Does Performance Pay Reduce Teachers' Intrinsic Motivation?

Evidence from the New York City Teacher Bonus Program

Jesse Margolis

City University of New York Graduate Center

February 16, 2015

Abstract: Analyses of the recent New York City teacher bonus program found lower student test scores at schools randomly chosen to receive performance pay than at control schools not eligible for bonuses. Several studies have suggested lower intrinsic motivation among teachers at treatment schools as a possible mechanism driving these surprising results. The program was abruptly ended after three years, allowing me to use post-trial data to study the program's persistent effect on intrinsic motivation absent any effect on extrinsic motivation. While I replicate prior results showing that the teacher bonus program had a negative impact on student test scores, a difference-in-difference analysis of teacher survey responses indicates that this was not likely caused by a change in teachers' intrinsic motivation. Moreover, a regression discontinuity (RD) study demonstrates that the observed "negative" effect on test scores was not driven by a decline in performance at treatment schools, but rather by an increase in performance at control schools. This finding highlights a risk with RCT experiments in the social sciences: even with proper randomization, the control group may not be a valid counterfactual for the treatment group.

Note: I would like to thank Michael Grossman, David Jaeger, Ted Joyce, Steve O'Connell and seminar participants at the CUNY Graduate Center for helpful comments on an earlier version of this research.

1. Introduction

Merit pay, or pay-for-performance, is common in many industries (Lemeiux, MacLeod and Parent, 2009). Salespeople are often paid on commission, waiters receive a substantial portion of their pay in tips, and bankers receive year-end bonuses based, in part, on their supervisors' assessment of their performance. By contrast, public school teachers in the U.S. are usually paid according to a uniform, district-wide pay scale that takes into account years of experience, level of education, and little else. Over the last decade, many schools and districts have started to experiment with merit pay, often paying teachers bonuses based on their students' test scores. In 2009, the U.S. Department of Education said it would judge schools applying for the Federal Race to the Top grants, in part, on the degree to which they used teacher and principal evaluations "to provide opportunities for highly effective teachers and principals...to obtain additional compensation...".[1] When the Round 1 winners – Tennessee and Delaware – were announced on March 29, 2010, both states had teacher bonus programs as part of their proposals.

Internationally, evaluations of teacher bonus programs have found largely positive impacts on student test scores and other outcomes (see Lavy, 2009 in Israel; Glewwe, Ilias, and Kremer, 2010 in Kenya; and Muralidharan and Sundararaman, 2011, in India). However, within the United States, the two large teacher bonus programs that have been rigorously evaluated have shown no positive impact on student test scores or other outcomes. In Nashville, the Project on Incentives in Teaching (POINT) program provided middle school math teachers with the potential to win individual bonuses of up to $15,000 annually based on improvement in their students' math test scores. After three years, students in randomly selected treatment classrooms – whose teachers were eligible for bonuses – showed no improvement in test scores over students in control classrooms (Springer et al., 2010). In New York City, the School-Wide

[1] Federal Register, November 18, 2009, pp. 59836-59872

Performance Bonus program – hereafter referred to as the NYC teacher bonus program – provided teachers and other educators at high-needs schools with the opportunity to win a roughly $3,000 annual bonus based on their school's performance on the Progress Report (an annual evaluation of test scores and other outcomes). After three years, students at schools randomly selected to be in the treatment group showed no better – and in some cases, worse – outcomes on standardized tests and other measures (Marsh et. al., 2011; Fryer, 2013; Goodman and Turner, 2013).

These results run counter to classical economic theory, which posits a strong link between performance-based-pay and improved performance. In traditional economic models, both greater consumption and greater leisure are assumed to increase utility. If one's consumption does not increase with greater work effort – since pay is unrelated to effort – then one will work as little as possible to maximize leisure. Researchers in social psychology, however, have theorized that people are intrinsically motivated to undertake many activities and providing performance-based pay for an intrinsically motivating activity can be counterproductive. Performance-based pay can decrease intrinsic motivation and potentially performance, a finding that has been repeatedly demonstrated in laboratory settings.

In a pioneering study, college students were provided with the opportunity to work on a series of challenging puzzles, and half of the students were paid based on the number of puzzles they solved. Those who were paid proved to be less motivated to work on these puzzles at a later point, during their free time, than students who were unpaid (Deci, 1971). Edward Deci, who conducted the study, said it "supported the hypothesis that if monetary rewards are given to subjects for doing an intrinsically motivated activity, and if the rewards are made contingent on their performance, their intrinsic motivation for the activity will decrease" (Deci, 1975, p. 132).

In a related study, Uri Gneezy and Aldo Rustichini find that compared to an unpaid control group, college students do worse on an IQ test if paid a small sum of money per right answer, though better if paid a large sum (2000a).

One field study showing behavior consistent with a loss of intrinsic motivation is sometimes called the Israeli Daycare Study (Gneezy and Rustichini, 2000b). In an effort to encourage parents to pick up their children on time, ten daycare centers in Israel began levying a small fine on those who were late. After levying the fine, the proportion of parents arriving late to pick up their children increased, prompting the daycares to remove the fine. However, even when the fine was removed, the proportion of parents arriving late stayed at the new higher level.

The design of the New York City teacher bonus program offers an opportunity to test for this phenomenon – financial incentives reducing intrinsic motivation – in a rigorous, real-world setting. Starting in the 2007/08 school year – hereafter referred to as 2008 – the NYC teacher bonus program was run as a three-year randomized controlled trial, with 212 treatment schools and 184 control schools (of which 175 treatment schools and 127 control schools – the focus of this study – had elementary or middle school testing data in all years between 2006 and 2013). The entire group of 396 schools were not, themselves, a random sub-sample New York City's 1,200+ schools. Rather, they represented the 396 highest-need schools as defined by their Progress Report peer index, a continuous index based on student demographic characteristics and prior test scores.[2]

At treatment schools, all educators affiliated with the local teacher's union (the United

[2] The peer index is intended to be a measure of fixed student characteristics at a school. At elementary and K-8 schools, the peer index is based on the percentage of students who require special education services, are Black or Hispanic, are English Language Learners, or qualify for free or reduced price lunch. At the middle school level, the peer index is based on the average Math and ELA test scores students earned in 4th grade, prior to entering the middle school. At the high school level, the peer index is based on the average Math and ELA test scores students earned in 8th grade, prior to entering the high school.

Federation of Teachers) were eligible to receive bonuses averaging $3,000 per person if their school met its target on the annual New York City Progress Report.[3] At the K-8 level, a school's score on the Progress Report was determined largely by students' test scores and growth in test scores (85%), attendance (5%) and the results of parent, teacher, and student surveys (10%). At the high school level, a portion of the weight was removed from test score measures and placed on the school's graduation rate and students' credit accumulation.

Educators at 62% of treatment schools won a bonus during the first year of the program (2008), a rate that rose to 88% during the second year (2009), and 13% during the third year (2010). Over the three years of the program, New York City paid out nearly $56 million in bonuses (Marsh et al. 2011). On January 20, 2011, the New York City Department of Education (NYCDOE) announced it was suspending the teacher bonus program, based on uncertain benefits and budget constraints.[4] In July, 2011, after the RAND Corporation released its final report, the NYCDOE permanently ended the teacher bonus program[5] (Marsh et al., 2011). After the conclusion of the three-year pilot, 2011 was the first school year in which teachers at treatment schools no longer had the possibility of receiving a performance bonus.

In explaining the neutral to negative results of the New York City bonus program, researchers have pointed to a decrease in intrinsic motivation as one possible cause. Referring to their finding of a negative and statistically significant effect of the bonus program on math test scores at large schools, Goodman and Turner (2013) say that "One explanation is that the bonus program crowded out teachers' intrinsic motivation…" Fryer (2013), who finds a negative effect on the bonus program on middle school math and language arts test scores, says that "…some

[3] A committee of two teachers and two principal appointees determined the exact distribution of the bonus among all eligible recipients. In 2008, 52% of staff who received an award received exactly $3,000 (Marsh et. al., 2011)
[4] http://www.nytimes.com/2011/01/21/nyregion/21bonuses.html.
[5] http://www.nytimes.com/2011/07/18/education/18rand.html

argue that teacher incentives can decrease a teacher's intrinsic motivation…" And Marsh et al (2011), who conducted extensive interviews with New York City teachers as part of the program's official evaluation, report that "In many schools, staff members also typically attributed their hard work and their efforts to improve their practices to intrinsic motivations far above any external pressures or incentives."

In my study, I seek to answer the following question: did the New York City teacher bonus program lower teachers' intrinsic motivation? I do this by extending prior research in three ways. First, I use the RCT assignment to assess the impact of the bonus program on student test scores in the three years following the suspension of the program. To the extent the bonus has a *persistent* negative effect on intrinsic motivation – as prior literature indicates it might – studying the post-suspension period allows me to disentangle this effect from any positive effect on extrinsic motivation during the program period itself. Second, I conduct an item-level analysis of the NYC teacher survey using all questions asked consistently before, during, and after the teacher bonus program. Using a difference-in-differences (DD) methodology, I assess the impact of the bonus program on teacher responses to questions related to intrinsic motivation and compare this to the impact of the bonus program on questions unrelated to intrinsic motivation.[6] Finally, I complement the RCT with a regression discontinuity (RD) study of student test scores, taking advantage of the fact that the entire experimental sample – both treatment and control schools –represent the highest-need one-third of schools in New York, as defined by a continuous index with a rigid cut point. This allows me to use non-eligible schools as a counterfactual for *both* the treatment and control schools and answer the following question: do

[6] The canonical DD study has one first difference over time, whereas the two differences I note here are between the treatment and control group and between questions related and unrelated to intrinsic motivation. I address the time dimension in my study in two ways. First, I calculate a lagged dependent variable model where I control for the difference in survey scores in 2007, before the bonus program was introduced. Second, I calculate a triple difference model, where the third difference is over time. The results, which are shown in an appendix, are similar.

we observe a negative effect of the bonus program because the treatment schools went down relative to their (unobserved) counterfactual or, rather, because the control schools improved?

In analyzing the post-program period, I find that the negative effects of the bonus program on student test scores continued and grew after the program was suspended. The bonus program caused treatment schools to perform between 0.13 and 0.17 standard deviations worse than control schools in math and between 0.08 and 0.13 standard deviations worse in language arts in the three years following the program's suspension. While this result, taken alone, would be consistent with a decrease in intrinsic motivation among teachers at treatment schools, the next two analyses provide evidence against such an interpretation. In analyzing the teacher survey, I again show that the bonus program had an effect: teacher survey scores at treatment schools were lower during the post-program period. However, the trend is the same for questions related and unrelated to intrinsic motivation, a pattern one wouldn't expect if changes in intrinsic motivation were driving this result. Finally, in the RD analysis – where I use non-eligible schools as a counterfactual for both the treatment and control schools – I replicate the main results from the RCT analysis: students at treatment schools had lower test scores than students at control schools, especially in the period following the suspension of the bonus program. However, this does not appear to be caused by a decline in performance at the treatment schools, but rather by an improvement at control schools. This finding is inconsistent with a decline in intrinsic motivation among teachers at treatment schools and suggests another mechanism may be driving the NYC teacher bonus program results observed in this and other studies. Overall, I find little evidence to suggest that intrinsic motivation declined at schools participating in the New York City teacher bonus program.

2. Methodology & Data

In this paper, I analyze the impact of the NYC teacher bonus program on teachers' intrinsic motivation in three ways. First, I take advantage of the fact that the program was designed as a Randomized Controlled Trial (RCT) to assess its impact on student test scores. I compare average school-wide student test scores in math and language arts in treatment schools to those in control schools. Second, I use a difference-in-differences methodology to study the impact of the program on teacher attitudes. Making use of the RCT, I assess the impact of the bonus program on teacher responses to survey questions related to intrinsic motivation and compare that to the impact of the bonus program on teacher survey questions unrelated to intrinsic motivation. Third, I use a Regression Discontinuity (RD) design to better understand the mechanism behind the observed RCT results. In particular, I analyze whether the test score results observed in the RCT are due to a decline in performance at the treatment schools or an increase in performance at control schools.

In each analysis, I study the impact of the bonus program over eight years, which are divided into three distinct periods in Table 1. In the pre-program period, comprised of the years 2006 and 2007, I expect no impact of the bonus program on test scores or teacher attitudes because the program had not yet been announced. Analyzing these years serves to validate the randomization and provide a falsification test. During the three-year program period (2008 to 2010), even if the program reduced teachers' intrinsic motivation, the predicted impact on both student test scores and teacher survey results is indeterminate. For student test scores, any negative impact of the program due to teachers' lower intrinsic motivation might have been offset by a positive impact of the reward on teachers' extrinsic motivation. For teacher survey results, the tendency to report lower levels of intrinsic motivation might have been offset by the desire to actually win a bonus: teacher survey results counted as 3-5% of a school's Progress Report score on which the bonus was based. Once the bonus program was suspended in 2011, however, the extrinsic reward was removed, leaving only any persistent change in intrinsic motivation. As indicated by the results of the Israeli Daycare Study, once a person's intrinsic motivation has been lowered by an external reward/penalty, it may not quickly return to its previous higher level. If true in the case of the New York City teacher bonus program, we would be most likely to observe results consistent with a decrease in teachers' intrinsic motivation *after* the bonus was removed.

Table 1 – Expected effect of the NYC teacher bonus program on test scores and survey results

Period	Years	Expected Effect	Rationale
Pre-program (placobo)	2006 2007	None	Program not yet announced
Program	2008 2009 2010	Indeterminate	Positive effect of extrinsic reward countered by negative effect of reduced intrinsic motivation
Post-program	2011 2012 2013	Negative	Loss of intrinsic motivation only (to the extent it persists)

Note: each year listed refers to the end of the school year (i.e. 2013 represents the 2012/13 school year)

2.1 Student Test Scores (Randomized Controlled Trial)

In line with prior research on the New York City teacher bonus program, I focus first on student test scores in math and ELA in grades 3 through 8. In addition to connecting with prior literature, student test scores are a useful starting point for two reasons. First, a large body of research has established that teachers have a strong influence on student test scores, so it is plausible to think that a change in teachers' intrinsic motivation would be reflected in their students' test scores (Rockoff, 2004, Rivkin, Hanushek, and Kain, 2005, Kane and Staiger, 2008). Second, for many policy makers, a change in teachers' intrinsic motivation is important only if it affects student outcomes (of which test scores are one measure). Assuming policy makers are primarily concerned with improving student outcomes, and only secondarily in the welfare of the education workforce, a change in teacher attitudes that affects student performance will be of paramount importance.

In New York State, all third through eighth graders who are not severely disabled take standardized tests in the second half of the year. For the years 2006 through 2013, the NYCDOE web site provides information on average test scores and number of students tested by grade by school. I create a single average test score variable for each school in each year, weighting by the number of students in each grade. To create the average test score variable, I normalize the scale scores within each grade to have mean zero and standard deviation one – based on the citywide distribution of average school-level test scores – before combining into a single school-level average. In the results below, I restrict my sample to those schools that have ELA test score data for all years in the study period: 2006 to 2013. This excludes two schools that opened during this period and ten schools that closed during this period. My results are robust to including these schools.

To assess the impact of the NYC teacher bonus program on student test scores, I fit the simple linear model shown in Equation 1 using Ordinary Least Squares (OLS).

$$Test_{st} = \beta_0 + \beta_1 D_s + \beta_2 Test_{s,t=2007} + \beta_3 X_{st} + \epsilon_{st} \qquad (1)$$

In Equation 1, $Test_{st}$ represents the average school-wide test score at school s in year t, D_s is a dummy variable equal to one if school s was invited to be in the treatment group and zero if school s was placed in the control group, $Test_{s,t=2007}$ is the pre-treatment average test score, and X_{st} represents a vector of school-level control variables. Assuming random assignment to the treatment group, β_1 represents the causal effect on student test scores of being invited to participate in the teacher bonus program. Since not all schools invited to participate chose to do so, β_1 should be viewed as an Intention to Treat (ITT) estimate.[7] Neither $Test_{s,t=2007}$ nor X_{st} are necessary to identify the causal impact of the bonus program on student test scores. They are included only to reduce residual variance and allow for a more precise impact estimate. In the base specification, I restrict both β_2 and β_3 to be zero.

2.2 Teacher Survey Responses (Difference-In-Differences)

To assess the impact of the teacher bonus program on a more direct measure of intrinsic motivation, I use teacher responses to the New York City School Survey. Every year, in March and April, all parents, all teachers, and students in grade 6 and above are invited to fill out the NYC School Survey. When the surveys were first collected in April, 2007, the response rate was 26% for parents, 44% for teachers, and 65% for students. The surveys collected in April, 2013, had a response rate of 54% for parents and 83% for teachers and students. On the teacher survey, many questions have been asked consistently for most or all of the time period under study. From 2007 to 2012, for example, teachers were asked for their agreement with the statement: “Teachers in my school recognize and respect colleagues who are the most effective teachers.” Teachers could respond that they Strongly Agreed, Agreed, Disagreed, or Strongly

[7] To be part of the program, 55% of UFT represented staff had to vote to participate. According to Marsh et. al. (2011), 87% of schools invited to be part of the bonus program elected to do so.

Disagreed to this statement. For each question in each school, the NYCDOE assigned a score from 0 to 10 along a Likert-like scale, with 0 corresponding to Strongly Disagree and 10 corresponding to Strongly Agree (or the reverse where Strongly Agree would be a negative response). Collections of questions were grouped into four categories – academic expectations, communication, engagement, and safety – and average scores for each category were calculated. For each category, the average score across all respondent groups – teachers, students, and parents – appears on the Progress Report and counts for 10% of a school's letter grade. Detailed item-level data by school are made available on the NYCDOE's web site.

I focus on the 45 questions that were asked consistently between 2007 and 2012. This allows me to control for a pre-treatment year (2007) in some specifications and observe teacher responses over time through the end of the teacher bonus program. I exclude 2013 since the survey was redesigned that year and the majority of questions changed. I identify questions related to intrinsic motivation in two ways. First, I review the questions and subjectively divide them into nine questions that I consider to be most closely related to intrinsic motivation and 36 remaining questions (see Table A1 in the appendix). Second, I select the one question that was most closely related to the items on widely-used *Intrinsic Motivation Inventory* (Deci & Ryan, 2014): "I usually look forward to each working day at my school." Since this question was only asked in 2012 and 2013, I develop a broader group of intrinsic motivation-related questions through an exploratory factor analysis of survey results in 2012. Following Deci & Ryan (2014), I identify three categories of questions by grouping together items that have a factor loading of at least 0.6 on one category and less than 0.4 on the others. I consider the category that contains the "look forward" question as the one most closely related to intrinsic motivation. The other two categories appear to contain questions that are broadly related to teacher collaboration and school safety. To estimate the effect of the bonus program on teachers' survey responses, I fit the equation shown below using Ordinary Least Squares:

$$z_{st} = \beta_0 + \beta_1 D_s + \beta_2 z_{s,t=2007} + \epsilon_{st} \qquad (2)$$

where

$$z_{st} = x_{st} - y_{st} \qquad (3)$$

Here x_{st} is the average teacher score on the intrinsic motivation questions for school s in year t, y_{st} is the average score on non-intrinsic motivation questions, and z_{st} is the difference between the two. With β_2 set to zero, β_1 provides the straightforward difference-in-differences (DD) estimate where the first difference is between the two types of survey questions and the second difference is between the treatment and control group. Unlike the canonical DD model, neither difference is related to time, though I add a time dimension in two ways. First, I estimate the coefficient β_2 on a lagged dependent variable, $z_{s,t=2007}$. Second, I recalculate the dependent variable as the difference between the current year difference and the 2007 difference, thus estimating a difference-in-difference-in-difference (DDD) model.

2.3 Regression Discontinuity Design

To validate and extend the RCT results, I analyze the student test score data using a Regression Discontinuity (RD) Design, taking advantage of the rigid discontinuity in school eligibility for the bonus program. Prior to randomization, 430 high-need schools were selected to be eligible for the bonus based on their peer index on the NYC Progress Report.[8] These *bonus-eligible* schools – a subset of all 1,217 NYC schools that received a Progress Report in 2007 – were selected entirely based on their Progress Report peer index. Within each of five school types – elementary, K-8, middle, high, and transfer – schools above a particular peer index score were chosen to be eligible for the teacher bonus program and schools below that score were ineligible.[9] Conceptually, an RD design compares bonus-eligible schools just above the peer index cut point to bonus-ineligible schools just below the cut point. Assuming that bonus-ineligible schools just below the cut point present a valid counterfactual for bonus-eligible schools just above the cut point, once the peer index (i.e. forcing variable) is controlled for, any difference in outcomes reflects the impact of being eligible for the bonus program.

I use several specifications to implement the RD design based on the equation below:

$$Test_{st} = \beta_0 + \beta_1 E_s + \beta_2 f\left(PeerIndex_{s,t=2007}\right) + \beta_3 E_s \times f\left(PeerIndex_{s,t=2007}\right) + \epsilon_{st} \quad (4)$$

In Equation 4, $Test_{st}$ is the average math or ELA test score for school s in year t, E_s is an indicator for whether school s was among original 430 bonus eligible schools with one indicating eligible and zero indicating ineligible, and $\left(PeerIndex_{s,t=2007}\right)$ is a polynomial function of the school's peer index in 2007, the forcing variable by which school eligibility for the bonus program was determined. The polynomial function is interacted with E_s to allow it to be fit separately on either side of the eligibility cut point.

[8] Note that my study focuses on the 302 elementary, middle, and K-8 schools that had consistent testing data from 2007 to 2013 and were not barred from participation by the UFT (discussed below).

[9] For middle schools, high schools, and transfer schools, a lower peer index meant a higher-need school. Technically, therefore, schools *below* a certain peer index threshold were eligible for the bonus program. In this analysis, I normalize all peer indices by calculating z-scores within school type and then defining a higher z-score to be a higher-need schools (multiplying by -1 where needed).

In keeping with recommendations in the RD literature, I estimate various versions of this equation (Lee and Lemieux, 2010). I allow $f\left(PeerIndex_{s,t=2007}\right)$ to vary from a first through fourth order polynomial function and I restrict the sample to bandwidths increasingly close to the cut point. I also run a local linear regression using a triangular kernel (Nichols, 2011). In the tables presented in the main paper, I focus on the results using all of the data where $f\left(PeerIndex_{s,t=2007}\right)$ is a linear function for two reasons. First, as shown in Figure 3, there appears to be a strong linear relationship between peer index and test scores. Second, the linear regression performs best on "placebo" tests using data from 2006 and 2007, prior to the implementation of the bonus program. Results from the local linear regression – which are very similar to those from a full sample linear regression and perform only slightly worse on placebo tests – are presented in an appendix.

3. Descriptive Results

3.1 Confirmation of Randomization

Table 1 shows basic demographic and performance characteristics for the treatment and control schools during the period prior to the bonus program. Overall, we can see that the schools in the study have a high proportion of Black or Hispanic students (96%) and a large share of students who qualify for free or reduced price lunch. The average test scores for these schools are substantially below the New York City mean, with z-scores of between -0.75 and -0.78, depending on the subject and the group. This lower-than-average performance in the pre-treatment period is consistent with the eligibility requirement for the bonus program: schools had to have among the highest peer index scores (i.e. highest need) on the 2007 Progress Report.

Table 2 – Pre-Bonus Program Characteristics at Treatment and Control Schools

Variable	Treatment	Control	Difference	p-Value
% Elementary Schools (2008)	0.64	0.61	0.03	0.59
% Middle Schools (2008)	0.24	0.25	-0.02	0.76
% K-8 Schools (2008)	0.13	0.14	-0.02	0.70
% English Language Learner (2008)	0.22	0.22	0.00	0.85
% Special Education (2008)	0.18	0.19	0.00	0.51
% Free or Reduced Price Lunch (2008)	0.87	0.86	0.01	0.55
% Black or Hispanic (2008)	0.96	0.96	0.00	0.82
Progress Report Peer Index (2007, z-score)	0.93	0.95	-0.02	0.42
Progress Report Overall Score (2008)	53.8	53.6	0.2	0.93
Math Test Score (2007, z-score)	-0.75	-0.76	0.00	0.95
Change in Math Score (2006 --> 2007)	0.03	0.01	0.02	0.57
ELA Test Score (2007, z-score)	-0.78	-0.76	-0.02	0.71
Change in ELA Score (2006 --> 2007)	-0.03	0.00	-0.03	0.43
Enrollment (2008)	585	574	11	0.70
Count	175	127		

Note: the table shows the mean values for the treatment and control schools. The difference is the coefficient of a regression of each variable on a dummy variable for the treatment group. The regression is weighted by enrollment for each of the first nine variables. The regression is weighted by the number of test takers for the following four variables, which represent test score outcomes. Variables listed as 2008 are based on an enrollment snapshot from 10/31/07, approximately concurrent with the start of the bonus program.

When comparing the 175 treatment schools to the 127 control schools, we see that the two groups are balanced along each dimension. The p-value of a test for the equality of means is far from traditional levels of statistical significance for each variable, giving confidence in the randomization.[10] This is consistent with prior research into the New York City teacher bonus program, which finds evidence consistent with proper randomization (Fryer, 2013; Goodman and Turner, 2013; Marsh et. al., 2011).

3.2 Student Test Scores

Figure 1 shows the distribution of school-wide average math and ELA test scores in the

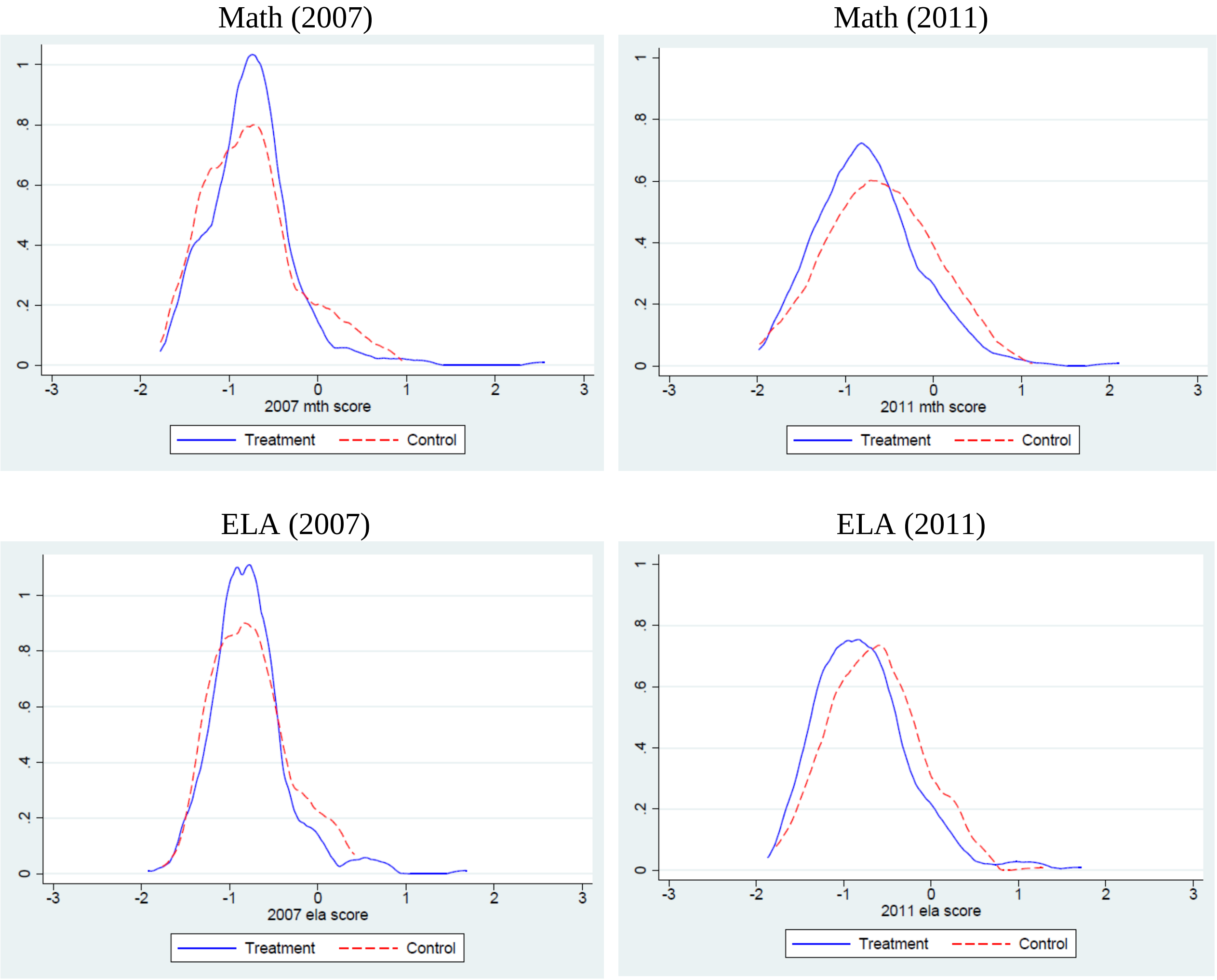
[10] The randomization of schools into the treatment and control groups was conducted by Roland Fryer, who at the time was serving as the Chief Equality Officer of the NYCDOE.

treatment and control groups. As before, the test scores are measured in z-scores, standardized based on the NYC-wide mean and standard deviation within each year/grade/subject combination. The left side of the chart shows the distribution in 2007, prior to the implementation of the bonus program. In both subjects, the treatment and control groups appear to be fairly similar to one another. Both distributions are centered below zero, consistent with the selection criteria for the program.

The right side of the figure shows the distributions in 2011, the first year after the bonus program was suspended. We can see that both the treatment and control distributions have widened when compared to 2007. In particular, the distributions have moved to the right somewhat, with a larger number of schools having above average test scores (z-scores > 0). Given that schools were selected to be eligible for the study based on their low incoming achievement levels and high-need demographic characteristics from 2007, it makes sense that by 2011, the schools would be less similar to one another (within each group). Since both groups of schools were, by definition, among the lowest achieving schools in 2007, mean reversion would tend to lead to an improvement in their standing over time relative to all NYC schools.

We can also see that by 2011, the treatment group distribution is below the control group distribution. This is true for both subjects and appears to be true across the majority of the distribution (i.e. the shape of the two graphs are fairly similar). While clearly visible, the difference between the two distributions is small, representing what appears to be a fraction of standard deviation in test scores.

Figure 1 – Test Scores Distributions Before (2007) and After (2011) the Bonus Program


Math (2007)

Math (2011)

ELA (2007)

ELA (2011)

3.3 Teacher Surveys

In addition to validating randomization, when analyzing the teacher survey, we must also address the possibility of response bias. While it is administered as a census, teachers are not required to take the survey and not all do. Since it was first administered in 2007, teacher response rates have increased from 44% to 83%. To test whether the bonus program itself had an influence on teacher response rates I run a simple regression along the lines of Equation 2, where z_{st} represents each school s's response rate in year t. The results are shown in Table 3.

Table 3 – Teacher Response Rates to the NYC School Survey

	Pre-Pgm.	Program			Post-Program		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
D_i Coefficient	2007	2008	2009	2010	2011	2012	2013
Teacher, No Controls	-0.029	-0.014	0.022	0.038	0.010	0.005	0.008
	(0.020)	(0.026)	(0.025)	(0.024)	(0.021)	(0.023)	(0.020)
Teacher, Control for 2007		-0.002	0.036	0.047**	0.022	0.013	0.017
		(0.026)	(0.024)	(0.024)	(0.021)	(0.023)	(0.020)
N	299	302	302	302	302	299	302

Note: each coefficient is the result of a separate regression with school-level teacher survey response rate as the dependent variable and treatment (D_i) as the independent variable. Regressions with controls adjust for the survey response rate in 2007. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

In Table 3, the coefficients show the percentage point difference in response rates between the treatment and control groups. When looking at the regression without controls in the first row, none of the differences are statistically significant. We can see, however, that the point estimate for the treatment effect in 2007 – prior to the start of the bonus program – is negative, while the point estimate for 2010 is positive. In the second row, I control for 2007 survey scores in each regression, which leads the 2010 coefficient to become a positive and statistically significant 4.7 percentage points. Overall, there is some evidence that teacher response rates increased at bonus eligible schools *during* the bonus program – perhaps unsurprising since teacher surveys played a small role in determining whether a school won a bonus – but no evidence of a lasting effect. In the three years since the bonus program ended, the difference between teacher response rates at treatment and control schools has been indistinguishable from zero.

3.4 Regression Discontinuity Design

A necessary condition for the validity of an RD design is that the forcing variable influences assignment to treatment. Figure 2 shows this to be the case for the NYC teacher bonus program. The x-axis shows the distance from the peer index cut point in units of standard deviation, where schools are grouped into bins of width 0.1. The y-axis shows the percentage of schools in each bin that were invited to participate in the bonus program. No schools below the eligibility cut point were invited to participate in the bonus program and slightly more than 50% of schools above the eligibility cut point were invited to participate.

Figure 2 – Probability of Treatment by Distance from Peer Index Cut Point

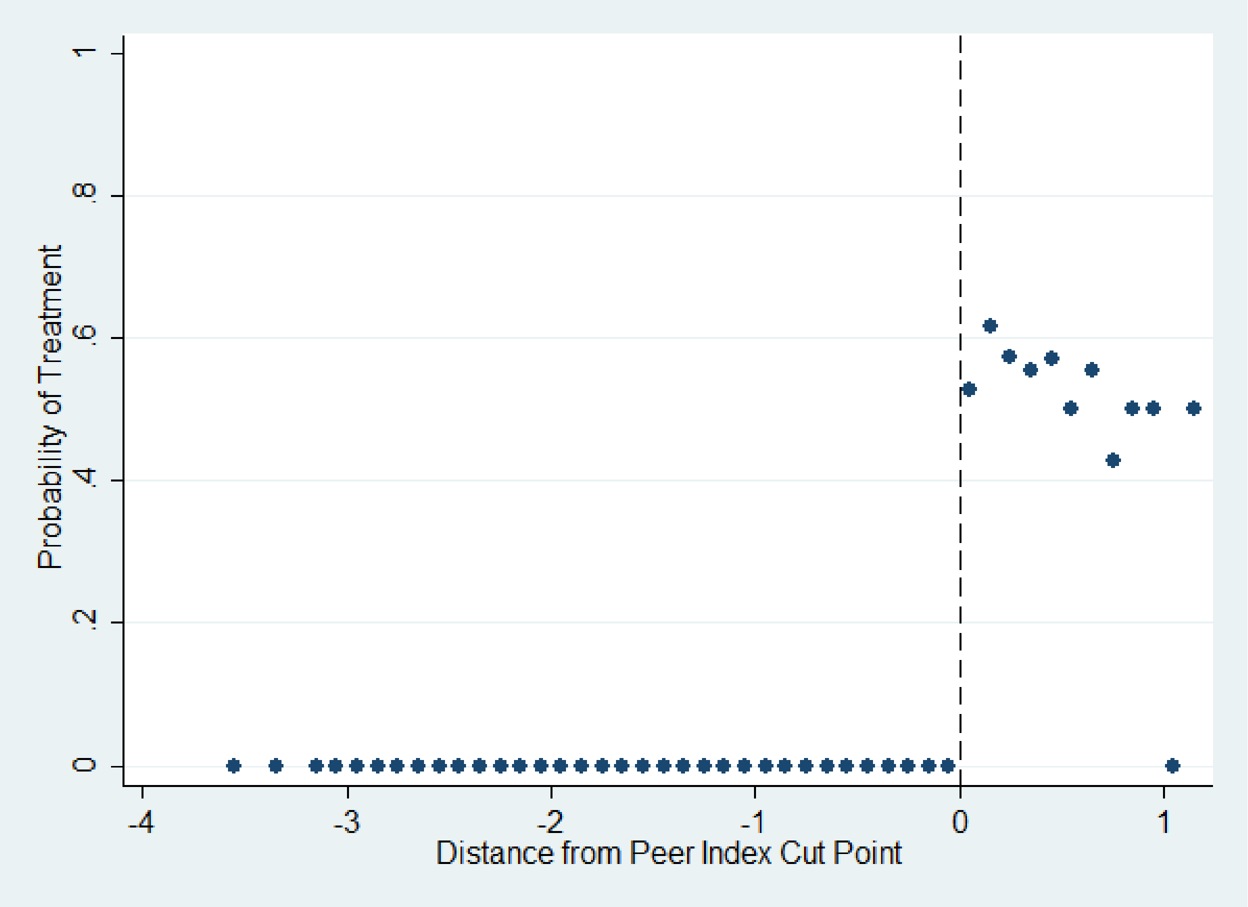
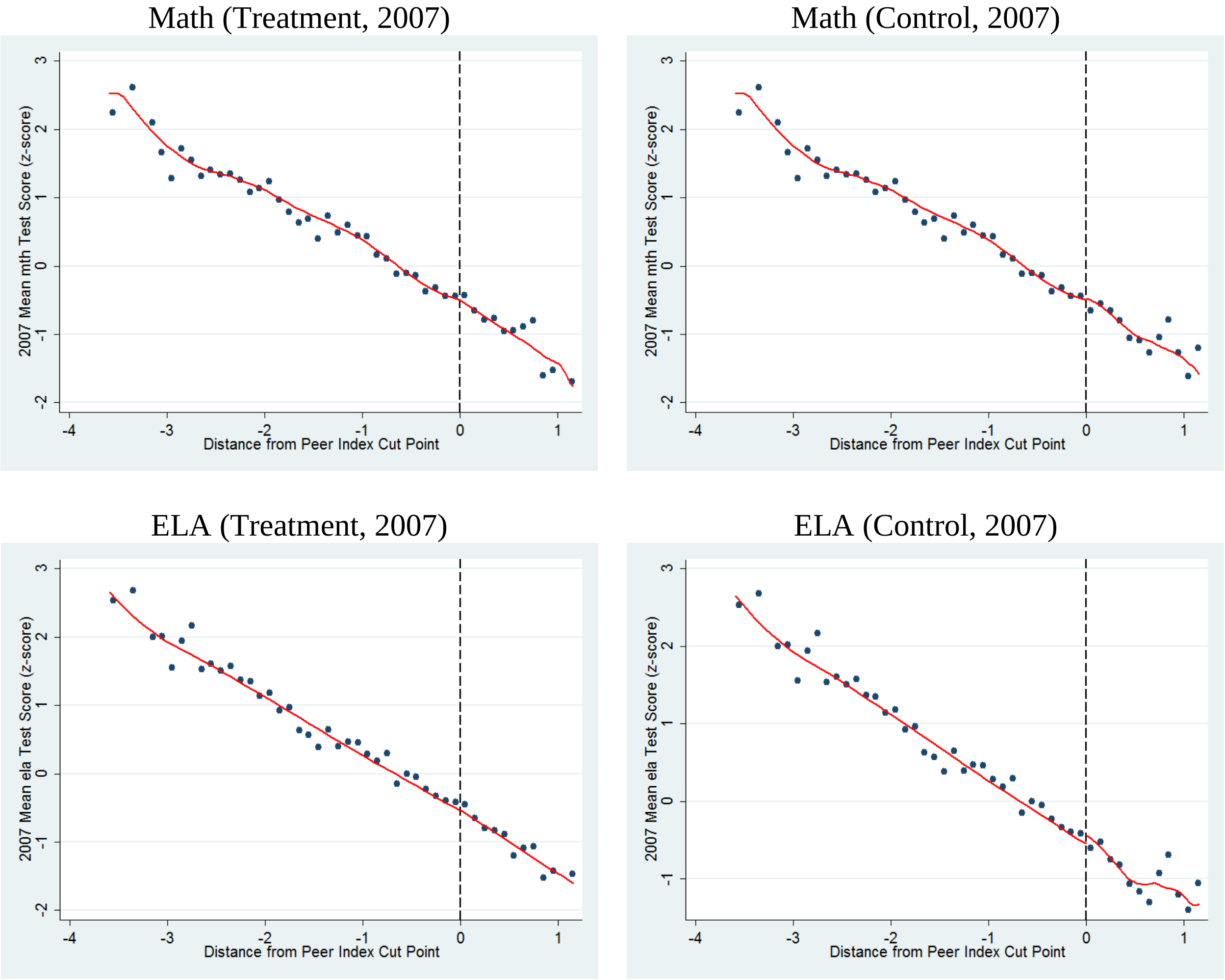


Figure 3 – Test Scores by Distance from Peer Index Cut Point



Math (Treatment, 2007)

Math (Control, 2007)

ELA (Treatment, 2007)

ELA (Control, 2007)

A second condition for the validity of an RD design is that observations (in this case, schools) do not have precise control over the assignment variable (Lee and Lemieux, 2010). One test is to compare pre-treatment covariates to the forcing variable. This is shown for one important covariate – the average 2007 test score – in Figure 3. As in Figure 2, the x-axis shows the forcing variable – the peer index – grouped into bins of size 0.1. The left panel of Figure 3 restricts eligible schools to be those in the treatment group while the right panel restricts eligible schools to those in the control group. A line based on a local linear regression is fit separately on either side of the cutoff and is superimposed on the scatterplot. For both groups and subjects, we can see a strong negative relationship between peer index and average math test scores prior to the introduction of the bonus program. The relationship appears to be fairly linear and, as we would expect, has no notable discontinuity at the cut point for eligibility for the bonus program.

4. Results

4.1. Student Test Scores

To test for results consistent with a loss of intrinsic motivation among teachers, I first look at student test scores. Table 4 shows the results of 30 separate regressions, each fit using Equation 1.[11] In each regression, the dependent variable is a school's average test score in math or ELA, measured in units of standard deviation. The sample of schools is restricted to be those that were eligible to for randomization into the bonus program. In the rows labeled "No Controls," the regression is run on a constant and a single indicator variable that is equal to one if a school was randomly selected to be part of the treatment group and zero otherwise. In the rows labeled "With Controls," additional independent variables are added to the regression to control for pre-treatment test scores and demographic variables. In all cases, the coefficients should be viewed as Intention to Treat (ITT) estimates as most (87%), but not all, invited schools elected to participate in the program (Marsh et al., 2011).

In Table 4, as expected, we see no impact of the bonus program on student test scores prior to the implementation of the program. In both 2006 and 2007, all coefficient estimates are close to zero, with three positive and three negative. These results serve as a placebo test and, in combination with Table 2, provide evidence that the randomization procedure was effective. While adding controls to the regression in 2007 reduces the size of the standard error, it does not

[11] Results are weighted by the number of test takers in each school. Unweighted results are similar.

markedly change the size of the coefficients.[12]

Table 4 – Impact of the Bonus Program on Average Student Test Scores

	Pre-Program		Program			Post-Program		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
D_i Coefficient	2006	2007	2008	2009	2010	2011	2012	2013
Math, No Controls	-0.03	0.00	-0.04	-0.09	-0.12*	-0.14*	-0.16**	-0.14*
	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)	(0.08)	(0.08)	(0.07)
Math, With Controls		0.02	-0.04	-0.09**	-0.13***	-0.15***	-0.17***	-0.13**
		(0.03)	(0.03)	(0.04)	(0.05)	(0.05)	(0.05)	(0.06)
ELA, No Controls	0.01	-0.02	-0.02	-0.06	-0.07	-0.13**	-0.11*	-0.12*
		(0.05)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.07)
ELA, With Controls		-0.02	-0.01	-0.05	-0.06	-0.12**	-0.08*	-0.08
		(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.05)	(0.05)
R-Squared								
Math, No Controls	0.00	0.00	0.00	0.01	0.01	0.01	0.02	0.02
Math, With Controls		0.77	0.78	0.63	0.53	0.53	0.49	0.43
ELA, No Controls	0.00	0.00	0.00	0.00	0.01	0.02	0.01	0.01
ELA, With Controls		0.75	0.73	0.59	0.49	0.45	0.46	0.44
N	302	302	302	302	302	302	302	302

Note: each coefficient is the result of a separate regression with school-level mean test score as the dependent variable and treatment (D_i) as the key independent variable. Regressions with controls adjust for the average test score in the same subject in 2007 (or 2006 in the case when the 2007 test score is the dependent variable), the Progress Report Overall Score in 2007 (excluded in the case when the 2007 test score is the dependent variable), the Progress Report Peer Index in 2007, indicators for the school level (K8, ES, MS) in 2008, and the percentage of students in 2008 who were: English Language Learners, Special Education, eligible for free or reduced price lunches, and black or Hispanic. Regressions are weighted by the number of test takers in each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

During the treatment period – corresponding to the years 2008 through 2010 – we see a small but growing negative effect of the teacher bonus program (though often not statistically distinguishable from zero). In math, by 2010, treatment schools have test scores that are 0.12 standard deviations lower than control schools in the baseline specification and 0.13 standard deviations lower in the specification with controls. The former estimate is significant at the 0.10

[12] It is not possible control for pre-treatment test scores in the 2006 regression because 2006 is the first year that the NYCDOE provides test score data on its web site.

level and the latter at the 0.01 level. The increased statistical significance in the regression with controls is largely caused by a reduction in the standard error, as the coefficient changes little. In ELA, we see negative coefficients throughout the 2008 to 2010 time period that are not statistically distinguishable from zero. These results are broadly consistent with those of other researchers who have studied the effect of the NYC teacher bonus program on student outcomes, despite the fact that two studies (Marsh et. al., 2011; Fryer, 2013) use individual student data and one (Goodman and Turner, 2013) explores only the first two years of the program.

With respect to student test scores, one contribution of this paper is to extend the analysis beyond the conclusion of the teacher bonus program and look for results consistent with a loss of intrinsic motivation. In the last three columns of Table 4 – corresponding to the years 2011 to 2013 – we see such results. In math, the negative point estimates become larger in magnitude and more strongly statistically significant. In the uncontrolled regression, the math point estimate reaches a low point at -0.16 standard deviations in 2012. In ELA, the negative point estimates roughly double in magnitude between 2010 and 2011 and become statistically significant at the 0.05 level, whether or not one includes controls to improve precision. In the uncontrolled regression, the ELA point estimate reaches a low of -0.13 standard deviations in 2011, the year immediately following the conclusion of the bonus program.

These results are consistent with a story in which the introduction of the teacher bonus program reduced intrinsic motivation among teachers at the treatment schools. However, these results could be consistent with other potential explanations. Perhaps the bonus program caused teachers at treatment schools to change their practice in a way that was detrimental to student learning. Or, perhaps teachers at control schools – who may have been aware of the bonus program through news reports – increased their productivity in an effort to compete with treatment schools and prove that merit pay doesn't improve performance in New York City.

4.2 Teacher Survey Responses

To further explore whether the observed test score results are caused by a decline in teachers' intrinsic motivation, I analyze teacher responses to the NYC School Survey. As discussed earlier, teachers' responses were converted to Likert-like scale with 0 corresponding to the most negative response and 10 to the most positive. Each school received a score for each question in each year when that question was asked. Table 5 shows the impact of the bonus program on teacher responses to one sample question, where teachers were asked for their agreement with the following statement: "I usually look forward to each working day at my school." It is fit using a simplified version of Equation 3, where z_{st} corresponds to school s's score on the question in year t. While this question was only asked in 2012 and 2013, two and three years after the completion of the bonus program, respectively, it is arguably the most direct measure of intrinsic motivation on the survey.

Table 5 – Impact of the Bonus Program on Teacher Agreement with the Statement: "I usually look forward to each working day at my school"

	(1)	(2)
	2012	2013
D_i Coefficient	-0.22	-0.04
	(0.15)	(0.15)
N	297	301

Note: each coefficient is the result of a separate regression with teacher scores as the dependent variable and treatment (D_i) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

The coefficients are both negative – indicating that teachers at treatment schools were less likely to agree with this statement – but they are statistically indistinguishable from zero. For context, in 2012 this question had a mean score of 7.01 and a standard deviation of 1.22 in the control group. To better understand the trajectory of survey responses before, during, and after the bonus program, I focus on the 45 questions that were consistently asked on the teacher survey between 2007 and 2012. I divide these questions into those related and unrelated to intrinsic motivation in two ways. First, I subjectively divide them based on whether or not I consider the question to be related to intrinsic motivation (see Table A1 in the appendix for a full list). Second, I conduct an exploratory factor analysis using data from the 2012 survey and identify those questions measure the same underlying construct as the "look forward" question shown in Table 5.

Table 6 - Intrinsic Motivation Questions vs. Other Questions (Subjective Division)

	Pre-Pgm.	Program			Post-Program	
	(1)	(2)	(3)	(4)	(5)	(6)
D_i Coefficient	2007	2008	2009	2010	2011	2012
Intrinsic Motivation Questions	-0.16	-0.08	0.05	-0.07	-0.24**	-0.23**
	(0.12)	(0.11)	(0.12)	(0.11)	(0.11)	(0.11)
Other Questions	-0.15	-0.06	0.03	-0.09	-0.25**	-0.23*
	(0.13)	(0.11)	(0.12)	(0.12)	(0.13)	(0.13)
Difference (Intrinsic - Other)	-0.01	-0.02	0.02	0.02	0.00	0.00
	(0.06)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)
N	299	302	302	302	302	297

Note: each coefficient is the result of a separate regression of the mean school-level teacher survey scores as the dependent variable and treatment (D_i) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Tables 6 shows the results when questions are subjectively divided into those that are, and are not, related to intrinsic motivation. In the top row, we see that the treatment group had intrinsic motivation survey scores similar to the control group both before and during the bonus program. In 2011 and 2012, after the bonus program concluded, the intrinsic motivation survey scores for the treatment group were between 0.23 and 0.24 points lower than for the control group, a result that is statistically significant at the 0.05 level. This fact, taken alone, is consistent with a reduction in teachers' intrinsic motivation at schools invited to participate in eh bonus program. However, in the second row of the table, we see a very similar pattern emerge among questions that are unrelated to intrinsic motivation. In the third row, we see results from a difference-in-differences regression where the first difference is between the treatment and control group and the second difference is between those questions related to intrinsic motivation and those that are unrelated.[13] In this row, we see precisely estimated zeros, indicating that it is unlikely that the bonus program had a greater effect on teacher responses to intrinsic motivation questions than on teacher responses to other questions.

In Table 7, I show a similar set of regressions coefficients where the survey questions have been divided based on the results of an exploratory factor analysis. Those labeled "Intrinsic Motivation Questions" in Table 5 are a set of 18 questions that loaded onto the same factor as the "look forward question" discussed above. As a comparison group, in row 2, I select a group of six questions that load onto a factor that appears to be related to school safety.[14] The results in Table 7 are very similar to those in Table 6, though the coefficients are less precisely estimated.

[13] This specification is not a standard difference-in-differences regression, since neither difference is related to time. In the appendix, however, I take time into account in two ways. First, I used a lagged dependent variable model to control pre-program scores in 2007. Second, I use a triple-difference model where the first difference is between treatment and control, the second difference is between intrinsic motivation questions and non-intrinsic motivation questions, and the third difference is between each year and 2007. Once 2007 is controlled for, the results in rows 1 and 2 are similar in direction but no longer statistically significant at the 0.05 level. The results in row 3 continue to be precisely-estimated zeros.

[14] For example, one survey item in this group is: "Gang activity is a problem in my school."

In both tables, it appears that the participating in the bonus program may have had a long term negative effect on teacher survey scores. However there is no indication that intrinsic motivation was more affected than other attitudes measured on the survey.

Table 7 - Intrinsic Motivation Questions vs. Other Questions (Factor Analysis Division)

	Pre-Pgm.	Program			Post-Program	
	(1)	(2)	(3)	(4)	(5)	(6)
D_i Coefficient	2007	2008	2009	2010	2011	2012
Intrinsic Motivation Questions	-0.20	-0.05	0.07	-0.06	-0.27*	-0.27*
	(0.16)	(0.13)	(0.14)	(0.14)	(0.15)	(0.15)
School Safety Questions	-0.19	-0.16	-0.08	-0.21*	-0.25*	-0.25*
	(0.15)	(0.13)	(0.13)	(0.12)	(0.14)	(0.13)
Difference (Intrinsic - Safety)	-0.01	0.10	0.15	0.15	-0.02	-0.01
	(0.16)	(0.13)	(0.11)	(0.12)	(0.12)	(0.12)
N	299	302	302	302	302	297

Note: each coefficient is the result of a separate regression of the mean school-level teacher survey scores as the dependent variable and treatment (D_i) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

4.3. Regression Discontinuity Evidence

In the test score and survey results shown above, I present the estimated coefficients on a *treatment* indicator – equal to 1 for treatment schools and 0 for control schools – rather than on a *control* indicator. Doing so implicitly suggests that results at the treatment schools declined relative to their counterfactual. However, it is also plausible that the treatment schools held steady – on the path they would have had the bonus program not been implemented – and the control schools improved. To test for such a phenomenon, which would be inconsistent with a decline in intrinsic motivation among treatment teachers, I use a Regression Discontinuity (RD) design.

To implement the RD study, I take advantage of the rigid discontinuity in school eligibility for the bonus program. Prior to randomization, 430 high-need schools were selected to be eligible for the bonus based on their peer index on the NYC Progress Report.[15] These *bonus-eligible* schools – a subset of all 1,217 NYC schools that received a Progress Report in 2007 – were selected entirely based on their Progress Report peer index. Within each of five school types – elementary, K-8, middle, high, and transfer – schools above a particular peer index score were chosen to be eligible for the teacher bonus program and schools below that score were ineligible.[16] Conceptually, for an RD study, I want to compare bonus-eligible schools just above the peer index cut point to bonus-ineligible schools just below the cut point. Assuming that bonus-ineligible schools just below the cut point present a valid counterfactual for bonus-eligible schools just above the cut point, once the peer index (i.e. forcing variable) is controlled for, any difference in outcomes reflects the impact of being eligible for the bonus program.

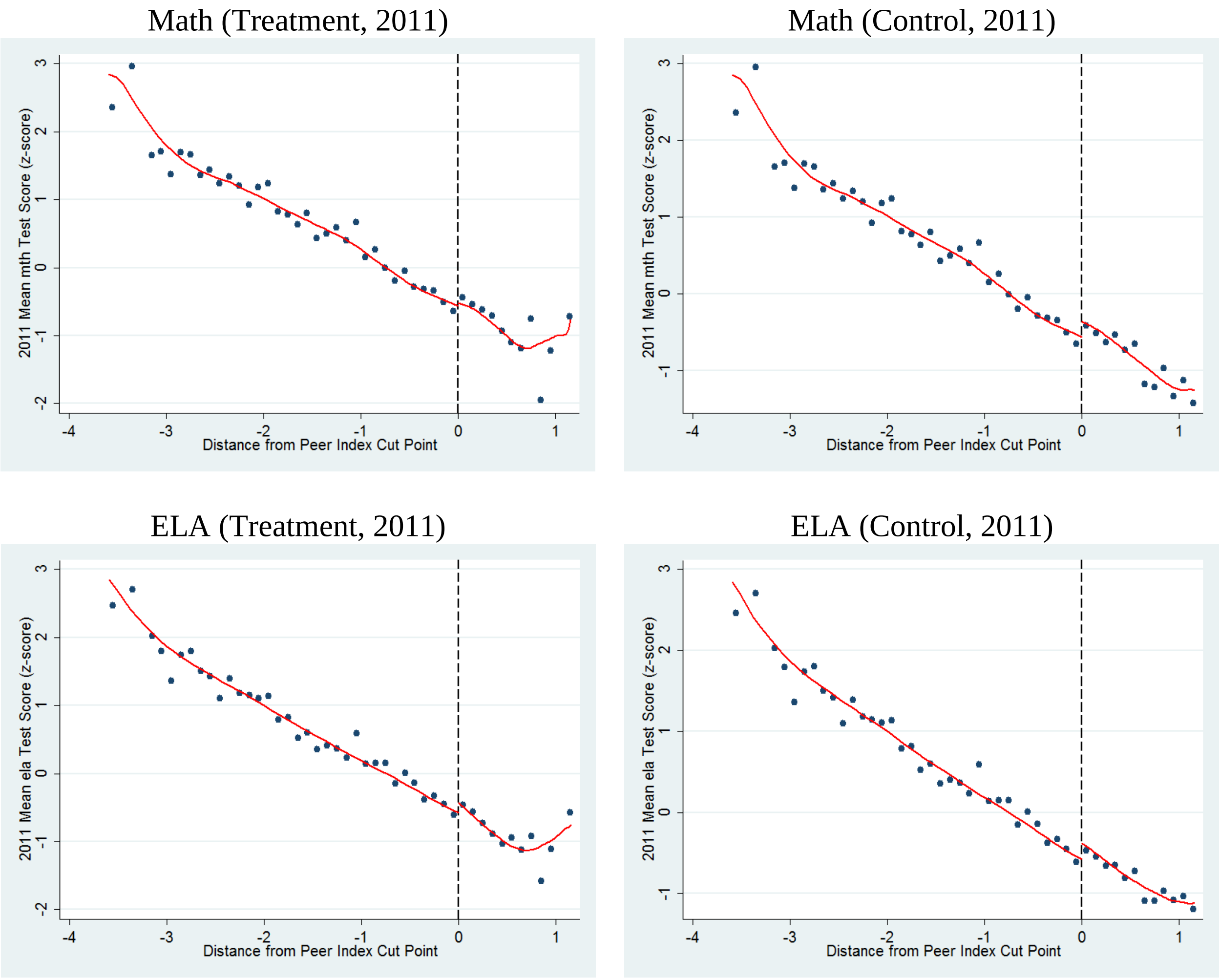
While the 302 elementary, middle, and K-8 schools I use in my main analysis are valid for an RCT, there are two adjustments I need to make to ensure the validity of the RD. First, as documented in Fryer (2013), prior to randomization, 34 bonus eligible schools were barred by the teachers' union (UFT) for "unknown reasons." I add back the 15 elementary, middle, and K-8 schools among these 34 to ensure that any effect found by the RD is not caused by the UFT excluding these schools from the eligible group (since they made no similar exclusion among the ineligible group). Second, schools that spanned the middle and high school grades had two

[15] Note that my study focuses on the 302 elementary, middle, and K-8 schools that had consistent testing data from 2007 to 2013 and were not barred from participation by the UFT.

[16] For middle schools, high schools, and transfer schools, a lower peer index meant a higher-need school. Technically, therefore, schools *below* a certain peer index threshold were eligible for the bonus program. In this analysis, I normalize all peer indices by calculating z-scores within school type and then defining a higher z-score to be a higher-need schools (multiplying by -1 where needed).

chances to become bonus eligible, since they received a separate peer index for their middle and high school portions, and qualified if either peer index was above the cutoff. Prior to the RD, I remove all schools that received both middle and high school peer indices from the analysis, whether or not they appeared in the bonus eligible group. This reduced the bonus eligible group by six schools and the non-eligible group by 24 schools. In total, this leaves 311 bonus eligible elementary, middle, and K-8 schools which I compare to 607 non-eligible schools.

Figure 4 – Test Scores by Distance from Peer Index Cut Point (2011)



Math (Treatment, 2011)

Math (Control, 2011)

ELA (Treatment, 2011)

ELA (Control, 2011)

Figure 4 presents the results of the RD graphically for 2011, the first year after the bonus program was concluded. The graph is organized in the same way as Figure 3, with math test scores on the top ELA test scores on the bottom. In all four panels, to the left of the dotted line we see average test scores for schools that were not eligible for randomization based on their peer index, which serve as the counterfactual. In the left two panels, which compare non-eligible schools to treatment schools, we see little notable discontinuity in 2011 test scores between the two groups. To the extent there is any discontinuity, it appears that treatment schools may have outperformed non-eligible schools. In the right two panels, which compare non-eligible schools to control schools, we see a more noticeable discontinuity. In particular, controls schools near the cut point – just to the right of the dotted line – appear to have higher 2011 average test scores than would be predicted based on their peer index.

Table 8 provides regression evidence to support the visual evidence in Figure 4. Each entry in Table 8 shows the coefficient and standard error for E_s– a dummy variable indicating eligibility to be randomized for the teacher bonus program – from linear regression to fit Equation 4. As explained above, I fit a linear regression model because the data – even in the pre-program period – have a strong linear relationship and higher-order polynomials appear to overfit the data, failing the placebo tests. Consistent with the recommendation of Gelman and Imbens (2014), I also present RD results using a local linear regression in an appendix, and the results are very similar to those presented in Table 8.

In Table 8, we see that eligibility for the bonus program had no impact on test scores in either the treatment or control group in the two years prior to the program's announcement (2006 and 2007). This result, consistent with the visual evidence in Figure 3, serves as a falsification test. In the top two rows of Table 8, we see little impact of an effect of the bonus program on treatment school test scores after the program began in 2008. The treatment-school estimates are generally fairly small not significantly different from zero. If anything, the coefficients are somewhat more likely to be positive than negative, a finding that is inconsistent with what one would expect if teachers' intrinsic motivation had declined in treatment schools.

Table 8 – Impact of Bonus Eligibility on Student Test Scores (Regression Discontinuity)

	Pre-Program		Program			Post-Program			
E_s	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
Coefficient	2006	2007	2008	2009	2010	2011	2012	2013	N =
Math, Treatment	-0.01	0.00	-0.01	0.08	0.06	0.12	0.07	-0.05	793
	(0.07)	(0.07)	(0.07)	(0.08)	(0.09)	(0.09)	(0.10)	(0.09)	
ELA, Treatment	0.01	0.03	0.08	0.08	0.14*	0.11	0.02	-0.04	793
	(0.07)	(0.06)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	
Math, Control	0.05	0.05	0.06	0.16*	0.17*	0.28***	0.24**	0.09	745
	(0.07)	(0.08)	(0.08)	(0.09)	(0.09)	(0.10)	(0.09)	(0.09)	
ELA, Control	0.08	0.07	0.15**	0.16**	0.20***	0.21***	0.15**	0.08	745
	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)	

Note: Each number is the coefficient on E_s in a separate regression of school-level mean test score on distance from the eligiblity cut point, an indicator for bonus eligiblity (E_s), and an interaction of the two (allowing the slopes to vary on either side of the cut point). Regressions are weighted by the number of test takers in each school. Robust standard errors are in parentheses. The sample size is the same in each year because only schools with testing data in each year were included. *** p<0.01, ** p<0.05, * p<0.1

The bottom two rows of Table 8 restrict the eligible sample to control schools only. When running the RD, I am then comparing ineligible schools – which never had the opportunity to be randomized into the bonus program – with eligible schools that were randomly selected to be control schools. As we see in Table 8, control schools right around the eligibility cut point had significantly higher test scores than ineligible schools with similar peer indices. Between 2009 and 2012, the estimated coefficients range from a 0.15 to a 0.28 school-level standard deviation increase in test scores, and all results are significant at the 0.10 level with most significant at the 0.05 level. These results – when combined with the null effect observed in treatment schools – provide more nuance to the negative RCT results observed earlier in this paper. It appears that the negative impact of the bonus program was not so much the result of a decline in performance among treated schools, but rather an improvement in performance among control schools.

Table 9 – Comparison of RD and RCT Results

	Pre-Program		Program			Post-Program		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	2006	2007	2008	2009	2010	2011	2012	2013
A. RDD (Treatment Schools)								
Math (All Schools)	-0.01	0.00	-0.01	0.08	0.06	0.12	0.07	-0.05
ELA (All Schools)	0.01	0.03	0.08	0.08	0.14*	0.11	0.02	-0.04
B. RDD (Control Schools)								
Math (All Schools)	0.05	0.05	0.06	0.16*	0.17*	0.28***	0.24**	0.09
ELA (All Schools)	0.08	0.07	0.15**	0.16**	0.20***	0.21***	0.15**	0.08
C. Treatment - Control								
Math (All Schools)	-0.06	-0.05	-0.07	-0.08	-0.11	-0.16	-0.17	-0.14
ELA (All Schools)	-0.07	-0.04	-0.07	-0.08	-0.06	-0.10	-0.13	-0.12
D. RCT Results (Treatment Effect)								
Math, No Controls	-0.03	0.00	-0.04	-0.09	-0.12*	-0.14*	-0.16**	-0.14*
ELA, No Controls	0.01	-0.02	-0.02	-0.06	-0.07	-0.13**	-0.11*	-0.12*
Sample Size								
Panel A (Treatment)	793	793	793	793	793	793	793	793
Panel B (Control)	745	745	745	745	745	745	745	745

Note: Each number in Panel A and B is the coefficient on a separate regression of school-level mean test score on distance from the eligiblity cut point, an indicator for bonus eligiblity (Ei), and an interaction of the two (allowing the slopes to vary on either side of the cut point). Panel A restricts eligible schools to include treatment schools only. Panel B restricts eligible schools to include control schools only. Panel C shows the difference between the estimates in Panel A and Panel B. Panel D reproduces the treatment coefficient from the "no control" regressions in Table 2. Regressions are weighted by the number of test takers in each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 9 explicitly shows the degree to which the RD and RCT results align with one another. In Panel A and B, I replicate the coefficients from Table 8, omitting standard errors for clarity. In Panel C, I subtract the RD coefficients for control schools from the RD coefficients for treatment schools, obtaining one estimate of the difference between treatment and control schools. These estimates are very close to the RCT results from Table 4, which I have replicated (for the regression with no controls) in Panel D of Table 9. While the RCT and the RD show largely the same estimated difference between the treatment and control group, the RD gives us more insight into the cause of this difference. Rather than a decline in performance among treated schools – as would be consistent with a decline in teachers' intrinsic motivation – it appears likely that control schools improved their performance over what would have happened in the absence of the bonus program.

5. Conclusion

In this paper, I test whether the implementation of a large teacher bonus program in New York City had a detrimental effect on teachers' intrinsic motivation. Initially reviewing data from the Randomized Controlled Trial, I find some evidence consistent with this hypothesis. Teachers at treatment schools responded more negatively to questions related to intrinsic motivation and students at treatment schools scored lower on standardized tests than their counterparts at randomly assigned control schools. However, exploring further, it appears unlikely that a decline in teachers' intrinsic motivation drove these results. When looking at teacher survey scores, we see that scores at treatment schools dropped by a similar amount for questions unrelated to intrinsic motivation. And when looking at student test scores, a Regression Discontinuity analysis shows that students at treatment schools did not, in fact, decline relative to their (unobserved) counterfactual, as would be predicted if the bonus program lowered teachers' intrinsic motivation. Rather, it appears that students at control schools scored higher than their (also unobserved) counterfactual.

This finding – that control schools appear to have improved because of the bonus program – is not one that I expected to find when I began this research. Moreover, it relies on the assumption that control schools knew they were control schools – as opposed to ineligible schools – and did something to improve their students' performance as a result. As such, a Bayesian statistician might suggest I place relatively little weight on the "statistically significant" results for control schools in Tables 8 and 9, since my prior belief was strongly against finding an improvement in the control schools (they were, after all, meant to provide a counterfactual for the treatment schools).

While unlikely, however, it is not infeasible that teachers at control schools might have known they were working in control schools and improved their performance. The bonus program was widely publicized when it was implemented, and the NYCDOE published a full list of invited treatment schools that, if combined with information on their peer index, would have enabled schools to know if they were in the control group with reasonable certainty.[17] [18] Moreover, it has been widely observed that the behavior of experimental units may change once they are aware they are participating in a study, a phenomenon broadly referred to as a Hawthorne Effect. Once specific form is known as the "John Henry Effect," where a control group is aware they are the control group for an experiment and seeks to work harder as a result.

[17] See, for example, a *New York Times* article in 2008 announcing the distribution of $14.2 million dollars to teachers after the first year of the program (http://www.nytimes.com/2008/09/19/education/19bonus.html)

[18] The announcement with the list of invited treatment schools was available at the following link until recently: http://schools.nyc.gov/Offices/mediarelations/NewsandSpeeches/2007-2008/20071218_performance_pay.htm. I have an archived copy and can provide it upon request.

Such an effect, if it took place during the NYC teacher bonus program, might explain the results observed in this paper.

More broadly, the RD results serve as an important reminder when interpreting RCT results. An RCT is generally considered the "gold standard" for causal evidence in the social sciences because the control group is assumed to be a valid counterfactual for the treatment group. The control group answers the question: what would have happened to the treatment group had the experiment not be implemented? However, the results in this paper remind us that, even in an RCT, the treatment group counterfactual is never actually observed. Even with proper randomization, as by all accounts occurred in the NYC teacher bonus program, the control group is still an imperfect estimate of what would have happened in the treatment group had the experiment not occurred.

References

Deci, E. (1971): “Effects of externally mediated rewards on intrinsic motivation.” *Journal of Personality and Social Psychology*, 18(1), 105-115.

Deci, E. (1975): Intrinsic Motivation. New York: Plenum Publishing Co.

Deci, E. and Ryan, R (2014): Intrinsic Motivation Inventory. Downloaded from http://www.selfdeterminationtheory.org/ on 7/14/2014

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

Gelman, A and G. Imbens (2014): “Why High-Order Polynomials Should Not be Used in Regression Discontinuity Designs.” *NBER Working Paper 20405*.

Glewwe, P., N. Ilias, and M. Kremer (2010): “Teacher Incentives.” *American Economic Journal: Applied Economics*, 2, 1-25.

Gneezy, U. and A. Rustichini (2000a): “Pay Enough or Don't Pay At All.” *Quarterly Journal of Economics*. August 2000, 791-810.

Gneezy, U. and A. Rustichini (2000b): “A Fine is a Price.” *Journal of Legal Studies*, vol. XXIX, part 1, 1-18.

Goodman, S. and L. 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), 409-20.

Kane T., and D. Staiger (2008). “Estimating Teacher Impacts on Student Achievement: An Experimental Validation.” NBER Working Paper No. 14607.

Lee, D. and T Lemieux (2010): “Regression Discontinuity Designs in Economics” *Journal of Economic Literature* 48: 281–355

Lemieux, T., MacLeod, W. B., and Parent, D (2009). “Performance Pay and Wage Inequality” *Quarterly Journal of Economics*, 124 (1), 1-49

Marsh, J., M. Springer, D. McCaffrey, K. Yuan, S. Epstein, J. Koppich, N. Kalra, C. DiMartino, A. Peng (2011): A Big Apple for Educators: New York City’s Experiment with Schoolwide Performance Bonuses: Final Evaluation Report. Santa Monica, CA: RAND Corporation http://www.rand.org/pubs/monographs/MG1114.html

Muralidharan, K., and V. Sundararaman (2011): “Teacher Performance Pay: Experimental Evidence from India," *The Journal of Political Economy*, 119(1), 39-77.

Nichols, A. (2011): “rd 2.0: Revised Stata module for regression discontinuity estimation.” http://ideas.repec.org/c/boc/bocode/s456888.html

Rivkin, S., E. Hanushek, and J. Kain (2005). “Teachers, Schools, and Academic Achievement.” *Econometrica*, 73(2), 417-458.

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

Springer, M., D. Ballou, L. Hamilton, V. Le, J. Lockwood, D. McCaffrey, M. Pepper and B. Stecher (2010): Teacher Pay For Performance: Experimental Evidence from the Project on Incentives in Teaching. Santa Monica, CA: RAND Corporation, ttp://www.rand.org/pubs/reprints/RP1416.

Table A1 – Subjective Division of Teacher Survey Questions

Intrinsic Motivation Questions (Subjective Division)

Number	Statement
t_q2b	Teachers in this school set high standards for student work in their classes
t_q5a	To what extent do you feel supported by: your principal
t_q5c	To what extent do you feel supported by: other teachers at your school
t_q6b	School leaders invite teachers to play a meaningful role in setting goals and making important decisions for this school
t_q6d	Teachers in my school respect teachers who take the lead in school improvement efforts
t_q6e	Teachers in my school trust each other
t_q6f	Teachers in my school recognize and respect colleagues who are the most effective teachers
t_q6h	School leaders give me regular and helpful feedback about my teaching
t_q6i	School leaders place a high priority on the quality of teaching at this school

Other Questions (Subjective Division)

Number	Statement
t_q1a	School leaders communicate a clear vision for this school
t_q1c	School leaders encourage open communication on important school issues
t_q1d	Curriculum, instruction, and assessment are aligned within and across the grade levels at this school
t_q1e	The principal places the learning needs of children ahead of other interests
t_q1f	The principal is an effective manager who makes the school run smoothly
t_q1g	I trust the principal at his or her word
t_q2c	My school has clear measures of progress for student achievement throughout the year
t_q2d	This school makes it a priority to help students develop challenging learning goals
t_q2e	This school makes it a priority to help students find the best ways to achieve their learning goals
t_q4	My school offers a wide enough variety of courses to keep students at my school engaged
t_q6a	The principal has confidence in the expertise of the teachers
t_q6c	School leaders provide time for collaboration among teachers
t_q6g	School leaders visit classrooms to observe the quality of teaching at this school
t_q6k	Most teachers in my school work together on teams to improve their instructional practice

t_q6l	Teachers in my school use student achievement data to improve instructional decisions
t_q7a	The professional development I received this year provided me with teaching strategies to better meet the needs of my students
t_q7b	I have sufficient materials to teach my class(es), including: books, audio/visual equipment, maps, and/or calculators
t_q7d	This year, I received helpful training on the use of student achievement data to improve teaching and learning
t_q7e	The professional development I received this year provided me with content support in my subject area
t_q8a	Obtaining information from parents about student learning needs is a priority at my school
t_q8b	Teachers and administrators in my school use information from parents to improve instructional practices and meet student learning needs
t_q8c	My school communicates effectively with parents when students misbehave
t_q10d	How often during this school year have you: sent parents written information on what you are teaching and what students are expected to learn
t_q10e	How often during this school year have you: sent home information on services to help students or parents such as: tutoring, after-school programs, or workshops adults can attend to help their children in school
t_q11a	Order and discipline are maintained at my school
t_q11b	I can get the help I need at my school to address student behavior and discipline problems
t_q11c	I am safe at my school
t_q11d	Crime and violence are a problem in my school
t_q11e	Students in my school are often threatened or bullied
t_q11g	Most students at my school treat teachers with respect
t_q11h	Most parents treat teachers at this school with respect
t_q11j	Students' use of alcohol and illegal drugs in school is a problem at my school
t_q11k	There are conflicts at my school based on race, color, creed, ethnicity, national origin, citizenship/immigration status, religion, gender, gender identity, gender expression, sexual orientation or disability
t_q11l	There is a person or a program in my school that helps students resolve conflicts
t_q11m	Gang activity is a problem in my school
t_q11n	My school is kept clean

Table A2 - Intrinsic Motivation Questions vs. Other Questions (Subjective Division, Controls for 2007)

	Pre-Pgm.	Program			Post-Program	
	(1)	(2)	(3)	(4)	(5)	(6)
D_i Coefficient	2007	2008	2009	2010	2011	2012
Intrinsic Motivation Questions	N/A	0.07	0.17*	0.05	-0.15	-0.14
		(0.08)	(0.09)	(0.09)	(0.09)	(0.10)
Other Questions	N/A	0.08	0.15	0.03	-0.14	-0.12
		(0.08)	(0.09)	(0.09)	(0.11)	(0.11)
Difference (Intrinsic - Other)	N/A	-0.01	0.02	0.03	-0.00	-0.01
		(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
N	N/A	299	299	299	299	294

Note: each coefficient is the result of a separate regression of the mean school-level teacher survey scores as the dependent variable and treatment (D_i) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A3 - Intrinsic Motivation Questions vs. Other Questions (Factor Analysis Division, Controls for 2007)

	Pre-Pgm.	Program			Post-Program	
	(1)	(2)	(3)	(4)	(5)	(6)
D_i Coefficient	2007	2008	2009	2010	2011	2012
Intrinsic Motivation Questions	N/A	0.11	0.21*	0.07	-0.16	-0.17
		(0.10)	(0.12)	(0.12)	(0.13)	(0.13)
School Safety Questions	N/A	0.01	0.06	-0.06	-0.12	-0.11
		(0.08)	(0.09)	(0.09)	(0.10)	(0.10)
Difference (Intrinsic - Safety)	N/A	0.14	0.17*	0.18*	-0.00	-0.00
		(0.09)	(0.09)	(0.10)	(0.10)	(0.10)
N	N/A	299	299	299	299	294

Note: each coefficient is the result of a separate regression of the mean school-level teacher survey scores as the dependent variable and treatment (D_i) as the independent variable. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A4 – Teacher Survey Triple Difference Regression

	Pre-Pgm.	Program			Post-Program	
	(1)	(2)	(3)	(4)	(5)	(6)
D_i Coefficient	2007	2008	2009	2010	2011	2012
Subjective Division	N/A	-0.01	0.02	0.03	-0.00	-0.01
		(0.04)	(0.05)	(0.05)	(0.05)	(0.06)
Factor Analysis Division	N/A	0.16	0.19	0.20	0.02	0.02
		(0.11)	(0.12)	(0.12)	(0.13)	(0.14)
N	N/A	302	302	302	302	297

Note: each coefficient is the result of a separate regression. The dependent variable is a difference-in-difference variable based on school-level survey scores. The first difference is between survey scores related to intrinsic motivation and those unrelated. The second difference is between the year shown and 2007. The independent variable is a dummy (D_i) indicating whether or not the school was randomly offered treatment. Regressions are weighted by the number of responders at each school. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table A5 – Impact of Bonus Eligibility on Student Test Scores (RD, Local Linear Regression)

	Pre-Program		Program			Post-Program			
E_s	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
Coefficient	2006	2007	2008	2009	2010	2011	2012	2013	N =
Math, Treatment	0.05	0.08	0.07	0.19**	0.16	0.16	0.12	0.07	793
	(0.08)	(0.08)	(0.08)	(0.09)	(0.10)	(0.10)	(0.11)	(0.10)	
ELA, Treatment	0.05	0.02	0.07	0.07	0.12	0.11	0.06	0.04	793
	(0.07)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)	(0.09)	
Math, Control	0.13*	0.14	0.15*	0.25***	0.25**	0.31***	0.27**	0.21**	745
	(0.08)	(0.09)	(0.08)	(0.09)	(0.10)	(0.11)	(0.11)	(0.10)	
ELA, Control	0.13*	0.08	0.15**	0.15**	0.17**	0.20**	0.15**	0.12	745
	(0.08)	(0.07)	(0.07)	(0.08)	(0.07)	(0.08)	(0.08)	(0.08)	

Note: Each number is the coefficient on E_s in a separate local linear regression of school-level mean test score on distance from the eligiblity cut point and an indicator for bonus eligiblity (E_s). The regressions are fit separately on either side of the cut point, and are weighted by the number of test takers in each school. Robust standard errors are in parentheses. The sample size is the same in each year because only schools with testing data in each year were included. *** p<0.01, ** p<0.05, * p<0.1