure outcomes in scaled scores, however.

Similar tables for Reading can be found in the Appendix; Appendix Table 1 reports results for ThinkLink, and Appendix Table 2 reports ISAT and ITBS estimates. Just as in Math, the Team Loss treatment seems to have the strongest effects on reading scores, though the effects disappear when combined with the negative Individual Loss effects in the pooled specifications. The spillover-adjusted estimates are also higher than the unadjusted estimates in all regressions. While large standard errors prevent definitive conclusions, we view this as further suggestive evidence that our treatments had beneficial effects for students in certain control classrooms.

To investigate the heterogeneity of the program's effects, Table 5A presents results split out across genders, races, and baseline test performance. Because of the large number of coefficients presented, we show results only for the percentile improvement or decline on ThinkLink.[21] Column 1 presents results for the full sample, repeating the estimates shown in Table 3. The next two columns break down the results by gender. The following two columns divide the sample by race, and the last two columns according to whether a student's baseline test score was above or below the median baseline score in his grade. There is little systematic heterogeneity that we are able to detect. In no case is the comparison of treatment effects statistically different across the sub-groups.

An interesting pattern emerges when we separate students by grade level however. As Table 5B demonstrates, each treatment has large and statistically significant effects among students in second grade and below – ranging from 14.487 (6.403) percentile points in Team Gain to 17.437 (6.126) in Individual Loss. Among 3$^{rd}$-8$^{th}$ graders, the effects are more muted, and only the loss treatment effects are differentiable from zero. Pooling the loss treatments yields a treatment effect of 5.791 (2.555) percentile points among older students, as opposed to -0.740 (3.266) in pooled gain.

[21] Full results are available from the authors on request.

## VI. Interpretation

The results presented thus far suggest that loss aversion can play an important role in amplifying the power of teacher incentives. When we offer teachers performance rewards paid at the end of the year, we replicate the null findings in Springer et al. (2010) and Fryer (forthcoming). However, when we provide an advance on the payment and re-frame the incentive as avoidance of a loss, we observe treatment effects in excess of $0.201\sigma$ and some as high as $0.398\sigma$—similar to the increase in student achievement associated with a more than one standard deviation increase in teacher quality. In all of our specifications, we can reject the null hypothesis of equal effects in the Loss and Gain treatments with 95% confidence.

In this section we identify and respond to three potential threats to our interpretation: the influence of attrition, the use of cash-in-advance to change the education production function, and the potential for cheating.

### Attrition

It is possible that teachers in the loss treatment resorted to means other than teaching to increase their rewards. One might worry that they found ways to discourage their weaker students from taking exams, for instance. Our incentive structure complicates this strategy to some extent, at least, as rewards are based on a student performance relative to other students with similar baseline scores and demographic profiles. However, teachers could potentially weed out students who have made relatively less progress during the year.

To investigate the prevalence of attrition in our sample, we run a probit regression on all of our covariates where the dependent variable is an indicator for missing any of the exams we consider. Table 6 reports the marginal effects of our Gain and Loss treatments, evaluated at the mean.

There is little evidence that attrition is a threat to our results. Students in the Gain treatments are 2.1 (1.3) percentage points less likely than the control group to be missing ThinkLink Math scores, but the difference is not statistically significant. Similarly, they are more likely to be missing ISAT/ITBS scores by roughly the same magnitude – 2.8 (2.0) percentage points in Math. The Loss treatment demonstrates smaller and

insignificant results. This is important because if a student misses a state exam, a score of zero is imputed when calculating the school-wide average. Given that we see almost identical treatment effects across ThinkLink and the state exams for which our teachers were not directly incentivized, it seems unlikely that attrition is driving our results.

Liquidity Constraints

Suppose that teachers are able to purchase certain materials that facilitate instruction such as new workbooks or dry-erase markers. Incentivizing student achievement creates an additional incentive to make these investments, provided that the gains (and the associated rewards) are large enough to justify the expense. Under perfect credit markets, we would not expect any difference in effects between our gain and loss treatments. Teachers in the loss group could use their upfront check if necessary, and cash-strapped teachers in the gain treatments could borrow money to finance their purchases.

If teachers are liquidity-constrained, however, the loss treatment effectively gives them access to a form of financing unavailable to teachers in the gain treatment. It is possible that this mechanism could create the effects that we observe.

Survey evidence, however, suggests that this explanation is unlikely. Table 7 reports treatment effect estimates on several survey outcomes. When asked how much of their personal money they spent on classroom materials that year, teachers in the loss group reported spending less than teachers in the gain treatment or the control group (though the difference is not statistically significant). What's more, 69% of teachers in the loss treatment report that they had not spent any money from their check when they were surveyed in March 2011 (three quarters of the way through the experiment).

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.

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. It seems unlikely that the pattern would repeat itself on the ITBS and ISAT tests if differential cheating practices across treatment and control groups were driving our results.

## VII. Conclusion

In this study, we 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 more than one standard deviation. Our findings have implications not only within education, but more broadly. While there is overwhelming laboratory evidence for loss aversion, there have been few prior field experimental demonstrations of this phenomenon.

Our results, along with those of Hossain and List (2009) suggest that there may be significant potential for exploiting loss aversion in the pursuit of both optimal public policy and the pursuit of profits.

References

Aaronson, D., Barrow, L. & Sander, W. (2007). "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics*, 25(1): 95-135.

Bandiera, Oriana & Barankay, Iwan & Rasul, Imran, 2012. "Team Incentives: Evidence from a Firm-Level Experiment," IZA Discussion Papers 6279, Institute for the Study of Labor (IZA).

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

Barlevy and Neal. 2011. "Pay for Percentile." NBER Working Paper No. 17194.

Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir and Jonathan Zinman. 2010. "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment," *Quarterly Journal of Economics* 125(1), pp. 263-305.

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.

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

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2012). "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," NBER Working Paper 17699.

Corcoran, Sean P., William N. Evans, and Robert M. Schwab. 2004. "Changing Labor-Market Opportunities for Women and the Quality of Teachers, 1957-2000." *American Economic Review*, 94(2): 230-235

Duflo, Esther, Rema Hanna, and Stephen P. Ryan (forthcoming). "Incentives Work: Getting Teachers to Come to School." *American Economic Review*.

Fryer, Roland G. (2011). "Teacher Incentives and Student Achievement: Evidence form New York City Public Schools," *Journal of Labor Economics*, forthcoming.

Ganzach, Yoav, and Nili Karsahi, "Message Framing and Buying Behavior: A Field Experiment," *Journal of Business Research*, 32 (1995), 11–17.

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

Mathematica Policy Research, Inc.

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

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

Hanemann, W. Michael. 1991. "Willigness To Pay and Willingness To Accept: How Much Can They Differ?" *American Economic Review*, 81(3): 635-647.

Hoxby, Caroline M. and Andrew Leigh. 2004. "Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States." *American Economic Review*, 94(2):236-240.

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.

Hossain, Tanjim and John A. List. 2009. "The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations." NBER Working Paper 15623.

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

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.

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

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

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

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

Levitt, Steven D., John A. List, Susanne Neckermann and Sally Sadoff. 2012. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." NBER Working Paper No. 18165.

List, John A. and Anya C. Savikhin. 2012. “The Behavioralist as Dietician: Leveraging Behavioral Economics to Improve Child Food Choice and Consumption.” Working Paper.

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

Neal, Derek. 2011. “The Design of Performance Pay Systems in Education.” NBER Working Paper No. 16710.

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

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

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

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane and Douglas O. Staiger, 2008. “Can You Recognize an Effective Teacher when You Recruit One?” NBER Working Paper no. 14485

Samuelson, William and Richard Zeckhauser. 1988. “Status Quo Bias in Decision Making.” *Journal of Risk and Uncertainty*, 1(1): 7-59.

Springer, Matthew G., Dale Ballou, Laura S. Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. “Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching.” Conference paper, National Center on Performance Incentives.

Thaler, Richard. 1980. “Toward a Positive Theory of Consumer Choice.” *Journal of Economic Behavior & Organization*, 1(1): 39-60.

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

Tversky, Amos and Daniel Kahneman. 1991. “Loss Aversion in Riskless Choice: A Reference-Dependent Model.” *The Quarterly Journal of Economics*, 106(4): 1039-1061.

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

**Figure 1: Effect of Teacher Performance Pay Programs on Student Achievement**

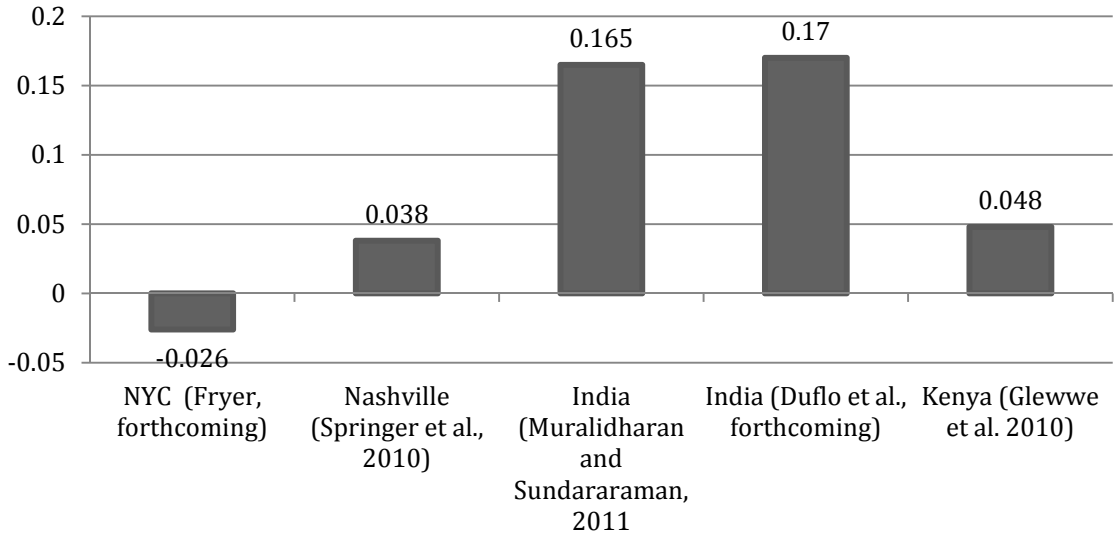

Table 1: Summary of Teacher Incentive Program

*Panel A: Overview*

| | |
|---|---|
| Schools | Nine K-8 schools in Chicago Heights, IL |
| Operations | $632,960 distributed in incentive payments, 90% consent rate. |

*Panel B: Outcomes of Interest*

| | Subjects and Grades | Date of Assessment |
|---|---|---|
| ThinkLink Learning Diagnostic Assessment | Math (K-8th), Reading (K-8th), and Science (4th and 7th) | May 2011 |
| Illinios Standards Achievement Test | Math (3rd-8th), Reading (3rd-8th), and Science (4th and 7th) | March 2011 |
| Iowa Test of Basic Skills | Math (K-2nd) and Reading (K-2nd) | March 2011 |

*Panel C: Treatment Details*

| | Timing of Rewards | Basis For Rewards |
|---|---|---|
| Individual Loss | Teachers Receive $4,000 check in September 2010; must pay back difference in June 2011 | Teacher's own students |
| Individual Gain | Teachers paid in full in June 2011 | Teacher's own students |
| Team Loss | Teachers Receive $4,000 check in September 2010; must pay back difference in June 2011 | Teacher's and teammate's students |
| Team Gain | Teachers paid in full in June 2011 | Teacher's and teammate's students |

| | |
|---|---|
| *All Treatments* | Treated teachers earned between $0 and $8,000 in bonus payments 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. |

Table 2: Summary Statistics by Treatment Arm

| | Control | Gain | Loss | *p-val* |
|---|---|---|---|---|
| Female | 0.501 | 0.489 | 0.484 | 0.791 |
| White | 0.043 | 0.048 | 0.044 | 0.957 |
| Black | 0.357 | 0.359 | 0.365 | 0.978 |
| Hispanic | 0.598 | 0.573 | 0.576 | 0.953 |
| Free or Reduced Price Lunch | 0.988 | 0.981 | 0.972 | 0.172 |
| Limited English Proficiency | 0.168 | 0.151 | 0.121 | 0.866 |
| Special Education | 0.121 | 0.096 | 0.121 | 0.440 |
| Baseline Thinklink Math Score | -0.069 | 0.059 | 0.037 | 0.451 |
| Baseline Thinklink Reading Score | -0.040 | 0.040 | 0.021 | 0.722 |
| Observations | 656 | 981 | 982 | |
| Homerooms | 38 | 48 | 59 | |

Notes: This table presents summary statistics for the control group and our two treatment arms. Thinklink scores are standardized to have mean zero and standard deviation one. Column (4) displays a p-value from a test of equal means in the three groups, with standard errors clustered at the homeroom level.

Table 3: The Effect of Treatment on Thinklink Math Scores (ITT)

| | *Percentile Rank* | | | *Scaled Score* | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Pooled Loss | 6.866** | 6.840*** | 9.559*** | 0.222*** | 0.220*** | 0.338*** |
| | (2.677) | (2.554) | (3.585) | (0.070) | (0.065) | (0.101) |
| Individual Loss | 5.827* | 6.284** | 9.128** | 0.212** | 0.220*** | 0.346*** |
| | (3.090) | (3.002) | (4.114) | (0.081) | (0.078) | (0.117) |
| Team Loss | 8.201*** | 7.471** | 10.011*** | 0.238*** | 0.221*** | 0.335*** |
| | (3.111) | (2.889) | (3.690) | (0.084) | (0.076) | (0.102) |
| Pooled Gain | 1.263 | 1.884 | 3.901 | 0.078 | 0.092 | 0.185* |
| | (2.888) | (2.843) | (3.508) | (0.072) | (0.070) | (0.097) |
| Individual Gain | 0.942 | 1.939 | 3.706 | 0.066 | 0.092 | 0.178* |
| | (3.299) | (3.293) | (3.889) | (0.083) | (0.081) | (0.106) |
| Team Gain | 1.777 | 1.846 | 4.070 | 0.095 | 0.093 | 0.196* |
| | (3.440) | (3.343) | (3.891) | (0.085) | (0.082) | (0.104) |
| Controls? | N | Y | Y | N | Y | Y |
| Spillovers Removed? | N | N | Y | N | N | Y |
| Pr(Gain = Loss) | 0.019 | 0.031 | 0.015 | 0.019 | 0.027 | 0.008 |
| Observations | 2311 | 2311 | 1986 | 2311 | 2311 | 1986 |
| Clusters | 141 | 141 | 125 | 141 | 141 | 125 |

Notes: This table presents estimates of the effectiveness of various teacher incentive structures on Thinklink math scores. Results are shown for two outcome measures: a student's percentile rank on an end-of-year test relative to her nine nearest neighbors, and her score on that test normalized to have mean zero and standard deviation one within each grade. All regressions include dummy variables for each students' school and grade as well as baseline test scores. Columns (2), (3), (5), and (6) add controls for students' race, gender, age, free-lunch status, English proficiency, and special education status. To minimize spillover effects, columns (3) and (6) and drop students whose teachers were rewarded based on the performance of other students in the sample. We also present a p-value on the null hypothesis of equal treatment effects in pooled loss and pooled gain. Standard errors are clustered at the homeroom level throughout. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 4: The Effect of Treatment on ISAT and ITBS Math Scores (ITT)

| | *Percentile Rank* | | | *Scaled Score* | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Pooled Loss | 6.867** | 6.680** | 10.838** | 0.213*** | 0.201*** | 0.398*** |
| | (3.269) | (3.194) | (5.470) | (0.078) | (0.076) | (0.129) |
| Individual Loss | 2.915 | 3.240 | 7.333 | 0.157* | 0.149 | 0.357** |
| | (3.918) | (3.917) | (6.311) | (0.090) | (0.092) | (0.146) |
| Team Loss | 11.686*** | 10.803*** | 14.025** | 0.292*** | 0.271*** | 0.444*** |
| | (3.472) | (3.350) | (5.564) | (0.092) | (0.085) | (0.133) |
| Pooled Gain | 0.228 | 1.012 | 4.556 | -0.013 | -0.003 | 0.175 |
| | (3.402) | (3.351) | (5.198) | (0.085) | (0.084) | (0.128) |
| Individual Gain | 0.205 | 0.991 | 3.883 | -0.052 | -0.036 | 0.137 |
| | (3.818) | (3.770) | (5.346) | (0.098) | (0.096) | (0.134) |
| Team Gain | 0.666 | 1.207 | 4.569 | 0.042 | 0.039 | 0.211 |
| | (4.283) | (4.251) | (6.007) | (0.115) | (0.113) | (0.154) |
| Controls? | N | Y | Y | N | Y | Y |
| Spillovers Removed? | N | N | Y | N | N | Y |
| Pr(Gain = Loss) | 0.017 | 0.042 | 0.025 | 0.003 | 0.006 | 0.004 |
| Observations | 2144 | 2144 | 1828 | 2144 | 2144 | 1828 |
| Clusters | 140 | 140 | 124 | 140 | 140 | 124 |

Notes: This table presents estimates of the effectiveness of various teacher incentive structures on ISAT/ITBS math scores. Results are shown for two outcome measures: a student's percentile rank on an end-of-year test relative to her nine nearest neighbors, and her score on that test normalized to have mean zero and standard deviation one within each grade. All regressions include dummy variables for each students' school and grade as well as baseline test scores. Columns (2), (3), (5), and (6) add controls for students' race, gender, age, free-lunch status, English proficiency, and special education status. To minimize spillover effects, columns (3) and (6) and drop students whose teachers were rewarded based on the performance of other students in the sample. We also present a p-value on the null hypothesis of equal treatment effects in pooled loss and pooled gain. Standard errors are clustered at the homeroom level throughout. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 5A: The Impact of Treatment on Math Percentile Scores Within Demographic Subgroups

| | *Whole* | *Gender* | | | *Race* | | | *Baseline Score* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Sample* | Male | Female | p-val | Black | Hispanic | p-val | Above Median | Below Median | p-val |
| Pooled Loss | 6.840*** | 5.942** | 8.681*** | | 10.181** | 9.531*** | | 7.421** | 5.718* | |
| | (2.554) | (2.893) | (3.230) | 0.404 | (4.029) | (3.183) | 0.892 | (3.484) | (3.346) | 0.704 |
| Individual Loss | 6.284** | 6.624* | 7.516** | | 9.388** | 10.354*** | | 4.914 | 7.761** | |
| | (3.002) | (3.388) | (3.696) | 0.805 | (4.167) | (3.783) | 0.841 | (4.066) | (3.652) | 0.565 |
| Team Loss | 7.471** | 5.424 | 10.001*** | | 11.027** | 8.894** | | 11.338*** | 3.403 | |
| | (2.889) | (3.481) | (3.573) | 0.250 | (5.337) | (3.683) | 0.727 | (3.636) | (3.873) | 0.097 |
| Pooled Gain | 1.884 | 2.993 | 2.028 | | 0.449 | 5.498 | | 2.567 | 0.256 | |
| | (2.843) | (2.836) | (3.640) | 0.759 | (3.397) | (3.333) | 0.241 | (3.920) | (3.343) | 0.618 |
| Individual Gain | 1.939 | 2.306 | 2.662 | | 0.274 | 4.440 | | 1.363 | 1.914 | |
| | (3.293) | (3.274) | (4.231) | 0.917 | (4.179) | (3.996) | 0.410 | (4.499) | (3.927) | 0.916 |
| Team Gain | 1.846 | 3.892 | 1.372 | | 0.504 | 6.588 | | 4.247 | -1.714 | |
| | (3.343) | (3.578) | (4.008) | 0.494 | (4.278) | (3.971) | 0.180 | (4.215) | (4.213) | 0.256 |
| Observations | 2311 | 1155 | 1122 | | 584 | 941 | | 1225 | 1078 | |
| Clusters | 141 | 141 | 135 | | 95 | 102 | | 134 | 139 | |

Notes: This table reports treatment effects for various subgroups in the data. The outcome variable is the student's percentile rank within a group of her nine nearest neighbors. Columns (5), (8), and (11) report p-values resulting from a test of equal coefficients between the gender, race, and baseline test score groups, respectively. Standard errors (clustered at the homeroom level) are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 5B: The Impact of Treatment on Math
Percentile Scores Within Grade Level Subgroups

| | *Whole* | *Grade Level* | | |
|---|---|---|---|---|
| | *Sample* | K-2 | 3-8 | p-val |
| Pooled Loss | 6.840*** | 17.745*** | 5.791** | |
| | (2.554) | (6.126) | (2.555) | 0.065 |
| Individual Loss | 6.284** | 17.437** | 6.603** | |
| | (3.002) | (6.922) | (2.871) | 0.137 |
| Team Loss | 7.471** | 17.328** | 5.690* | |
| | (2.889) | (6.931) | (3.157) | 0.118 |
| Pooled Gain | 1.884 | 15.515*** | -0.740 | |
| | (2.843) | (5.359) | (3.266) | 0.008 |
| Individual Gain | 1.939 | 15.859*** | -3.286 | |
| | (3.293) | (5.247) | (4.244) | 0.004 |
| Team Gain | 1.846 | 14.487** | 2.544 | |
| | (3.343) | (6.403) | (3.538) | 0.093 |
| Observations | 2311 | 638 | 1673 | |
| Clusters | 141 | 45 | 97 | |

Notes: This table reports treatment effects for grade-level subgroups in the data. The outcome variable is the student's percentile rank within a group of her nine nearest neighbors. The rightmost column reports the p-values resulting from a test of equal coefficients between the two groups. Standard errors (clustered at the homeroom level) are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 6: Attrition

| | Gain | Loss |
|---|---|---|
| Missing Thinklink Math Score | -0.021 | 0.006 |
| | (0.013) | (0.017) |
| | N = 2511 | |
| Missing Thinklink Reading Score | -0.024 | -0.001 |
| | (0.014) | (0.013) |
| | N = 2419 | |
| Missing ISAT Math Score | 0.028 | 0.001 |
| | (0.020) | (0.022) |
| | N = 2511 | |
| Missing ISAT Reading Score | 0.027 | 0.010 |
| | (0.023) | (0.025) |
| | N = 2419 | |

Notes: This table presents the increase in the probability of several measures of attrition associated with our gain and loss treatments. The results shown are the marginal effects calculated from a probit regression of the relevant dependent on treatment indicators and our full list of control variables. Standard errors (in parentheses) are clustered at the homeroom level. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Table 7: Teacher Survey Results

| Outcome | Gain | Loss |
|---|---|---|
| Hours Lesson Planning | 0.026 | -0.461 |
| | (1.256) | (1.258) |
| Hours Grading | -0.328 | -1.647 |
| | (1.438) | (1.440) |
| Hours Calling or Meeting w/ Parents | 0.138 | -0.071 |
| | (0.458) | (0.459) |
| Hours Tutoring Outside of Class | 1.092 | 0.953 |
| | (1.742) | (1.744) |
| Hours Leading Extracurricular Activities | 0.585 | 0.555 |
| | (1.330) | (1.332) |
| Hours Completing Administrative Work | -1.587* | -1.258 |
| | (0.936) | (0.938) |
| Hours Completing Professional Development Coursework | 0.464 | 0.084 |
| | (1.392) | (1.394) |
| Personal Money Spent on Class Materials ($) | 19.744 | -77.895 |
| | (101.722) | (101.874) |

Notes: This table presents results gathered from surveys of teachers in our experimental group at the end of the school year. All coefficients are derived by regression the outcome variable in the first column on a dummy for participation in the Gain or Loss treatment. The sample size for each regression is 82 teachers. Teachers are considered to have participated in either type of treatment if they receive that type of incentive based on any of their students' performance. Standard errors are reported in parentheses. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 1: The Effect of Treatment on Thinklink Reading Scores (ITT)

| | *Percentile Rank* | | | *Scaled Score* | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Pooled Loss | -0.006 | 0.266 | 4.649 | 0.006 | 0.018 | 0.179 |
| | (2.778) | (2.761) | (4.093) | (0.067) | (0.064) | (0.115) |
| Individual Loss | -8.060** | -7.875** | -4.047 | -0.189** | -0.174* | -0.028 |
| | (3.739) | (3.796) | (4.268) | (0.095) | (0.090) | (0.118) |
| Team Loss | 5.640** | 5.945** | 13.637*** | 0.139** | 0.150** | 0.388*** |
| | (2.795) | (2.814) | (4.450) | (0.064) | (0.065) | (0.125) |
| Pooled Gain | -0.095 | -0.082 | 6.061 | -0.030 | -0.036 | 0.159 |
| | (2.649) | (2.581) | (3.693) | (0.063) | (0.060) | (0.104) |
| Individual Gain | 2.708 | 2.656 | 10.647*** | 0.054 | 0.044 | 0.285** |
| | (3.381) | (3.360) | (4.031) | (0.082) | (0.080) | (0.116) |
| Team Gain | -5.451* | -5.395* | 1.585 | -0.182** | -0.182** | 0.027 |
| | (3.095) | (3.051) | (4.106) | (0.080) | (0.078) | (0.112) |
| Controls? | N | Y | Y | N | Y | Y |
| Spillovers Removed? | N | N | Y | N | N | Y |
| Pr(Gain = Loss) | 0.974 | 0.899 | 0.604 | 0.632 | 0.464 | 0.779 |
| Observations | 2242 | 2242 | 1848 | 2242 | 2242 | 1848 |
| Clusters | 140 | 140 | 119 | 140 | 140 | 119 |

Notes: This table presents estimates of the effectiveness of various teacher incentive structures on Thinklink reading scores. Results are shown for two outcome measures: a student's percentile rank on an end-of-year test relative to her nine nearest neighbors, and her score on that test normalized to have mean zero and standard deviation one within each grade. All regressions include dummy variables for each students' school and grade as well as baseline test scores. Columns (2), (3), (5), and (6) add controls for students' race, gender, age, free-lunch status, English proficiency, and special education status. To minimize spillover effects, columns (3) and (6) and drop students whose teachers were rewarded based on the performance of other students in the sample. We also present a p-value on the null hypothesis of equal treatment effects in pooled loss and pooled gain. Standard errors are clustered at the homeroom level throughout. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

Appendix Table 2: The Effect of Treatment on ISAT/ITBS Reading Scores (ITT)

| | *Percentile Rank* | | | *Scaled Score* | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Pooled Loss | -4.102 | -3.848 | -0.222 | 0.008 | 0.026 | 0.079 |
| | (3.285) | (3.194) | (4.937) | (0.085) | (0.075) | (0.115) |
| Individual Loss | -8.844** | -8.088* | -4.878 | -0.096 | -0.074 | -0.039 |
| | (4.358) | (4.310) | (5.474) | (0.111) | (0.099) | (0.128) |
| Team Loss | -0.792 | -0.924 | 4.813 | 0.078 | 0.093 | 0.206 |
| | (3.935) | (3.837) | (5.223) | (0.107) | (0.096) | (0.125) |
| Pooled Gain | -2.218 | -2.526 | 1.272 | -0.004 | -0.022 | 0.033 |
| | (2.869) | (2.816) | (4.210) | (0.076) | (0.070) | (0.106) |
| Individual Gain | -0.872 | -1.039 | 4.522 | 0.043 | 0.027 | 0.122 |
| | (3.352) | (3.243) | (4.139) | (0.086) | (0.076) | (0.100) |
| Team Gain | -5.167 | -5.586 | -1.988 | -0.095 | -0.115 | -0.059 |
| | (3.470) | (3.548) | (4.655) | (0.100) | (0.095) | (0.123) |
| Controls? | N | Y | Y | N | Y | Y |
| Spillovers Removed? | N | N | Y | N | N | Y |
| Pr(Gain = Loss) | 0.556 | 0.675 | 0.637 | 0.886 | 0.541 | 0.579 |
| Observations | 2347 | 2347 | 1919 | 2347 | 2347 | 1919 |
| Clusters | 144 | 144 | 123 | 144 | 144 | 123 |

Notes: This table presents estimates of the effectiveness of various teacher incentive structures on ISAT and ITBS reading scores. Results are shown for two outcome measures: a student's percentile rank on an end-of-year test relative to her nine nearest neighbors, and her score on that test normalized to have mean zero and standard deviation one within each grade. All regressions include dummy variables for each students' school and grade as well as baseline test scores. Columns (2), (3), (5), and (6) add controls for students' race, gender, age, free-lunch status, English proficiency, and special education status. To minimize spillover effects, columns (3) and (6) and drop students whose teachers were rewarded based on the performance of other students in the sample. We also present a p-value on the null hypothesis of equal treatment effects in pooled loss and pooled gain. Standard errors are clustered at the homeroom level throughout. *, **, and *** denote significance at the 90%, 95%, and 99% confidence levels, respectively.

ONLINE APPENDIX – NOT FOR PUBLICATION

## A. Variable Construction

*Attrition*
For our experiment, we agreed to base teacher rewards on students' performance relative to other students of similar ability. Our experimental sample therefore includes all students for whom we observe enough baseline information to compute a "predicted" score at the time of treatment assignment. In other words, for a student to enter our experimental sample, we must observe either (a) a baseline ThinkLink score from Fall 2010 or (b) a full set of test scores from the pre-treatment year. Students in this group for whom we do not observe a final test score are considered attriters.

Students who enter the district in the middle of the year are not included in our sample.

*Free and Reduced Price Lunch*
We construct an indicator variable set to one if a student meets guidelines for free or reduced price lunch, as determined by District administrative files.

*Limited English Proficiency*
We set the LEP indicator to one if the District considers a student to be a "Limited English Speaker" or a "Non English Speaker" and zero if the student is a "Fluent English Speaker" or "Native English Speaker."

*Race/Ethnicity*
We code the race variables such that the five categories – white, black, Hispanic, Asian, and other – are complete and mutually exclusive. Hispanic ethnicity is an absorbing state. Hence "white" implies non-Hispanic white, "black" non-Hispanic black, and so on.

*Special Education*
We construct an indicator variable for students who receive special instruction according to an Individualized Education Plan, as recorded by the District.

*Test Scores*
For grades K-8, we observe scaled Math and Reading scores on four ThinkLink examinations taken during the 2010-11 school year (baseline, fall, winter, and spring). Teachers' rewards were determined by their student's performance on the final test, and it is our main outcome variable.

Students also take the Iowa Test of Basic Skills (grades K-2) and the Illinois Standard Achievement Test (grades 3-8). We observe math and reading scores for all grades. Where available, we use scores from the Spring 2010 exams as control variables.

To simplify interpretation, we normalize all scores to have mean zero and standard deviation one by grade, subject, and test. When scores are used as control variables, we leave them un-standardized and use the scale score interacted with the student's grade level.

*Treatment Variables*

To compute ITT effects, all treatment designations are based on the teacher(s) to which a student is initially assigned. When a student has multiple teachers incentivized on a given test, we assign treatment values according to the dosage assumption described in the text.