

**Jumping at the Chance:
The Effects of Teacher Bonuses on Student Achievement¹**

Douglas Lee Lauen
Assistant Professor of Public Policy
University of North Carolina at Chapel Hill

April 20, 2011

Working Paper
Please do not cite without permission

Abstract

Pay for performance plans are also spreading across the country due to encouragement from the Obama administration's \$4 billion Race to the Top initiative, which places a high priority on merit pay. Through a program that involved public accountability and bonuses, the state of North Carolina awarded over one billion dollars in school-based performance bonuses for meeting test score growth targets between 1997 and 2009. Using statewide student-level data on 5th graders from North Carolina, I examine the effects of failing to earn a bonus on reading, math, and science test scores in 2008, a year in which math and reading scores were "high-stakes," and science tests were "low stakes." Results from regression discontinuity estimates show for the average student failing the bonus target causes higher reading gains and no adverse effects on science scores, suggesting that the bonus program did not narrow the curriculum at the expense of science. Among low achieving students, however, negative effects on science scores emerge, which must be weighed against positive effects among low achieving students in both reading and math.

¹ The author thanks Ashu Handa, Gary Henry, Tom Cook, and Vivian Wong for their feedback, the North Carolina Education Research Data Center at Duke University for providing the data, the Spencer Foundation and UNC-Chapel Hill for financial support, and Michael Gaddis for research assistance. This paper was presented at the Society for Research on Educational Effectiveness in March 2011.

I. Introduction

During the 1960s and 1970s, scholars became quite pessimistic about the role schools play in reducing inequality (Coleman et al., 1966; Jencks, 1972). The current scholarly consensus is that while schools may not matter much, teachers do, and that to close test score gaps districts and schools must recruit, retain, and reward high quality teachers, especially those serving disadvantaged populations (Goldhaber, 2008; Hanushek, 1992; Rivkin, Hanushek, & Kain, 2005; Sanders & Rivers, 1996). This seemingly clear mandate begs a key question: how should we evaluate and reward teachers? The complexity and expense involved in evaluating individual teachers in the classroom and the desire to reward teachers for outputs rather than inputs has led to increased interest in paying teachers for their ability to raise student test scores. Pay for performance plans are also spreading across the country due to encouragement from the Obama administration's \$4 billion Race to the Top initiative, which places a high priority on merit pay.

Teacher bonus programs likely work through two mechanisms: teachers seeking to maximize income and minimize public "accountability threats" (Carnoy & Loeb, 2002; Figlio & Rouse, 2006; Hanushek & Raymond, 2005; Jacob, 2005). Existing research focuses mainly on the first mechanism, with careful designs to determine whether individual or group based incentives have positive effects on intended outcomes and minimal unintended consequences (Glewwe, Ilias, & Kremer, 2010; Lavy, 2002, 2009; McCaffrey, Pane, Springer, Burns, & Haas, 2011; Muralidharan & Sundararaman, 2009; Springer et al., 2010). Bonuses from most public sector programs, however, may not be large enough to induce substantial changes in teaching practice. Moreover, teacher bonus programs are a matter of public record and may be embedded within public accountability policies which seek to hold schools, principals, and groups or individual teachers publicly accountable for student performance. Therefore, in addition to small

scale randomized experiments, it is important for the research base on teacher incentives to include well designed studies of bonus programs linked to public accountability.

Through a program that involved public accountability and bonuses, the state of North Carolina awarded over one billion dollars in school-based performance bonuses for meeting test score growth targets between 1997-1998 and 2008-2009 (about \$100 million per year).² During this time period, teachers received \$750 each if their school met an “expected growth” target and \$1,500 if their school met a “high growth” target. Eligibility for both \$750 bonuses and public recognition was determined by an “expected growth” score calculated by the state department of public instruction’s accountability office. Because assignment to treatment was made by public officials on the basis of a continuous variable with a sharp threshold, this program presents an opportunity to examine the effects of teacher bonuses on student achievement in North Carolina with a regression discontinuity design.

Using statewide student-level data from North Carolina, I test whether failing to attain the \$750 bonus target in the spring of 2007 increases 5th grade math and reading test score gains in the spring of 2008, a year in which elementary schools were held accountable for 5th grade math and reading scores. I hypothesize that failing to achieve growth targets will induce educators to raise test scores the following year. In response to accountability threats educators may increase time on tested subjects, alter the allocation of teachers to students, implement new curriculum, and tutoring programs, to name just a few possibilities. These responses could have adverse affects on non-tested subjects or particular subgroups of students. Therefore, I hypothesize negative or null effects on low stakes test scores outside the accountability framework, and differential effects based on student position in the baseline test score distribution. I examine

² Personal email correspondence with Kristopher Nordstrom, North Carolina Fiscal Research, September 13, 2010.

whether bonuses induce incentive effects to narrow the curriculum at the expense of 5th grade science achievement, a subject tested for the first time in 2008 and not part of the school accountability system. The comparison of effects across the three subjects is thus a comparison of the effects of bonuses on high and low stakes tests. To determine whether bonuses widen or narrow test score gaps, I estimate differential effects on each subject test score for students low, average, and high in the within-school reading and math achievement distribution.

I find that failing the bonus target causes higher reading gains, slightly lower science scores for low achievers, larger gains for low and high achieving students in reading (relative to average students), and larger gains for low and average students in math. This evidence suggests that even modest group-based monetary incentives combined with public accountability holds some promise as a reform strategy. Low performing students in particular appear to benefit from this approach. These positive differential effects for low achievers in reading and math, however, come at the expense of small declines in science scores.

This study complements existing research on pay for performance by testing whether a statewide program of school-based incentives embedded in a long-standing public accountability system has had positive intended effects and minimal negative unintended effects. It does so using a design with both strong internal and external validity. As federal, state, and district officials move toward implementing teacher bonus programs, there is a pressing need for evidence on the effectiveness of teacher bonus programs. Given the budget crisis in most states, teacher pay for performance is likely to be a small supplement to base pay and linked to public accountability programs to enhance transparency and public support. Therefore, North Carolina's recent experience with teacher bonuses should be instructive for policy design.

II. Background and Evidence

Teachers are the most important and most expensive school-based resource (Goldhaber, 2008; Hanushek, 1992; Rivkin, et al., 2005; Sanders & Rivers, 1996). Moreover, teachers vary considerably in their ability to teach the academic skills tested on standardized tests. An early study found that the most effective teachers produced a full year's additional test score growth relative to the least effective teachers (Hanushek, 1992). Another study found a difference of fifty percentile points in achievement between students who had three consecutive years of highly effective teachers and students who had three consecutive years of the least effective teachers (Sanders & Rivers, 1996). Though teachers vary in their ability to raise student test scores, virtually all U.S. teachers are paid by years of experience, degree, and continuing education credits, factors that are not strongly predictive of teacher value added (Goldhaber, 2008). Recent reforms to the single salary schedule include paying teachers for knowledge and skills that experts believe to be connected to student achievement and paying teachers for performance (Podgursky & Springer, 2007). Pay for performance plans, on the other hand, are designed to compensate teachers for their output rather than their inputs. Most programs implemented in the U.S. are hybrids that include both pay for knowledge and skills and pay for performance (ibid).

With public sector unions, especially teachers unions, under attack, the increasingly influential role of the Gates and other foundations in education, and the Obama administration pushing pay for performance in Race to the Top stimulus funding, evaluating and paying teachers based on their value added to student test scores is now a central focus of education policy reform. As states and districts move toward implementing new teacher evaluation systems, there are several design elements to consider. First, can pay for performance plans be

designed to avoid the multitasking problem (Campbell, 1979; Dixit, 2002; Heckman, Heinrich, & Smith, 2002)? Teaching is a complex task and tests are imperfect instruments of student learning (Koretz, 2002). Pay for performance plans risk setting up incentives for teachers to focus on short-run objectives, such as reading comprehension and math computation skills, rather than skills and knowledge that may be crucial for a student's academic development in the long run, such as critical thinking skills, a love of learning, and social and scientific knowledge. Research on high stakes accountability has found increases in test-specific skills, but not generalizable knowledge of academic content (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998), a tendency to narrow and fragment the curriculum, and rote, teacher-directed instruction (Amrien & Berliner, 2002; Darling-Hammond, 2004; Linn, 2000; S. L. Nichols & Berliner, 2007; Orfield & Kornhaber, 2001; Valenzuela, 2005). Second, does pay for performance set up incentives to cream skim or triage based on how easy or difficult a student is to educate? Incentive plans based on the percentage of students at grade level, for example, could create incentives to focus on kids just below grade level at the expense of students in other parts of the achievement distribution. Plans based on test score growth, could create incentives for teachers to target resources at students with the greatest past history of growth. In schools facing accountability pressure, teachers and principals may manipulate the test-taking pool through selective disciplinary practices and reclassifying students as requiring special educational services, thereby making them ineligible for tests (Figlio, 2006; Heilig & Darling-Hammond, 2008; Jacob, 2005). Schools under accountability pressure may focus instruction and extra resources on those students most likely to improve a school's external standing (Booher-Jennings, 2005; Ladd & Lauen, 2010; Neal & Schanzenbach, 2010; Weitz & Rosenbaum, 2007). Third, should awards be based on individual or group-based performance? Individual-based

incentives have the advantage of being directly tied to a particular teacher and therefore the unit of accountability and the alignment of incentives is relatively straightforward. On the other hand, bonuses paid to individual teachers could harm professional community and to the extent that this is an important resource for school-wide performance (Bryk & Schneider, 2002), individual-based incentives could be less optimal.

At least three well designed randomized controlled trials or regression discontinuity studies have tested the effects of teacher-specific bonuses. The first, based in India, randomly assigned 100 schools to an individual-based incentive, 100 to a school-based incentive, 200 to receive extra resources, and 100 to control. This study found positive effects of both individual and group-based incentives relative to the control group and no adverse effects on conceptual (as opposed to mechanical) components of test scores and morale. In the second year of the study, teachers in individual incentive schools outperformed teachers in group-based incentive schools (Muralidharan & Sundararaman, 2009). A second study, based in Israel, using an RD design and the fact that some schools were mistakenly assigned to treatment and control (N=49 high schools, 629 teachers), found that individual teacher incentives based on a tournament design (a fixed pot of funds and a fixed number of winners) increased test taking and test passing rates in math and English exams (Lavy, 2009). A strength of this study is that it measured and found changes in teaching methods, after-school teaching, and effort, which lessens concerns that gaming and opportunistic behavior produced the results. The most recent study, based in metro Nashville schools in the United States, randomly assigned about 300 middle school math teachers to two groups: a treatment group eligible for bonuses of up to \$15,000 for student math test score gains, and a control group that was not eligible for bonuses (Springer, et al., 2010). Teachers were randomized within blocks of schools (based on student achievement) and clusters

based on course groups. About two-thirds of teachers volunteered for the experiment, but attrition from the experiment was relatively high (only 148 teachers remained through the end of the third year). The study found no significant difference overall in the student math scores of fifth through eighth grade students assigned to treatment and control. The sign of the treatment effect was positive, though, and with a somewhat larger sample size would have been small and statistically significant. Moreover, the study found fairly strong and significant effects for 5th grade teachers. Given that most 5th grade teachers teach in self-contained classrooms and most 6th through 8th grade teachers do not, it is conceivable that the incentive changed the behavior of those teachers that had the flexibility to place a larger emphasis on math in their classrooms. The study found increases in 5th grade social studies and science test scores, which suggests that the incentives did not narrow the curriculum. On the other hand, the 5th grade treatment effect did not persist through sixth grade, which raises some doubts about the validity of the 5th grade gains.

Group or team-based pay is another alternative for pay for performance plans. This approach has the advantage of perhaps increasing team cohesion and information sharing and the disadvantage of being only indirectly tied to a teacher's individual performance, which could lead to shirking, especially in large groups or schools. There have been at least three well designed studies of the effects of group incentives in addition to (Muralidharan & Sundararaman, 2009), summarized above. The first, based in Israel, used an RD design (N=55 high schools) to test small school-wide incentives (\$200-\$715 in mid-1990s dollars) found positive effects of incentives on test scores, exit exams, and school graduation rates and larger treatment effects for disadvantaged students (Lavy, 2002). A second study, based in Kenya, randomly assigned 50 schools to the treatment, school-wide bonuses, and 50 schools to control and found positive short

run effects on student test scores, but that these effects faded out by the third year, perhaps due to gaming and opportunistic behavior on the part of teachers in the treatment schools (Glewwe, et al., 2010). Last, in Round Rock, Texas, a study involving nine middle schools assigned 78 teams of teachers (371 teachers teaching 8,361 students) to the treatment of being eligible for bonuses or the control of being ineligible for bonuses. Preliminary results from this study show no treatment-control difference in the student achievement in math, reading, science, or social studies of teams assigned to these groups (McCaffrey, et al., 2011).

Overlooked in previous literature on pay for performance is whether incentives should be a private administrative matter or embedded in a system of public accountability. Previous empirical research is dominated by small-scale pilot studies that offer bonuses to a limited number of teachers. In some cases bonus winners were probably announced; in at least one case, the Nashville experiment, they were kept confidential. These types of experiments are designed to test whether incentives spur teacher's extrinsic motivation to maximize income. But if teachers are not motivated by these factors, but instead by professional norms and pride, perhaps public accountability would be more effective than pay for performance. Public accountability pressure involves setting performance indicators for schools, publicly reporting performance on these indicators, and applying sanctions or rewards based on performance. Survey evidence suggests that high stakes accountability tends to increase time spent on reading and math, the two subjects most often tested and included in accountability systems (Hannaway & Hamilton, 2008; Ladd & Zelli, 2002). Given indications that teachers appear to be devoting more time to reading and math in response to accountability pressure, it is perhaps not surprising that accountability pressure tends to increase test scores on high stakes, and sometimes even low stakes, tests

(Carnoy & Loeb, 2002; Chiang, 2009; Figlio & Rouse, 2006; Hanushek & Raymond, 2005; Jacob, 2005; Jacob & Lefgren, 2004).

This study complements existing research on pay for performance by testing whether a statewide program of school-based incentives embedded in a long-standing public accountability system has the intended effects on math and reading test score gains, and no unintended effect of narrowing the curriculum at the expense of a non-high-stakes subject, science, and whether the program had the effect of closing or narrowing the test score gaps in these three subjects.

III. Policy Background – ABCs Performance Incentives

The North Carolina accountability program, the ABCs of Public Education, first implemented in the 1996-1997 school year, awarded schools bonuses based on the annual achievement growth of their students from one year to the next. This growth approach to accountability was feasible because the state had been testing all students in grades 3 through 8 annually in math and reading since the early 1990s. The growth formula in 2007 was based on changes in normalized test scores from the mean and standard deviation from the first year a particular test was used in the state. The academic change for an individual student was calculated as the student's normalized score minus the average of two prior-year academic scores, with the average discounted to account for reversion to the mean. If a school raised student achievement by more than was predicted for that school, all the school's teachers, aides, and certified staff received financial bonuses—\$1,500 for achieving high growth and \$750 for meeting expected achievement growth. Schools were publically recognized for their performance, and schools failing to make expected growth were subject to intervention from the state. The program, which applied to all elementary, middle, and high schools in the state (more than 100,000 teachers, 2,000 schools, and 1.4 million students), intended to induce each school

to provide its students with at least a year's worth of learning for a year's worth of education. I examine the effect of failing the expected growth target on 5th graders. I focus on this grade level because: 1) only 5th and 8th graders took the reading, math, and science tests to permit a high-stakes / low-stakes comparison, 2) there were more than twice as many elementary schools as there are middle schools, which increases the power of the design; and, 3) the expected growth target was closer to the middle of the distribution of schools than was the high growth target, which also improves power. In this year, 69 percent of elementary schools met at least their expected growth target and received at least \$750 bonuses. Twenty-two percent of elementary schools also met their high growth target and received an additional \$750 for a total of \$1500 in bonus pay.

IV. Data

The study examines the incentive effects of the 2006-2007 bonus program on all public schools enrolling 5th grade students in North Carolina in the spring of 2008. The study uses administrative records from the North Carolina Department of Public Instruction archived by the North Carolina Education Research Data Center at Duke University. The outcomes are spring 2008 student test scores. The treatment is whether or not the school received a bonus for 2006-2007 test score growth. The unit of analysis of the study is, therefore, student outcomes nested within a school-level treatment. Students are matched to their 2008 school and its 2007 bonus treatment status. In other words, the treatment status follows the school not the student. Students who switch schools between 4th and 5th grade receive their 5th grade school's treatment status, not their former school's treatment status. The full analysis sample size is approximately 95,000 5th

grade students nested within about 1,300 schools.³ Within typical bandwidth estimated by the RD models, the sample typically contains about 50,000-70,000 students in about 700-900 schools. The sample includes students from urban, suburban, small town, and rural communities, and is 56% white, 27% black, 10% Hispanic, 14% academically gifted, 13% with special education needs, 46% eligible for free or reduced priced lunch, and 28% with college educated parents (see table 1).

While there is some dependence across time in school likelihood of failing the bonus target, a school's prior status is by no means a deterministic predictor of failing the next year. Table 2 shows 2007 bonus status as a function of various definitions of prior bonus status. Among schools failing the expected growth target in 2006, about half failed the target in 2007. Meeting the target in a prior year is a stronger predictor of next year failure than is failing the target. Among schools meeting the target in 2006, 84% met the target in 2007. Expanding the definition of prior status to include multiple consecutive years of failure, increases the probability of failing, but does not reduce the probability of meeting the target in 2007 to zero. These data suggest that a school's bonus eligibility can change over time and that some portion of schools are able to implement educational interventions to meet test score targets. A key feature of the policy was releasing school's precise expected growth score to all school district testing coordinators. Anecdotal evidence suggests that testing coordinators shared these data with superintendents and principals. Therefore it is plausible that principals and teachers could use the

³ I begin with 109,827 students and drop 13,607 students (12.4%) who were either retained in 4th grade and would thus not take 5th grade tests or attended a school with no "expected growth" rating, or both. Included in this total of censored students are 34 dropped from two schools with very high expected growth scores, and 86 students in 15 schools which had fewer than ten 5th graders. Of the remaining 96,220 students, 92,254 (95.9%) had a reading gain score, 92,597 (96.2%) had a math gain score, and 93,143 (96.8%) had a science score. Out of a total of 1386 schools with a 5th grade that were not schools for students with disabilities, hospital, or schools in jails, 39 (3%) had no expected growth score.

distance to the prior year bonus target to guide improvement strategies in the following year. Systematic evidence on this point is unfortunately unavailable.

V. Hypotheses and Identification Strategy

This study examines the incentive effects of North Carolina’s practice of awarding performance bonuses on test score achievement. Bonuses were awarded based solely on whether a school exceeds a threshold on a performance metric. The study uses a sharp regression discontinuity design to examine three questions:

Do bonuses induce incentive effects to increase math or reading test score gains?—I hypothesize that students in schools just below the bonus threshold in 2007 will have higher reading and math test score gains on the state’s assessment in year 2008 than students in schools just above the threshold. During the period of the study, both math and reading were “high stakes” in the sense that bonuses were awarded based on these subjects.

Do bonuses promote a narrowing of the curriculum at the expense of science?—In 2008, the initial year of the 5th grade science assessment, teacher bonuses were based on reading and math alone. In other words in 2008, science was a “low stakes” test. I hypothesize that students in schools that barely missed bonus thresholds based on reading and math in 2007 will have lower science test scores in 2008 because schools facing accountability threats will substitute away from science instruction to spend additional time on reading and mathematics.

Do bonuses promote “educational triage” based on the achievement level of the student?—North Carolina’s system creates an incentive to focus on the students with the highest potential for growth. Prior work shows that responses to accountability threats vary by subject matter (Ladd & Lauen, 2010). Reading achievement is considered to be less amenable to instructional intervention than mathematics. When faced with short run accountability threats,

therefore, schools will likely focus interventions on low achieving students in math, the subject with the highest potential benefit. Therefore, I hypothesize that in schools just missing bonus thresholds, mathematics test score gains will be higher for low achieving students than for students with average or high achievement. Furthermore, I hypothesize that in schools just missing bonus thresholds, reading test score gains for low achievers in reading will be approximately equal to reading test score gains for high achievers in reading.

The study employs a regression discontinuity design (RD), which requires that the treatment assignment be either a deterministic or a probabilistic function of a “forcing” variable (Imbens & Lemieux, 2008; Thistlethwaite & Campbell, 1960). The present study uses as a forcing variable the spring 2007 accountability rating that determined eligibility for the expected growth bonus (\$750). This variable, a performance metric with a threshold, was the sole determinant of \$750 bonus payments. Elementary school bonuses in 2007 were based solely on reading and math achievement. Treatment assignment was “sharp” in the sense that expected growth ratings, and thus bonus payments, were legislatively determined based on this performance metric (see figure 1). Because schools were assigned bonuses based on a formula by state assessment officials during the summer following spring testing, there was no opportunity for manipulation of treatment assignment around the cutoff (see figure 2).⁴

I estimate RD models with parametric models and local polynomial regression (Fan & Gijbels, 1996) which imposes no functional form assumptions on the relationship between the forcing variable and the outcome. Rather than fitting a constant function, this approach fits

⁴ The forcing variable made available to me for this study is not without its quirks. Although it was most certainly at one time a continuous variable, it is rounded to the second decimal and has no schools with a value of zero. In addition, there is a slight fall off in density at -.01 and +.01.

smoothed non-parametric functions to the observations within a distance h on either side of the cutoff point, c , on the forcing variable, X (Imbens & Lemieux, 2008):

$$(1) \min_{\alpha_l: \beta_l} \sum_{i: c-h < X_i < c} (Y_i - \alpha_l - \beta_l(X_i - c))^2 \text{ and}$$

$$(2) \min_{\alpha_r: \beta_r} \sum_{i: c \leq X_i < c+h} (Y_i - \alpha_r - \beta_r(X_i - c))^2 .$$

The value of $\widehat{\mu_l(c)}$ and $\widehat{\mu_r(c)}$ are $\widehat{\mu_l(c)} = \widehat{\alpha}_l + \widehat{\beta}_l(X_i - c) = \widehat{\alpha}_l$ and $\widehat{\mu_r(c)} = \widehat{\alpha}_r + \widehat{\beta}_r(X_i - c) = \widehat{\alpha}_r$, where $\widehat{\alpha}_l$ and $\widehat{\alpha}_r$ are the left and right intercepts (i.e., at the cutoff). Therefore, the average treatment effect is $\widehat{\tau}_{SharpRD} = \widehat{\alpha}_r - \widehat{\alpha}_l$ (ibid). I estimate non-parametric regressions with both a triangular kernel, which gives more weight to observations close to the cutoff, and a rectangular kernel, which gives the same weight to all observations regardless of position relative to the cutoff. When this function is estimated with a rectangular kernel, an identical treatment effect can be obtained from β_1 from a linear OLS regression with a difference in slopes to the left and to the right of the cutoff (Lee & Lemieux, 2010):

$$(3) y = \alpha + \beta_1 D + \beta_2(X - c) + \beta_3 D(X - c) + \epsilon, \text{ where } c - h \leq X \leq c + h.$$

To correct for the non-independence of student observations within schools, I report cluster robust standard errors from parametric models and cluster bootstrapped standard errors from non-parametric models. The cluster bootstrap preserves within-cluster correlation by randomly drawing clusters and including all within-cluster observations from each selected cluster in each bootstrap replication.⁵ To address the bias-efficiency tradeoff in RD designs—the wider the bandwidth the smaller the standard error and the larger the bias—the study uses an approach proposed by Imbens & Kalyanaraman (2009) for optimal bandwidth selection. I report

⁵ Lee and Card (2008) recommend clustering on the assignment in cases of discrete assignment variables. This results in slightly smaller standard errors than clustering on school. In the interest of avoiding type I errors, I therefore cluster on school rather than bins of the forcing variable.

effects at the optimal bandwidth, h^* , and also at a variety of bandwidths both below and above the optimal bandwidth.

For the purpose of examining whether the bonus program has differential effects on students based on prior achievement, I test for moderator effects with a parametric RD model. I define students as low, average, or high in reading prior achievement based on their position in the spring 2007 *within-school* reading pre-test (i.e., 4th grade) score distribution. Students defined as low fall below -.5 SD below the within-school pretest mean, those defined as medium fall within the range of -.5 SD and +.5 SD, and those defined as high fall above +.5 SD. The same procedure is used to define student prior achievement in math. Because students did not take a science test in 4th grade, there is no pretest in this subject. In the 5th grade science moderator models, I use the student's 4th grade reading pretest as a proxy for 4th grade science pretest (5th grade reading and science test scores correlate at $r=.78$). I use the within-school rather than the statewide pretest distribution as a basis for these indicators because resource distribution is most likely made on the basis of the distributions teachers and principals face within their particular schools rather than on how their students perform relative to statewide averages. I estimate a fully saturated moderator model to statistically test the difference between prior achievement groups:

$$(4) \ y = \alpha + \beta_1 D + \beta_2 (X - c) + \beta_3 D(X - c) + \beta_4 Low + \beta_5 High + \beta_6 D(Low) + \beta_7 D(High) + \beta_8 (X - c)Low + \beta_9 (X - c)High + \beta_{10} D(X - c)Low + \beta_{11} D(X - c)High + \epsilon, \text{ where } c - h \leq X \leq c + h.$$

VI. Results

A. Average treatment effects

Figure 3 shows the distribution of the expected growth score expressed as both counts of students and schools within each bin of the forcing variable, expected growth. The most comparable schools on either side of the cutoff are those at $-.01$ and $+.01$ (recall that there are no schools at exactly 0). Due to the statewide nature of this sample and the fact that the forcing variable has been rounded to relatively large bins, there are 2,716 students in 35 schools at $-.01$ on expected growth and 3,124 students in 40 schools at $+.01$ on expected growth. If assignment to treatment is as good as random for schools this close to the cutoff, the causal effect of failing to get a bonus on student test score is the mean difference in test scores among students in schools at these two points. At a bandwidth of $.01$ ($-.01$ versus $+.01$, total N of 75 schools and 5,836 students), the mean difference in 2008 standardized reading gain is $.089$ SD ($.046$), which is marginally significant ($p=.055$), with the positive value favoring students in schools that failed to receive bonuses in 2007. This suggests that failing the bonus target produced an incentive effect in the schools that barely missed the threshold the prior year. The effect in standardized math gain score at the same bandwidth is larger, but not statistically distinguishable from zero at conventional levels, $.128$ SD ($.088$), $p=.149$. The effect on standardized science score level, on the other hand, is negative and quite close to zero $-.016$ SD ($.090$), $p=.860$. While these results have the appeal of being exactly at the cutoff, they provide no information about the functional form of the relationship between the forcing variable and the outcomes on either side of the cutoff. Therefore, it is difficult to assess the extent to which the counterfactual—a continuation of a trend—is likely to hold. The counterfactual for no bonus schools is the extrapolation of a trend based on data points to the left of figure 3 into the area to the right of the cutoff. The

counterfactual for bonus schools is the extrapolation of a trend based on data points to the right of figure 3 into the area to the left of the cutoff. Without extending the bandwidth and making assumptions about functional form, it is difficult to assess these counterfactual conditions.

Graphs showing the full range of outcomes with a non-parametric smoothed function, bin average scores, and treatment effects on standardized reading gain, math gain, and science level are shown in figures 4-6.⁶ On the Y axis is the 2008 student test score and on the X axis is the school's 2007 expected growth score. The optimal bandwidth (Imbens & Kalyanaraman, 2009) in reading gain and science is .12 on either side of the cutoff (N=70,963 students and 902 schools) and .10 in math (N=57,990 students and 738 schools). At extreme values of the expected growth score, there is evidence of a non-linear relationship, but within bandwidths of .10 or .12, however, the relationship appears to be quite linear. The treatment effect reported in the figures is estimated with a rectangular kernel and cluster bootstrapped standard errors. (Treatment effects from models with a triangular kernel are shown in table 3, discussed below.) The graphs show a discontinuity in both reading and math gains, but no discontinuity in science level. Recall that I hypothesized that failing to reach the bonus target in the prior year would induce an incentive effect on test scores in the following year. So the treatment effects reported below are for those failing (not meeting) the bonus target. Also recall that in 2007, North Carolina schools were held accountable for reading and math gains and not science test score levels. The treatment effect of being in a school that failed the bonus target on reading gain is .074 SD, and is statistically significant at conventional levels. The treatment effect on math is .062 SD, but is imprecisely estimated. In science the treatment effect is .017 SD and also non-significant. These findings suggest that schools that fail bonus targets in the prior year implement

⁶ Non-parametric estimates and graphs produced in Stata with `rd`, a free user-produced program (A. Nichols, 2007).

educational interventions that cause higher achievement gains in reading and at least have no detrimental effect on science. That reading interventions do not spillover to create a similar positive effect on science is somewhat surprising given the high correlation of test score achievement in reading and science in 5th grade.

As shown in table 3, these results are robust to alternative specifications. For example, in column 1 a non-parametric model with a triangular kernel instead of a rectangular kernel produces a treatment effect on reading gain of .076 instead of .074. Using a parametric specification (see equation 3, above) produces exactly the same treatment effects as the non-parametric model with the rectangular kernel, although the standard errors differ due to the cluster robust, rather than cluster bootstrap, standard error correction. The treatment effects on math gains from the non-parametric model with a rectangular kernel and from a parametric model, for example, are both .062 in column 3, but the standard error differs by .002 points (.062 versus .060). Estimating models based on school averages rather than individual student scores does not change the substantive thrust of the study's findings significantly, although the size of the effect on reading decreases somewhat, the size of the effect on math increases somewhat, and the effect on science flips signs (compare columns 2 to 1, 4 to 3 and 6 to 5). Standard errors increase in the reading models and decrease in the math and science models. The reading effects are significant at the .10 level, whereas the math and science effects are not significant at conventional levels.

At bandwidths larger than .10, these treatment effects estimated from a linear parametric model with a difference in slopes (see equation 3) are robust across varying bandwidths (see table 4). Narrower bandwidths produce inconsistent evidence about treatment effects in all three subjects, with effects larger in reading, but not always significantly different from zero; effects

larger in math, and usually not significant; and widely divergent, but always insignificant, effects in science. At bandwidths from .10 to .24, a consistent story emerges: effects on reading range from .04 SD to .07 SD and are always significant, effects on math are .05 SD to .09 SD and are only significant at very wide bandwidths (.22 to .24), and effects on science are essentially zero and are never significant.

Another approach to exploring the validity of functional form assumptions is to fit a parametric specification with polynomial terms. Estimating models with squared and cubed terms, and all the associated interaction terms with the treatment dummy, produces larger treatment effects on reading gain and no consistent evidence of precisely estimated treatment effects on math gain or science test score levels (see table 5). Along the rows of panel A. in table 5 are specification types defined by the order of the polynomial. A polynomial of order 2 (includes squared terms) and 3 (includes squared and cubed terms) are defined by equations 5 and 6, below:

$$(5) \ y = \alpha + \beta_1 D + \beta_2 (X - c) + \beta_3 D(X - c) + \beta_3 (X - c)^2 + \beta_4 D(X - c)^2 + \epsilon.$$

$$(6) \ y = \alpha + \beta_1 D + \beta_2 (X - c) + \beta_3 D(X - c) + \beta_3 (X - c)^2 + \beta_4 D(X - c)^2 + \beta_5 (X - c)^3 + \beta_6 D(X - c)^3 + \epsilon.$$

At wider bandwidths (.12 and .10), the equation with a polynomial order of 3 produces significant treatment effects on reading gain of .189 and .231. A more complex parametric specification permits the curve to the left of the cutoff to adapt to points closer to the cutoff (see figure 7). At bandwidths of .08 and .06, a specification with squared terms produces significant estimates on reading gain of about the same size: .175 and .223, respectively. As shown in table 4, at a bandwidth of .04, a linear specification with a difference in slope (polynomial order of 1, see equation 3, above), produces a significant treatment effect on reading gain of .156. In sum,

more complex parametric specifications allow points closer to the cutoff to carry more weight in the estimate by allowing the curve to more flexibly adapt to the data points along the forcing variable. This approach results in a larger treatment effect than the one reported in table 3 and figure 4. Therefore, depending on the bandwidth and assumptions about functional form, the treatment effect of failing to meet the bonus target on reading gain lies between .07 and .23 SD.

While estimating a more complex parametric specification produces larger treatment effects on reading gain, these specifications produce little to no evidence of consistent treatment effects on math gain or science levels. Despite some anomalous results that are not robust across bandwidths, the overall impression of the coefficients in panel B in table 5 is positive, but imprecisely estimated treatment effects on math gain, and in panel C treatment effects on science that have varying signs and are often close to zero.

B. Differential treatment effects

The average treatment effects reported thus far could be misleading if effects differ substantially by student position in the test score distribution. A zero average effect could be the result of a positive effect for low achievers and a negative effect for high achievers, for example. A positive effect could emerge because average and high achievers have gains, while low achievers show no gains. To more fully understand how incentive effects operate and whether public teacher bonuses work to close or widen test score gaps, I examine whether incentives have differential effects.

I estimate moderator models shown in equation 4, above, at a variety of bandwidths, but find that relatively wide bandwidths are required to gain enough precision to estimate consistent effects. I report results here for each subject at a bandwidth of .20 (N=89,711 students in 1160 schools). A different pattern of effects emerges for each subject. In contradiction to recent

research on status-based systems such as No Child Left Behind, the North Carolina system based on bonuses for test score growth, does not produce outsized reading gains for students in the middle of the within-school distribution; in fact low and high achieving students have significantly higher incentive-induced gains than average students (figure 8; the tables used to produce this figure are in the appendix in tables A1-A3). Specifically, low achieving students in schools that failed the bonus target have reading gains .08 SD higher than low achieving students in schools that met the bonus target. High achieving students in no bonus schools have reading gains that are .06 SD higher than high achieving students in schools that met the bonus target. In short, bonus-induced gains are not offsetting. Students with average prior achievement gain little from the bonus program, and while pressure from the bonus program increases test scores for low achieving students, it is probably not working to close the test score gap overall.

A different pattern emerges in the math test score gain model. In this model, bonuses have positive effects on low and average achievers and essentially no effect on high achievers. Low achieving students have significantly larger gains than high achieving students (.10 SD versus .01 SD). Average achieving students also have larger gains than high achieving students (.07 SD versus .01 SD). Therefore, the bonus program is probably contributing to closing test score gaps in math.

In science, the effects of the no bonus incentive are much smaller, and yet another pattern emerges. While the average treatment effect on science is small and negative, this masks an important difference: low achieving students do slightly worse than average and high achieving students in schools that failed the bonus threshold ($p=.091$). The treatment effect on science for students with low prior achievement in reading is $-.02$ SD, whereas the treatment effects on science scores for students with average and high reading achievement are $.02$ SD and $.04$ SD,

respectively. In other words, the bonus incentives appear to narrow the curriculum at the expense of science for those students with the largest deficit in reading, a subject within the accountability framework.

Estimates from these moderator models of varying bandwidths (not shown, but available from author upon request) show that differential effects reported above in reading and math gains are relatively consistent across bandwidths ranging from .18 to .24, whereas the science differences are found in a somewhat narrower range of bandwidths (from .16 to .20). The fact that these moderator models require wider bandwidths to obtain precise estimates raises concerns about the comparability of the samples of schools on either side of the cutoff. Therefore, these estimates provide suggestive, but perhaps not conclusive, evidence of causal effects on students at various points in the achievement distribution.

C. Specification checks

This study posits a discontinuous jump in the treatment effect at the cutoff (expected growth=0). Although it is not a violation of the RD assumption of a discontinuity at the actual cutoff to have discontinuities at other points of the forcing variable, it is certainly reassuring when the evidence suggests there are none. I test for placebo discontinuities at the median below the cutoff (expected growth=-.05) and the median above zero (expected growth=+.09). I use a parametric specification (shown in equation 3) with cluster robust standard errors at bandwidths of .01 to .04 for the lower median test and .01 to .09 for the upper median test (the range between 0 and the upper and lower medians differ because the cutoff is not centered on middle of the distribution). To avoid confounding a placebo effect with the discontinuities at the actual cutoff, in the lower median test, I exclude observations above 0 (the actual cutoff); in the upper median test, I exclude observations below 0. In a total of 39 tests (three dependent variables, two placebo

points, thirteen bandwidths), I find no placebo treatment effect that is statistically significant at the .10 level or below (results from author upon request).

I also tested for discontinuities at the actual cutoff (expected growth=0) in school level baseline covariates that if they were also discontinuous at the cutoff could confound the treatment effect. I examined school 4th grade average reading and math test scores, percentage black students, percentage poor students (as measured by free and reduced lunch eligibility), percent novice teachers, and teacher turnover rate. I found no convincing evidence of discontinuities in any of these pre-treatment school-level variables. Out of a total of 144 tests, I reject the null of no treatment at the .10 level effect six times, for a rejection rate of .0417 (results from author upon request).

VII. Policy Implications and Conclusions

This study complements existing research on teacher pay for performance by examining the effects of a mature statewide program of school-based incentives embedded in a public accountability system. It is the first conducted on a recent statewide program at a time when state and district policymakers are implementing teacher incentive programs as part of the federal Race to the Top initiative. This work complements randomized controlled trials conducted in the U.S. that have examined whether small samples of teachers will change their behavior if offered the possibility of a bonus if their performance exceeds a threshold. In contrast to these pilot studies, the present study combines the strong internal validity of an RD design with the external validity of a long-standing program with a ten year history of paying out bonuses based on a relatively transparent performance indicator. This study asks whether teachers will respond to failing to meet a bonus target in the prior year by exerting more effort and implementing educational interventions to improve achievement gain in the following year.

In contradiction to theories and evidence that group-based incentives would have null effects due to free riding, this study provides some evidence that school wide incentives can increase test scores. In this sense, this work corresponds with two studies of the effects of group-based performance pay in Israel (Lavy, 2002) and India (Muralidharan & Sundararaman, 2009) and differs from recent work on team-based incentives from Texas (McCaffrey, et al., 2011). The positive effects of the bonus program could be offset if teachers face a multitasking problem in which other important aspects of teaching and learning are neglected in favor of those for which they are held accountable. I find mixed evidence of the multitasking problem on science test scores, a subject taught by 5th grade teachers in self-contained classrooms which was not part of the accountability system in 2008. Examining the effects of failing to meet the bonus target on a “low-stakes” test is a reasonable test of the multitasking problem because unlike middle school or high school teachers, 5th grade teachers have some flexibility in the time they allot to reading, math, science, and social studies. On average, no negative effects emerge in science test score levels, though small negative effects surface for low achieving students in schools that failed to meet the bonus threshold the prior year. This study tests for differential effects by student prior achievement to examine whether pressure to meet bonus targets has the potential for closing test score gaps between low and high achieving students. In reading, the effect of failing the bonus target has modest effects on low and high achieving students and virtually no effect on students in the middle of the within-school distribution. In math, on the other hand, the treatment effect has stronger effects on low and average students than high achieving students. Therefore, teacher pay for performance does not appear to work to close test score gaps in reading and may help close test score gaps in math. In summary, this study provides more evidence that teachers and schools respond to public accountability programs with interventions that increase test scores.

Establishing the short run effects of performance pay on student test scores in high and low stakes tests is an important first step in contributing to the research base on this type of human capital development policy. In light of the fade out effects found in Glewwe, et al (2010), important next steps include examining medium term effects on test scores to determine whether gains were the result of test preparation or more lasting educational interventions, investigating the effects of other important outcomes such as teacher turnover, and weighing the costs of the program against the benefits. Due to the statewide nature and retrospective nature of this study, I have no data on changes in classroom practices and therefore cannot explain why and how effects emerge. Future, prospectively designed, studies should include surveys, interviews, and observations of teachers to examine how teachers respond to performance incentives. I posit two mechanisms that could explain how performance pay could alter behavior: seeking to maximize income and aiming to avoid accountability threats. This study cannot tease apart these two explanations because the treatment testing in this study combines both mechanisms together: failing to attain the bonus is a publicly recognized status that deprives a teacher of additional income. This study hypothesizes that the combination of the two mechanisms can have positive effects. Future research should seek to determine whether public recognition is a complement or substitute for additional income.

References

- Amrien, A. L., & Berliner, D. C. (2002). High-Stakes Testing, Uncertainty, and Student Learning. *Education Policy Analysis Archives*, 10(18), Retrieved 12/23/08 from <http://epaa.asu.edu/epaa/v10n18/>.
- Booher-Jennings, J. (2005). Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, 42(2), 231-268.
- Bryk, A. S., & Schneider, B. L. (2002). *Trust in schools: A core resource for improvement*: Russell Sage Foundation Publications.
- Campbell, D. T. (1979). Assessing the Impact of Planned Social Change. *Evaluation and Program Planning*, 2, 67-90.
- Carnoy, M., & Loeb, S. (2002). Does External Accountability Affect Student Outcomes? A Cross-State Analysis. *Educational Evaluation and Policy Analysis*, 24(4), 305-331.
- Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93(9-10), 1045-1057.
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. J., Weinfeld, F. D., et al. (1966). *Equality of Educational Opportunity*. Washington: USGPO.
- Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record*, 106(6), 1047-1085.
- Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *The Journal of Human Resources*, 37(4), 696-727.
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications* (1st ed.). London ; New York: Chapman & Hall.
- Figlio, D. (2006). Testing, Crime and Punishment. *Journal of Public Economics*, 90(4), 837-851.

- Figlio, D., & Rouse, C. (2006). Do Accountability and Voucher Threats Improve Low-Performing Schools? *Journal of Public Economics*, 90(1-2), 239-255.
- Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher Incentives. *American Economic Journal: Applied Economics*, 2(3), 205-227.
- Goldhaber, D. (2008). Teachers matter, but effective teacher quality policies are elusive. *Handbook of research in education finance and policy*, 146–165.
- Hannaway, J., & Hamilton, L. (2008). *Performance-Based Accountability Policies: Implications for School and Classroom Practices*. The Urban Institute and RAND Corporation. Washington, DC.
- Hanushek, E. (1992). The trade-off between child quantity and quality. *The Journal of Political Economy*, 100(1), 84-117.
- Hanushek, E., & Raymond, M. A. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297-327.
- Heckman, J. J., Heinrich, C., & Smith, J. (2002). The performance of performance standards. *Journal of Human Resources*, 37(4), 778-811.
- Heilig, J. V., & Darling-Hammond, L. (2008). Accountability Texas-style: The progress and learning of urban minority students in a high-stakes testing context. *Educational Evaluation and Policy Analysis*, 30(2), 75-110.
- Imbens, G. W., & Kalyanaraman, K. (2009). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *NBER Working Papers*, 14726.
- Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2), 615-635.

- Jacob, B. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5-6), 761.
- Jacob, B., & Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, 86(1), 226-244.
- Jencks, C. (1972). Inequality: A Reassessment of the Effect of Family and Schooling in America.
- Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What Do Test Scores in Texas Tell Us? *Education Policy Analysis Archives*, 8(49).
- Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *The Journal of Human Resources*, 37(4), 752-777.
- Koretz, D. M., & Barron, S. (1998). The Validity of Gains in Scores on the Kentucky Instructional Results Information Systems (KIRIS). Santa Monica, CA: RAND Corporation.
- Ladd, H. F., & Lauen, D. (2010). Status versus Growth: The Distributional Effects of School Accountability Policies. *Journal of Policy Analysis and Management*, 29(3), 426-450.
- Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly*, 38(4), 494-529.
- Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *The Journal of Political Economy*, 110(6), 1286-1317.
- Lavy, V. (2009). Performance Pay and Teachers' Effort, Productivity, and Grading Ethics. *American Economic Review*, 99(5), 1979-2011.
- Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2), 655-674.

- Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48(2), 281-355.
- Linn, R. L. (2000). Assessments and Accountability. *Educational Researcher*, 29(2), 4-16.
- McCaffrey, D., Pane, J., Springer, M. G., Burns, S., & Haas, A. (2011). *Team Pay for Performance: Experimental Evidence from Round Rock's Project on Incentives in Teaching*. Paper presented at Society of Research on Educational Effectiveness, Washington, DC, March 4, 2011.
- Muralidharan, K., & Sundararaman, V. (2009). Teacher performance pay: Experimental evidence from India. *NBER Working Papers*, 15323.
- Neal, D., & Schanzenbach, D. W. (2010). Left Behind By Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics*, 92(2), 263-283.
- Nichols, A. (2007). rd: Stata module for regression discontinuity estimation:
<http://ideas.repec.org/c/boc/bocode/s456888.html>.
- Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage : how high-stakes testing corrupts America's schools*. Cambridge, Mass.: Harvard Education Press.
- Orfield, G., & Kornhaber, M. L. (2001). *Raising standards or raising barriers? : inequality and high-stakes testing in public education*. New York: Century Foundation Press.
- Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management*, 26(4), 909-950.
- Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.

- Sanders, W. L., & Rivers, J. C. (1996). Cumulative and residual effects of teachers on future student academic achievement: Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.
- Springer, M. G., Hamilton, L., McCaffrey, D. F., Ballou, D., Le, V. N., Pepper, M., et al. (2010). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives, Vanderbilt University.
- Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-Discontinuity Analysis - an Alternative to the Ex-Post-Facto Experiment. *Journal of Educational Psychology*, 51(6), 309-317.
- Valenzuela, A. (2005). *Leaving children behind : how "Texas-style" accountability fails Latino youth*. Albany: State University of New York Press.
- Weitz, K., & Rosenbaum, J. (2007). Inside the Black Box of Accountability: How High Stakes Accountability Alters School Culture and the Classification and Treatment of Students and Teachers. In A. S. e. al. (Ed.), *No Child Left Behind and the Reduction of the Achievement Gap* (pp. 97-116). New York and London: Routledge.

Figures

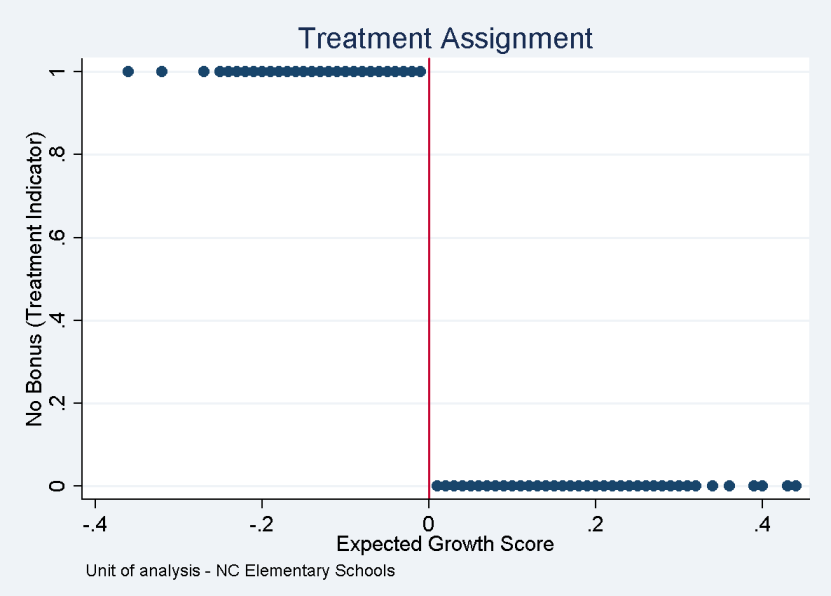


Figure 1. Treatment assignment as a function of the forcing variable, expected growth score.

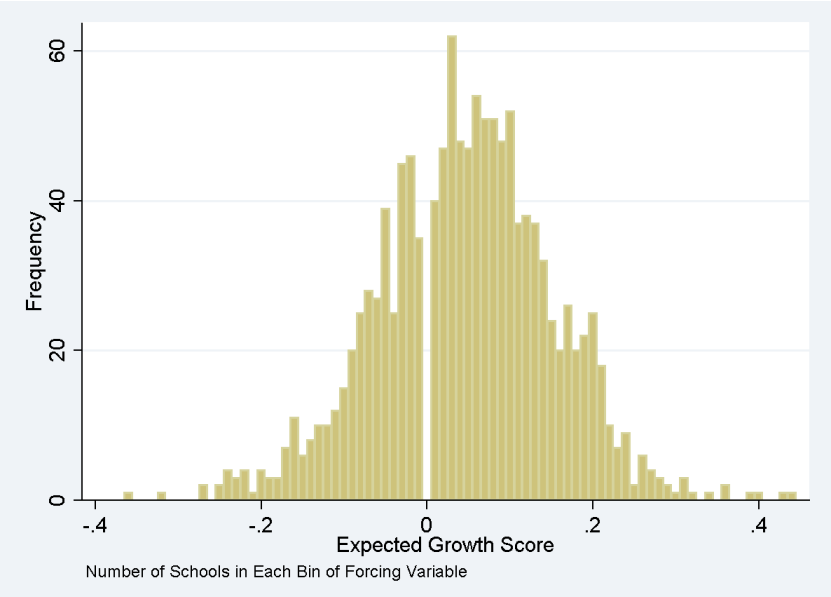


Figure 2. Density of schools at each bin of the forcing variable, expected growth score.

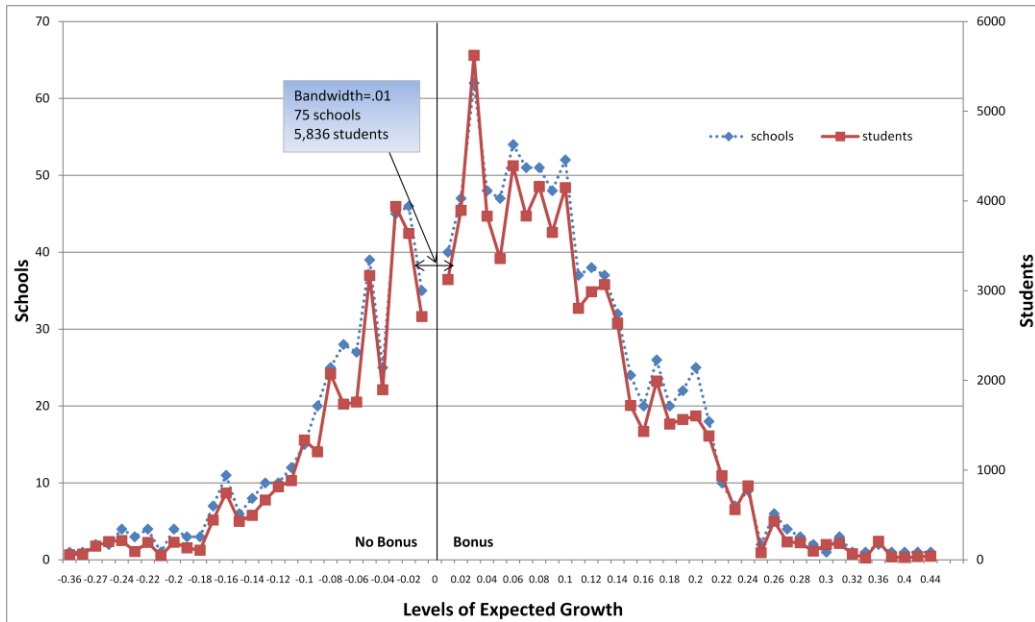


Figure 3. Density of schools and students at each level of the forcing variable, expected growth score and number of students and schools at bandwidth of .01.

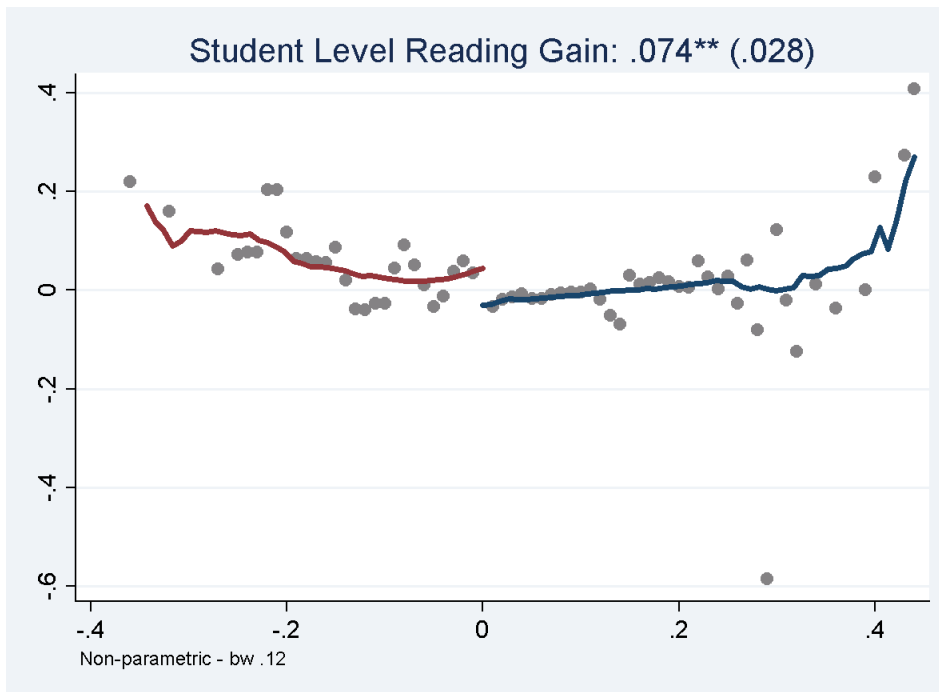


Figure 4. Treatment effect on '07-'08, 4th to 5th grade reading gain. Non-parametric RD model, rectangular kernel. Cluster bootstrapped SE. X-axis is 2007 growth score. Y in standardized units. Effect of failing to get a bonus: .074 (.028)**.

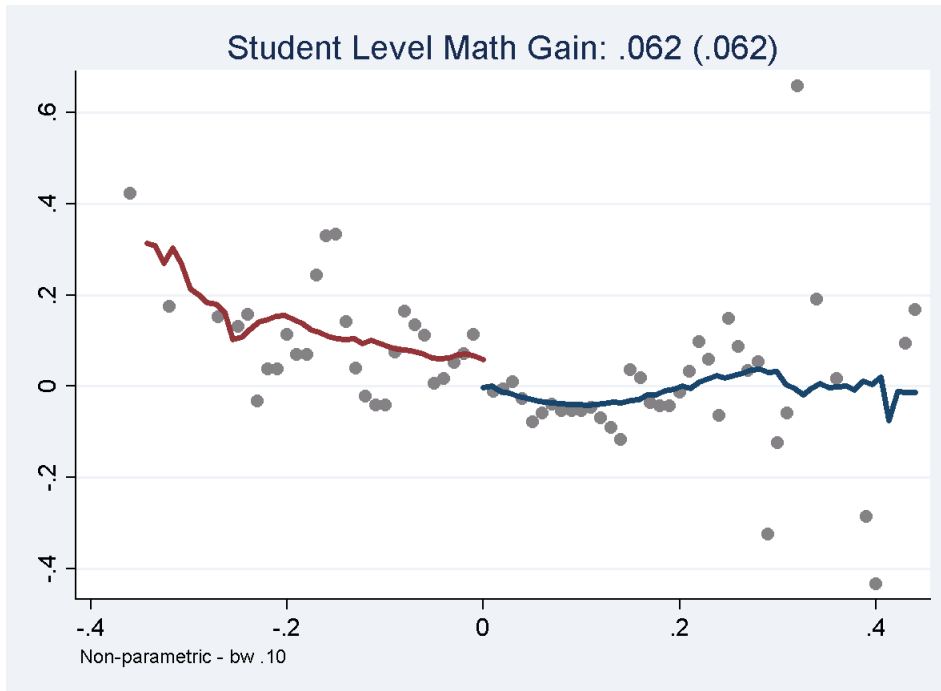


Figure 5. Treatment effect on '07-'08, 4th to 5th grade math gain. Non-parametric RD model, rectangular kernel. Cluster bootstrapped SE. X-axis is 2007 growth score. Y in standardized units. Effect of failing to get a bonus: .062 (.062).

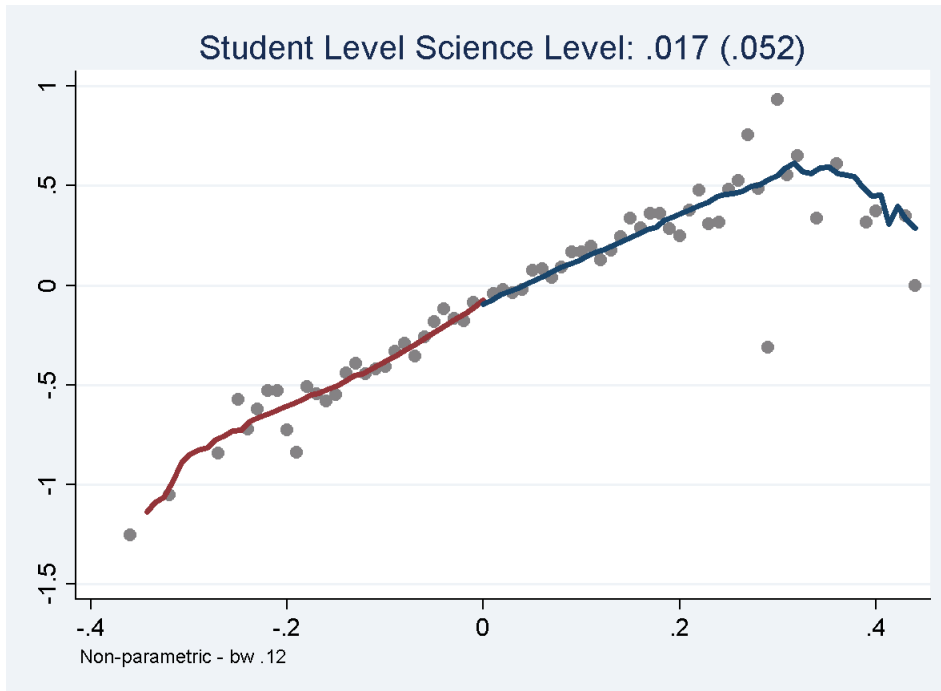


Figure 6. Treatment effect on '08, 5th grade science scale score. Non-parametric RD model, rectangular kernel. Cluster bootstrapped SE. X-axis is 2007 growth score. Y in standardized units. Effect of failing to get a bonus: .017 (.052).



Figure 7. Treatment effect on '07-'08, 4th to 5th grade reading gain. Parametric RD model, polynomial order 3, bandwidth .12. Cluster robust standard error. X-axis is 2007 growth score. Y in standardized units. Effect of failing to get a bonus: .189 (.078)*.

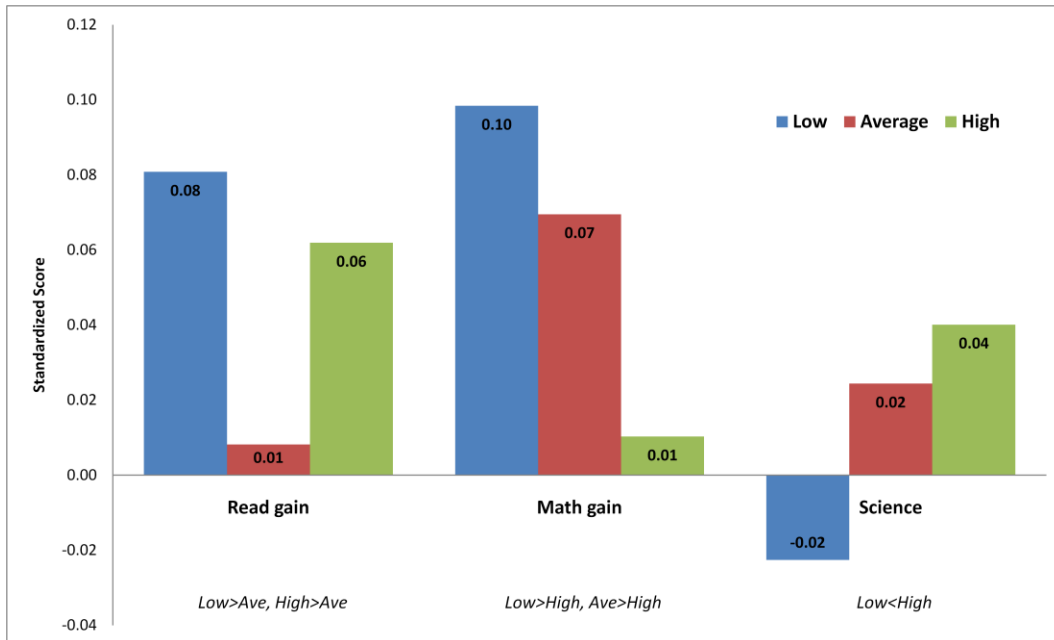


Figure 8. Differential effects of failing to meet the bonus target by student prior achievement.

Parametric models with cluster robust standard errors. Bandwidth=.20 for all models. Notes

below bars note which categories are statistically distinguishable from each other. Prior

achievement defined as position in the within-school 4th grade achievement distribution (see text

for details).

Tables

Table 1. Descriptive Statistics

Name	Description	N	Mean	SD	Min	Max
Dependent Variables						
zrgain	Standardized '07-'08, 4th to 5th grade reading gain	92254	0.00	1.00	-5.18	5.54
zmgain	Standardized '07-'08, 4th to 5th grade math gain	92597	0.00	1.00	-5.57	5.36
zsciscal2008	Standardized '08, 5th grade science level	93143	0.01	1	-3.19	3.07
Treatment Variables						
nobonus	Treatment dummy (Sch failed to meet bonus target)	96220	0.31	0.46	0.00	1.00
avegrow	Forcing variable (expected growth score)	96220	0.05	0.10	-0.36	0.44
Position in within School Achievement Distribution						
lowread07	Low 4th grade reading score	92836	0.30	0.46	0.00	1.00
highread07	High 4th grade reading score	92836	0.33	0.47	0.00	1.00
medread07	Med 4th grade reading score	92836	0.38	0.48	0.00	1.00
lowmath07	Low 4th grade math score	93248	0.31	0.46	0.00	1.00
highmath07	High 4th grade math score	93248	0.32	0.47	0.00	1.00
medmath07	Med 4th grade math score	93248	0.37	0.48	0.00	1.00
Interactions						
nobonusXavegrow		96220	-0.02	0.04	-0.36	0.00
ncbonusXlowread07		92836	0.09	0.29	0.00	1.00
ncbonusXhighread07		92836	0.10	0.30	0.00	1.00
ncbonusXlowmath07		93248	0.10	0.30	0.00	1.00
nxbonusXhighmath07		93248	0.10	0.30	0.00	1.00
avegrowXlowread		92836	0.01	0.06	-0.36	0.44
avegrowXhighread		92836	0.02	0.06	-0.36	0.44
avegrowXlowmath		93248	0.01	0.06	-0.36	0.44
avegrowXhighmath		93248	0.02	0.06	-0.36	0.44
nobonusXavegrowXlowread		92836	-0.01	0.03	-0.36	0
nobonusXavegrowXhighread		92836	-0.01	0.03	-0.36	0
nobonusXavegrowXlowmath		93248	-0.01	0.03	-0.36	0
nobonusXavegrowXhighmath		93248	-0.01	0.03	-0.36	0
Sample characteristics						
white	White student	96220	0.56	0.60	0.00	1.00
black	Black student	96220	0.27	0.44	0.00	1.00
hisp	Hispanic student	96220	0.10	0.30	0.00	1.00
gifted	Gifted student	96220	0.14	0.35	0.00	1.00
specialed	Student received special education services	91508	0.13	0.33	0.00	1.00
FRL	Student receives free or reduced price lunch	88088	0.46	0.50	0.00	1.00
bachormore	Parent has a bachelor's degree or greater	84871	0.28	0.45	0.00	1.00
Nsch	Total number of students in school	96220	96.93	47.49	10.00	362.00

Note: Low scores are defined as < -0.5 SD below the within-school mean, high as > +0.5 SD above the within-school mean, and med between -0.5 and 0.5 SD

Table 2. 2007 School Bonus Status by Prior Bonus Status

	2007 Bonus Status		
	Met	Failed	N
Prior Bonus Status			
Failed in 2006	50.7	49.3	1218
Met in 2006	84.4	15.6	1218
Failed in 2005 & 2006	43.4	56.7	1191
Met in 2005 & 2006	87.6	12.4	1191
Failed in 2004 & 2005 & 2006	36.0	64.0	1170
Met in 2004 & 2005 & 2006	88.6	11.4	1170
Failed in all years 2001-2006 ¹	--	--	--
Met in all years 2001-2006	91.0	9.0	1106

Note: Unit of analysis is elementary schools.

¹ No schools failed in all years between 2001 and 2006, inclusive.

Table 3. Summary of Main Treatment Effects from Parametric and Non-Parametric RD Models

	Reading Gain		Math Gain		Science Scale	
	1	2	3	4	5	6
	Student	School	Student	School	Student	School
Non Parametric	0.074**	0.061+	0.062	0.078	0.017	-0.026
Rectangle	(0.028)	(0.032)	(0.062)	(0.054)	(0.052)	(0.048)
Non Parametric	0.076*	0.074+	0.054	0.064	0.005	-0.012
Triangle	(0.030)	(0.038)	(0.067)	(0.060)	(0.056)	(0.050)
Parametric	0.074**	0.061+	0.062	0.078	0.017	-0.026
	(0.029)	(0.034)	(0.060)	(0.051)	(0.057)	(0.052)
Obs	68091	854	61098	949	68724	989
Bandwidth	0.12	0.11	0.10	0.13	0.12	0.15

Note: Columns 1, 3, and 5 are estimated with student-level data. Columns 2, 4, and 6 are estimated on school-level averages. Bandwidths chosen by Imbens & Kalyanaraman (2009) optimal bandwidth selection algorithm. Non-parametric regression standard errors are estimated with the cluster bootstrap. Parametric regression standard errors are estimated with the cluster robust standard error. + p<=.10, * p<=.05, ** p<=.01, *** p<=.001

Table 4. Average treatment effects of failing the bonus target across multiple bandwidths

	Reading Gain	Math Gain	Science Level
Bandwidth:			
0.02	0.0821	0.204	0.1960
0.04	0.156**	0.196 ⁺	-0.0943
0.06	0.0877*	0.0165	0.0346
0.08	0.0559	0.0248	-0.0136
0.10	0.0700*	0.0616	0.0097
0.12	0.0743*	0.0843	0.0168
0.14	0.0554*	0.0659	0.0028
0.16	0.0531*	0.0523	0.0163
0.18	0.0531*	0.0621	0.0186
0.20	0.0477*	0.0609	0.0081
0.22	0.0430 ⁺	0.0781 ⁺	-0.0006
0.24	0.0437 ⁺	0.0829*	-0.0105

Note: treatment effects estimated with a parametric regression (see eq. 3) and robust standard errors. + $p \leq .10$, * $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Table 5. Main Treatment Effects, polynomial specifications.

		Bandwidth				
		0.12	0.10	0.08	0.06	0.04
A. Reading Gain						
	Order of Polynomial					
	2	0.0792+ (0.0476)	0.101+ (0.0570)	0.175** (0.0636)	0.223** (0.0814)	0.0761 (0.123)
	3	0.189* (0.0782)	0.231* (0.0914)	0.177 (0.111)	-0.0309 (0.159)	-0.103 (0.361)
	N	68091	60881	50970	39597	27413
B. Math Gain						
	Order of Polynomial					
	2	0.0203 (0.0870)	0.0480 (0.0979)	0.125 (0.116)	0.312* (0.153)	0.0827 (0.235)
	3	0.150 (0.144)	0.233 (0.166)	0.338 (0.208)	0.141 (0.296)	0.564 (0.646)
	N	68325	61098	51153	39732	27497
C. Science Level						
	Order of Polynomial					
	2	-0.0144 (0.0946)	-0.0290 (0.105)	0.00818 (0.125)	-0.204 (0.164)	0.116 (0.251)
	3	-0.0781 (0.153)	-0.0573 (0.179)	-0.202 (0.230)	0.375 (0.323)	1.155+ (0.669)
	N	68724	61460	51470	39952	27643

Notes: All models allow for difference in slope; models with polynomial order of 3 include squared and cubed terms and associated interactions; models with polynomial order of 2 include squared terms and associated interactions; Cluster robust standard errors in parentheses; + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001.

Appendix Tables

Table A1. RD Models Predicting Standardized 5th Grader Reading Gain, .20 Bandwidth					
	(1) All	(2) All	(3) Low	(4) Med	(4) High
nobonus	0.0477*	0.00816	0.0808*	0.00816	0.0619*
	(0.0236)	(0.0274)	(0.0379)	(0.0274)	(0.0273)
avegrow	0.188	0.402*	-0.135	0.402*	0.314 ⁺
	(0.1310)	(0.1590)	(0.2170)	(0.1590)	(0.1640)
nobonusXavegrow	-0.205	-0.226	-0.149	-0.226	-0.103
	(0.2900)	(0.3450)	(0.4270)	(0.3450)	(0.3350)
nobonusXlowread07		0.0726*			
		(0.0358)			
nobonusXhighread07		0.0538 ⁺			
		(0.0290)			
highread07		-0.367***			
		(0.0185)			
lowread07		0.522***			
		(0.0240)			
avegrowXlowread		-0.537*			
		(0.2280)			
avegrowXhighread		-0.088			
		(0.1820)			
nobonusXavegrowXlowread		0.0769			
		(0.3820)			
nobonusXavegrowXhighread		0.123			
		(0.3680)			
_cons	-0.0275 ⁺	-0.0611***	0.460***	-0.0611***	-0.428***
	(0.0143)	(0.0170)	(0.0233)	(0.0170)	(0.0175)
<i>N</i>	86013	86013	25337	32436	28240
<i>R</i> ²	0	0.122	0.003	0.001	0
Cluster robust standard errors in parentheses (clustered on the school); + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.					
Lowread is an indicator for 4th grade reading score -0.5 SD or below the within-school average					
Highread is an indicator for 4th grade reading score $+0.5$ SD or above the within-school average					

Table A2. RD Models Predicting Standardized 5th Grader Math Gain, .20 Bandwidth

	(1) All	(2) All	(3) Low	(4) Med	(4) High
nobonus	0.0609 (0.0426)	0.0695 (0.0466)	0.0984 ⁺ (0.0504)	0.0695 (0.0466)	0.0103 (0.0436)
avegrow	-0.184 (0.2390)	0.102 (0.2510)	-0.586 ⁺ (0.3050)	0.102 (0.2510)	-0.000674 (0.2510)
nobonusXavegrow	-0.342 (0.5260)	-0.728 (0.5550)	0.14 (0.6010)	-0.728 (0.5550)	-0.228 (0.5640)
nobonusXlowmath07		0.0290 ^a (0.0351)			
nobonusXhighmath07		-0.0592 ^{+,a} (0.0303)			
highmath07		-0.324 ^{***} (0.0179)			
lowmath07		0.708 ^{***} (0.0219)			
avegrowXlowmath		-0.688 ^{**} (0.2200)			
avegrowXhighmath		-0.103 (0.1760)			
nobonusXavegrowXlowmath		0.868 [*] (0.4030)			
nobonusXavegrowXhighmath		0.5 (0.3620)			
_cons	-0.0228 (0.0251)	-0.136 ^{***} (0.0271)	0.572 ^{***} (0.0306)	-0.136 ^{***} (0.0271)	-0.460 ^{***} (0.0260)
<i>N</i>	86332	86332	26556	31789	27987
<i>R</i> ²	0.003	0.176	0.008	0.003	0

Cluster robust standard errors in parentheses (clustered on the school); + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

^a denotes a significant difference in the nobonus effect between low and high achieving students ($p=0.037$)

Lowmath is an indicator for 4th grade reading score -0.5 SD or below the within-school average

Highmath is an indicator for 4th grade reading score $+0.5$ SD or above the within-school average

Table A3. RD Models Predicting Standardized 5th Grader Science Level, .20 Bandwidth

	(1) All	(2) All	(3) Low	(4) Med	(4) High
nobonus	0.00809 (0.0461)	0.0244 (0.0515)	-0.0226 (0.0419)	0.0244 (0.0515)	0.0401 (0.0513)
avegrow	2.280*** (0.2770)	2.691*** (0.3230)	1.970*** (0.2660)	2.691*** (0.3230)	2.017*** (0.2840)
nobonusXavegrow	0.749 (0.4950)	0.539 (0.5560)	0.152 (0.4280)	0.539 (0.5560)	1.536** (0.5880)
nobonusXlowread07		-0.047 ^a (0.0306)			
nobonusXhighread07		0.0157 ^a (0.0276)			
highread07		0.805*** (0.0168)			
lowread07		-0.661*** (0.0188)			
avegrowXlowread		-0.721*** (0.1870)			
avegrowXhighread		-0.674*** (0.1690)			
nobonusXavegrowXlowread		-0.387 (0.3360)			
nobonusXavegrowXhighread		0.997** (0.3320)			
_cons	-0.0923** (0.0296)	-0.153*** (0.0337)	-0.813*** (0.0284)	-0.153*** (0.0337)	0.652*** (0.0306)
<i>N</i>	86829	85938	25339	32390	28209
<i>R</i> ²	0.048	0.388	0.056	0.084	0.067

Cluster robust standard errors in parentheses (clustered on the school); + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001.

^a denotes a significant difference in the nobonus effect between low and high achieving students (p=.091)

Lowread is an indicator for 4th grade reading score -.5 SD or below the within-school average

Highread is an indicator for 4th grade reading score +.5 SD or above the within-school average

Reading pretest used because no science only administered in 5th grade.