**Jumping at the Chance:**

**The Effects of Accountability Consequences on Student Achievement[[1]](#footnote-1)**

Douglas Lee Lauen

Assistant Professor of Public Policy

University of North Carolina at Chapel Hill

**Abstract**

Pay for performance plans are spreading across the country due to 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 from North Carolina, I examine the effects of accountability consequences on test scores in 2008, a year in which math and reading scores were “high-stakes,” and science tests were “low stakes.” Results from non-parametric discontinuity models show that at the margin accountability incentives cause higher reading gains and have no adverse effects on science scores or on low achieving students. Incentive effects on science are generally positive rather than negative. Effects on science are much stronger in low poverty schools, however, which suggests that interventions implemented in these schools to increase math and reading scores may have complemented, rather than substituted for, science instruction.

**Jumping at the Chance:**

**The Effects of Accountability Consequences on Student Achievement**

During the 1960s and 1970s, scholars became quite pessimistic about school’s role 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.

In the short run, teacher bonus programs may increase teacher productivity and thus student achievement through at least two mechanisms: teachers seeking to maximize income and teachers seeking to minimize public “accountability threats” (Carnoy & Loeb, 2002; Figlio & Rouse, 2006; Hanushek & Raymond, 2005; Jacob, 2005). In the long run, teacher bonus programs may also change the composition of the teaching force (Podgursky & Springer, 2007). Existing research focuses mainly on the hypothesis that teacher productivity will increase because teachers seek to maximize their income by increasing effort, with careful designs to determine whether individual or group based incentives have positive effects on intended outcomes and minimal unintended consequences (Fryer, 2011; Glewwe, Ilias, & Kremer, 2010; Lavy, 2002, 2009; Marsh et al., 2011; McCaffrey, Pane, Springer, Burns, & Haas, 2011; Muralidharan & Sundararaman, 2009; Springer et al., 2010). The present study, on the other hand, is designed to examine whether the consequences of accountability threats embedded in a long-standing statewide system of school wide bonuses for meeting test score growth targets cause test score increases in schools at the margin of bonus targets, narrow the curriculum at the expense of science, and narrow or widen test score gaps. A study of school responses to public accountability incentives linked to a bonus program is needed for several reasons. First, in the present time of budget austerity, bonuses from public sector programs will probably not be large enough to induce substantial changes in teaching practice;[[2]](#footnote-2) second, teacher bonus programs are a matter of public record and are most often embedded within public accountability policies which seek to hold schools, principals, and groups or individual teachers publicly accountable for student performance; and third, there is ample evidence from prior research that teachers and principals respond to accountability pressure in ways that raise test scores (Dee & Jacob, 2011; Jacob, 2005; Reback, 2008; Wong, Cook, & Steiner, 2009). Therefore, in addition to small scale randomized experiments, it is important for the research base on teacher incentives to include well designed studies of accountability incentives linked to public accountability programs.

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).[[3]](#footnote-3) 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. The state department of public instruction calculated quantitative scores to determine eligibility for both the expected growth and high growth bonuses, and the report card ratings that accompanied each bonus. Because public officials assigned schools to treatment on the basis of continuous variables with sharp thresholds, this study uses a regression discontinuity design to examine the effects of the accountability incentives embedded within this program.

Using statewide student-level data from North Carolina, I examine the 2008 math and reading test score gains for schools just below the margin of the spring 2007 bonus threshold to determine whether scores are higher relative to what they would have been had the school earned a bonus. Because schools were held accountable for math and reading test score gains in the 2007-2008 school year, I hypothesize that at the margin, failing to achieve growth targets will induce schools 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 on particular subgroups of students. Therefore, I hypothesize negative or null effects on low stakes test scores outside the accountability framework. Specifically, I examine whether accountability incentives narrow the curriculum at the expense of 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.

This study differs from existing pay for performance randomized controlled trials (RCTs). Such trials test whether offering bonuses for improving student performance has the effect of raising test scores. Because the North Carolina program was offered to all schools in the state, the present study cannot make inferences about the average differences between treatment and control schools. Instead, this study asks a more narrowly tailored question: do test score differences emerge between schools at the margin of a performance metric? A difference at the margin could indicate that schools respond to the positive and negative incentives embedded in a long-standing and statewide public accountability program.

Background and Evidence

Experimental and regression discontinuity (RD) evidence on the effects of teacher bonus programs on student achievement is mixed, with studies in India, Israel, and Kenya reporting positive results of individual or group-based teacher incentives (Glewwe, et al., 2010; Lavy, 2002, 2009; Muralidharan & Sundararaman, 2009) and no U.S.-based studies reporting positive effects (Fryer, 2011; Marsh, et al., 2011; McCaffrey, et al., 2011; Springer, et al., 2010). For example, a study 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 reports 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). Studies based in Israel using RD designs report positive effects of both individually-based (Lavy, 2009) and school-based (Lavy, 2002) bonuses.

Studies based in the U.S., however, paint a much more pessimistic picture of the promise of paying teachers bonuses for test score achievement. The POINT study, based in metro Nashville schools, randomly assigned about 300 middle school math teachers to two groups: a treatment group eligible for individually-based 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). 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.[[4]](#footnote-4) RCTs of group-based incentives in the U.S. have also produced null and even negative effects. A study in Round Rock, Texas, 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). The largest U.S. RCT to date is an evaluation of a school-based incentive program in New York City which between 2007 and 2010 distributed a total of about $75 million to over 20,000 teachers and other UFT members in about 200 treatment schools (Fryer, 2011) and produced null and negative effects of the bonus program on student achievement in reading and math (Fryer, 2011; Marsh, et al., 2011).

RCT evidence on the effects of being assigned to a teacher bonus program addresses the question: if teachers are offered more money, will they (or can they) change their effort or their practices to increase student achievement? Evidence from RCTs in the U.S. suggests that monetary incentives do not raise test scores. The present study is designed to address a different, but important, question: do test score differences emerge between schools that failed and met a bonus target the prior year? I posit that a difference will emerge because of the accountability threats embedded in the bonus/accountability program. 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; Dee & Jacob, 2011; Figlio & Rouse, 2006; Hanushek & Raymond, 2005; Jacob, 2005; Jacob & Lefgren, 2004; Reback, 2008; Wong, et al., 2009).

This accountability threat of the North Carolina bonus program is comprised of two incentives: extra pay and public recognition. These incentives can be framed as positive or negative. Schools failing to attain the target get the negative sanction of reduced pay and the negative sanction of “naming and shaming.” Schools meeting the target get the positive incentive of increased pay and the positive incentive of public recognition. If the findings show a spread in outcomes at the cutoff, a plausible interpretation of this discontinuity is that there is a *difference* in effort, which could emerge due to increased effort on the part of teachers and other actors in the production of educational achievement (e.g., principals, students, and parents) below the cutoff or reduced effort above it.

In short, overlooked in RCT pay for performance literature is whether monetary incentives embedded in a system of public accountability could have positive effects. With the exception of the NYC program, 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 publically 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 money, but instead by professional norms and public accountability, then perhaps tailoring incentives that tap intrinsic motivation, or a mix of intrinsic and extrinsic motivation, would be more effective than pay for performance alone.

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. The program, which applied to all elementary, middle, and high schools[[5]](#footnote-5) 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. The growth formula in 2007 was based on changes in test scores standardized based on 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. The school’s expected growth score was the average of student academic change scores in both reading and math. 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—$750 for meeting expected achievement growth. Schools meeting the expected growth standard and with at least 60% of students meeting their academic change expectation received an additional $750 bonus, for a total of $1500, which policymakers called a “high growth” bonus. Unlike some incentive programs that allow for local discretion about how to distribute bonuses, the state paid bonuses directly to eligible staff. Schools were publicly recognized for their performance and assigned labels based on performance based on growth category and their performance composite (the percent of students at or above grade level). A handful of schools each year received mandated state assistance if they failed to make expected growth and their performance composite was especially low. This assistance took the form of a team visiting the school and conducting a needs assessment. Assistance could take the form of funds to reduce class size, additional instructional support positions, and additional staff development.

This study examines the effect of failing the 2007 expected growth and high growth targets on 2008 test scores of elementary and middle school students. A North Carolina Department of Public Instruction press release dated September 7, 2007 officially announced the results of the 2006-2007 accountability ratings. Seventy percent of elementary and middle schools met at least their expected growth target and received at least $750 bonuses.[[6]](#footnote-6) Twenty-three percent of elementary schools also met their high growth target and received an additional $750 for a total of $1500 in bonus pay.

Data

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 sample includes students from urban, suburban, small town, and rural communities, and is 55% white, 28% black, 10% Hispanic, 13% academically gifted, 12% with special education needs, 46% eligible for free or reduced price lunch, and 28% with college educated parents (see table 1). 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 533,000 4th - 8th grade students nested within about 1,800 schools.[[7]](#footnote-7)  Table 2 displays counts of students and schools at each level of the forcing variable in the interval of -.10 to .10.

Hypotheses and Identification Strategy

This study examines whether failing to get a bonus and the negative publicity in failing to meet accountability standards has an effect on test scores in the following school year. 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 accountability consequences[[8]](#footnote-8) 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. I hypothesize that scores will be higher for bonus losers relative to bonus winners because prior research shows that educators tend to respond to accountability pressure by raising high stakes test scores. This may be due to exerting more effort and/or changing educational practices due to the accountability incentives embedded into the public bonus program. It could be, however, that motivation was just as high in schools barely meeting performance targets, especially if schools knew their precise score on the performance indicator used to assign bonuses (systematic evidence is unfortunately not available on this point). If schools that barely met the target feel just as much accountability pressure as schools that barely missed the target, then treatment effects at the margin could be null or at least suppressed. If, on the other hand, educators believe accountability ratings to be purely a function of random noise, then their incentive to implement improvement strategies could be reduced and we might expect null results. Of course, a positive effect of failing to achieve a bonus threshold could be also be interpreted as a negative effect of falling above it, which could happen if educator effort falls after attaining a bonus.

*Do accountability consequences promote a narrowing of the curriculum at the expense of science?*—In 2008, the initial year of the 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. On the other hand, if interventions to raise math and reading achievement have complementary spillover effects on science, accountability threats could increase science scores even though science was a low stakes test.

*Do accountability consequences have differential effects on low or high achieving students?*—North Carolina’s system creates an incentive to focus on the students with the highest potential for growth. A priori, it is not clear whether low or high achievers have the highest potential for growth in a particular year or whether incentive effects to direct resources towards particular groups of students would be greater in reading or math. Therefore, I examine distributional effects due to the policy relevance of determining whether incentive effects have effects on the test score gap between low and higher achieving students.

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 two forcing variables: 1) the spring 2007 expected growth score that determined eligibility for the expected growth bonus ($750) and 2) the change ratio that determined eligibility for the high growth bonus ($1500). These variables, performance metrics with thresholds, were the sole determinants of each type of bonus payment. School bonuses in 2007 were based solely on reading and math achievement. Treatment assignment was “sharp” in the sense that expected and high growth ratings, and thus bonus payments, were legislatively determined based on performance metrics. Because state assessment officials assigned schools bonuses based on a formula, there was no opportunity for manipulation of treatment assignment around the cutoff (see figure 1a and 1b).

To test for effects on math and reading gains and science scores, I estimate RD models with 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 smoothed non-parametric functions to the observations within a distance *h* on either side of the cutoff point, *c*, on the forcing variable, *X,* producing a difference of intercepts at the cutoff: (Imbens & Lemieux, 2008). 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. To correct for the non-independence of student observations within schools, I report cluster robust standard errors.[[9]](#footnote-9) As specification checks, I vary the bandwidth used to weight the data and include student and school control variables to the non-parametric regressions.

To test for differential effects on students based on prior achievement, I use a parametric RD model. The parametric RD model shown in equation (1) regresses a test score outcome, *y*, on a treatment indicator, *D* (coded *D*=1 for observations below the cutoff, and 0 at the cutoff and above), the forcing variable centered on 0, *X-c*, and the interaction of the treatment indicator and the forcing variable to allow for different slopes on either side of the cutoff:

1. , where .

From this equation I obtain the treatment effect of failing to meet the bonus target on next year’s test score, , and its standard error. When estimated on the same interval of observations around the cutoff as the bandwidth specified in the local polynomial regression, a linear OLS regression with a difference in slopes to the left and to the right of the cutoff produces an identical treatment effect as a non-parametric RD model with a rectangular kernel (Lee & Lemieux, 2010). To test for differential effects, I add variables coding student prior achievement to equation 1. I define students as low, average, or high in prior achievement based on their position in the spring 2007 *within-school* reading pre-test (e.g., 4th grade for 5th graders) score distribution. Students defined as *Low* in equation (2), below, 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. Because students took science for the first time in 2008 in only 5th and 8th grade there was no pretest in this subject. Therefore, I report two sets of science results, one with a reading pretest and one with a math pretest. 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:

1. , where .

In equation (2), is the effect of failing the bonus target on the test scores of students in the middle of the within-school prior test score distribution, is the test score difference between low and medium achieving students, and is the test score difference between high and medium achieving students.

Results

1. *Failing Expected Growth Treatment Effects*

The most comparable schools on either side of the cutoff are those within the window [-.01 to +.01] on the 2007 expected growth score, the assignment variable. Due to the statewide nature of this sample, there are about 18,500 4th-8th grade students in the estimation sample in 55 elementary and middle schools at -.01 on expected growth and about 31,000 students in 95 schools at 0 and +.01 on expected growth. Recall that schools at or above 0 on expected growth were eligible for a $750 bonus and schools at -.01 and below were not eligible for this bonus. If assignment to treatment is as good as random for students in schools this close to the cutoff, the causal effect of failing to attain the 2007 expected growth bonus target on 2008 student test score is the mean difference in test scores between students in schools at -.01 and students in schools at 0 and .01.

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 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. Within the interval of -.01 to +.01 on the forcing variable (a total N of 150 schools and about 49,500 students), the mean difference at the cutoff in 2008 standardized reading gain is .033 (.018), which is marginally significant (*p*=.067) with cluster robust standard errors, with the positive value favoring students in schools that failed to receive bonuses in 2007.[[10]](#footnote-10) This suggests that failing the bonus target produced a small incentive effect in the schools that barely missed the threshold the prior year. The effect in standardized math gain score within the same interval is about the same size, but not statistically distinguishable from zero at conventional levels, .034 (.034), *p*=.308. The effect on standardized science score level is positive and distinguishable from zero, .136 (.067), *p*=.045, suggesting that incentives did not narrow the curriculum at the expense of science.[[11]](#footnote-11)

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 below the cutoff into the area above the cutoff. The counterfactual for bonus schools is the extrapolation of a trend based on data points above the cutoff into the area below the cutoff. Without extending the bandwidth and evaluating functional form, it is difficult to assess these counterfactual conditions.

Figures 2-4 display for each dependent variable non-parametric smoothed functions on either side of the expected growth cutoff overlaid with bin average scores. On the Y-axis is the 2008 student test score and on the X-axis is the school’s 2007 expected growth score. The treatment effect reported in the note of each figure is estimated with a triangular kernel and cluster corrected standard errors. I estimate each model with a bandwidth of .05, but report estimates with other bandwidths below. The graphs show a positive treatment effect of failing the 2007 expected growth target on 2008 reading gain, math gain, and science test score. The treatment effect on reading gain of attending a school failing the bonus target is .063, and is statistically significant at conventional levels. The treatment effect on math is .048, but is imprecisely estimated. In science the treatment effect is .179 and is statistically significant at conventional levels. These findings are consistent with the hypothesis that schools that fail bonus targets in the prior year implement educational interventions that cause higher achievement gains in reading and do not narrow the curriculum at the expense of science, but instead have complementary effects on science scores.

Table 3 displays treatment effects from alternative specifications of the RD model. A non-parametric model with a rectangular kernel produces a treatment effect on reading gain of .083 instead of .063. A parametric specification (see equation 3, above) produces exactly the same treatment effects as the non-parametric model with the rectangular kernel. All reading results are significant at conventional levels. Using a rectangular kernel produces a larger treatment effect on math gain that approaches statistical significance, whereas this specification reduces the size and statistical significance of effect on science score at this particular bandwidth.

Perhaps more important than alternative specifications of the RD model is testing the robustness of effects across a variety of bandwidths and with covariates. The reading effects produced by a non-parametric model with a triangular kernel are statistically significant and approximately the same size at bandwidths .05-.10 (table 4, column 1). Adding student and school control variables does not affect the size or significance of these effects substantially (column 2). With and without control variables the math effects are positive, but small and largely not statistically significant at conventional levels. Without control variables, the science effects are positive, relatively large, and all significant at least at the 90% confidence level, which contradicts the hypothesis that incentives would narrow the curriculum at the expense of science (column 5). With control variables, however, the size of the effects falls by about half. Further investigation reveals that this is due to a heterogeneity of the treatment effect across high and low poverty schools, which I discuss in subsection D, below.

1. *Failing High Growth Treatment Effects*

Schools meeting the expected growth benchmark that also had at least 60% of students meeting their academic change expectations were eligible for $1500 “high growth” bonuses. To estimate the incentive effects around this cutoff, I drop all schools that failed expected growth and use the change ratio (the percent of students meeting their academic change expectations) as the forcing variable. In the reading model estimation samples, at bandwidths of .02, .05, and .10, there are 47, 106, and 224 schools, respectively. (The corresponding figures around the expected growth cutoff are 267, 582, and 1029 schools, respectively.) Not surprisingly, the reduction in sample size affects the precision of the high growth treatment effect estimates. For this reason, I extend the range of bandwidths somewhat to assess whether increasing sample size affects the statistical significance of results. The effects of failing the high growth target on reading gain are positive, about the same size as the effects of failing the expected growth target, and marginally significant between bandwidths of .08 and .13 (table 5). The effects of failing the high growth target on math gain are larger than the effects of failing the expected growth target close to the cutoff, but do not achieve statistical significance at any bandwidth. The effects on science are again positive, but do not attain statistical significance.

In summary, the incentive effects of failing to attain expected growth and high growth are all positive, even for science achievement, and are precisely estimated in the reading and science expected growth models. The incentive effects on math are imprecisely estimated in both sets of models. The incentive effects of failing high growth on reading gain are marginally significant, but only at wider bandwidths. The reading results are robust when school covariates are included, but the science results are not, which I discuss further below.

1. *Differential Treatment Effects of Failing Expected Growth*

I estimate moderator models shown in equation 4 at a variety of bandwidths and with control variables, but find almost no significant differences between low and average achievers or between low and high achievers (table 6; models without control variables also produce no differential effects). To take one example, within the interval of -.05 to +.05, the effect of failing the expected growth target on reading gain is .077 for low achievers, .054 for average achievers, and .050, for high achievers. Each coefficient is significantly different from zero, but no differences among the groups are statistically significant at conventional levels (panel A). Only two of the 108 comparisons (low vs. medium, high vs. medium, low vs. high) are statistically significant at p<.10, for a rejection rate of .0185, which is lower than would be expected by chance. Given that most treatment effect estimates across the three groups are roughly similar in size, I conclude that the incentives embedded in the bonus program had no differential effects on low achieving students, positive or negative. A specification with interactions with linear and squared terms of prior achievement produces qualitatively similar results with non-significant interaction terms. I also find no differential effects of failing the high growth standard (results not shown, but available from author upon request).

1. *Heterogeneity of the treatment effect on science*

As noted above, including student and school covariates reduced the size of the effects on science test scores. This is due to an unanticipated heterogeneity of the treatment effect on high and low poverty schools (see table 7). Specifically, in low poverty schools, the treatment effect of failing the expected growth target is positive and significant, whereas the effects in high poverty schools are positive, but much smaller and never significant (the effects in the two types of schools are significantly different in four out of the nine intervals: +/- .04, .06, .07, and .08).[[12]](#footnote-12) Larger science effects in low poverty schools may be explained by the fact that these schools had the organizational capacity to implement interventions that had positive spillovers science learning. Alternatively, low poverty schools may have been more likely to focus instruction on science in preparation for science becoming a high stakes subject in the 2008-2009 school year. Without data on instructional changes in high and low poverty schools, however, it is impossible to examine the plausibility of these explanations.

1. *Specification checks*

This study posits a discontinuous jump in the treatment effect at the cutoff. 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. Following a procedure outlined in Imbens and Lemieux (2008), 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 non-parametric specification with a triangular kernel with cluster robust standard errors at bandwidths of .02 to .10 for both tests. To avoid confounding a placebo effect with the discontinuities at the actual cutoff, in the lower median test, I exclude observations at 0 and above (the actual cutoff); in the upper median test, I exclude observations below 0. In a total of 54 tests (three dependent variables, two placebo points, nine bandwidths), I find five placebo treatment effects that are statistically significant at the .10 level or below for a rejection rate of .09, which is about what is expected by chance. I conduct the same test for median placebo points along the high growth forcing variable and produce a rejection rate of .03 (results from author upon request).

I also test for discontinuities at the actual cutoffs in school level baseline covariates that could confound the treatment effect if they were also discontinuous at the cutoff. I examine percentage black students, percentage poor students (as measured by free and reduced lunch eligibility), percent novice teachers, and teacher turnover rate. I find no evidence of discontinuities in percentage black students, percent novice teachers or teacher turnover rate. I find evidence of a discontinuity in school poverty at the cutoff, with *lower* poverty rates just below the cutoff than above it. Note that this is in the opposite direction of the treatment effects found in this study (positive effects on test scores for schools falling just below the cutoff). Given that schools could not feasibly change their poverty levels during the 2006-2007 in anticipation of meeting or failing the bonus target in the spring of, I do not consider this discontinuity a threat to the validity of the findings of reported above. Including school poverty as a covariate does not change the conclusions about the reading or math results, but does alter my conclusions about effects on science achievement, with heterogeneous effects, as discussed above.

Conclusion

U.S.-based performance pay RCTs test whether offering bonuses for improving student performance increases the effort and student test scores of the teachers randomized to treatment relative to those randomized to control. Evidence of individual and group based incentives in the U.S. suggests that monetary incentives do not raise test scores. This study examines whether test score differences emerge between schools at the margin of a performance metric used by policymakers to award bonuses to teachers and publically recognize and sanction schools.

I find that at the margin of accountability metrics, students in schools that failed to attain bonus thresholds generally outperformed students in schools that met such thresholds. All reported incentive treatment effects of failing a bonus target are positive. Evidence is strongest in reading, a high-stakes subject in the year of the study. I hypothesized that failure to meet test score growth targets in reading and math would narrow the curriculum at the expense of science, which was tested for the first time in 2008, and was thus, low stakes. I find generally positive rather than null or negative effects on science. An unanticipated result is that the effect of failing the bonus target varies by school poverty. In low poverty schools, the effect of failing the bonus target is positive, precisely estimated, and fairly large (about .20 effect size); in high poverty schools this effect is also positive, not statistically significant, and generally very close to zero. Explanations of these heterogeneous effects include that low poverty schools may have implemented instructional practices that had effects on both high and low stakes subjects and that low poverty schools may have focused teachers on improving science instruction in preparation for science becoming a high stakes subject in the 2008-2009 school year. Further research is needed to test the tenability of these explanations.

The findings of this study contribute to the literature on the effects of incentives in education, which shows that high stakes testing raises test scores on high stakes and sometimes low stakes tests (Carnoy & Loeb, 2002; Chiang, 2009; Dee & Jacob, 2011; Figlio & Rouse, 2006; Hanushek & Raymond, 2005; Jacob, 2005; Jacob & Lefgren, 2004; Reback, 2008; Wong, et al., 2009). The absence of a narrowing effect on science and the absence of adverse distributional effects for low achieving students suggests that the design of North Carolina’s system could be worth emulating as states implement performance pay and the next generation of accountability systems.

Establishing the short run effects of accountability incentives on student test scores in high and low stakes tests is an important step in contributing to the research base about educational accountability and performance pay. In light of the fade-out effects found in Glewwe, et al (2010), however, 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.

An important limitation of the study is its retrospective design, which includes no data on changes in educator effort and classroom practices and therefore cannot explain why effects emerge. To confirm that accountability incentives alter educator effort and practices, for example, would require measures of these things. Therefore, I am limited to posing plausible interpretations but not firm conclusions about why effects emerge. I posited two mechanisms that could explain how performance pay could alter behavior: seeking to maximize income and aiming to avoid accountability threats. Another limitation of this study is the inability to tease apart these two explanations. The treatment tested 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**

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.

Dee, T. S., & Jacob, B. (2011). The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management, 30*(3), 413-700.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications* (1st ed.). London ; New York: Chapman & Hall.

Figlio, D., & Rouse, C. (2006). Do Accountability and Voucher Threats Improve Low-Performing Schools? *Journal of Public Economics, 90*(1-2), 239-255.

Fryer, R. G. (2011). *Teacher Incentives and Student Achievement: Evidence from New York City Public Schools*. NBER Working Paper 16850 March 2011.

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.

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.

Ladd, H. F., & Zelli, A. (2002). School-based accountability in North Carolina: The responses of school principals. *Educational Administration Quarterly, 38*(4), 494-529. doi: Doi 10.1177/001316102237670

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.

Marsh, J. A., Springer, M. G., McCaffrey, D., Yuan, K., Epstein, S., Koppich, J., et al. (2011). A Big Apple for Educators: New York City's Experiment with Schoolwide Performance Bonuses (Vol. MG-1114-FPS): Rand Corporation.

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*.

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

Podgursky, M. J., & Springer, M. G. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management, 26*(4), 909-950.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics, 92*(5-6), 1394-1415.

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., & Campbell, D. T. (1960). Regression-Discontinuity Analysis - an Alternative to the Ex-Post-Facto Experiment. *Journal of Educational Psychology, 51*(6), 309-317.

Wong, M., Cook, T. D., & Steiner, P. M. (2009). *No Child Left Behind: An Interim Evaluation of its Effects on Learning using two Interrupted Time Series each with its own Non-Equivalent Comparison Series.* Working paper.

**Figures**

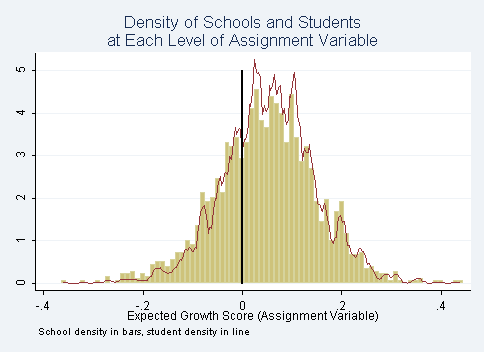


Figure 1a. Density of students and schools at each bin of the expected growth forcing variable.

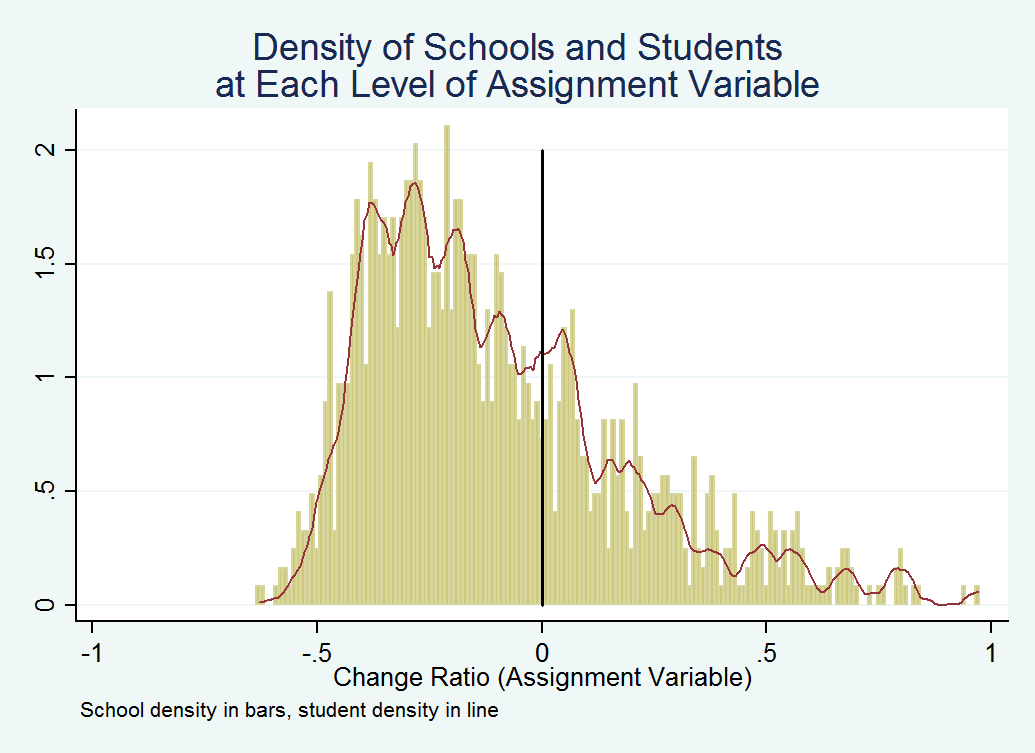


Figure 1b. Density of students and schools at each bin of the change ratio forcing variable.

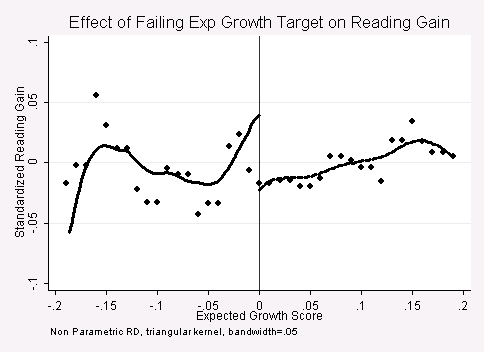


Figure 2. Treatment effect of failing expected growth on reading gain (4th - 8th graders). Non-parametric RD model, triangular kernel. Cluster robust SE. Effect of failing to get a bonus: .063 (.023)\*.

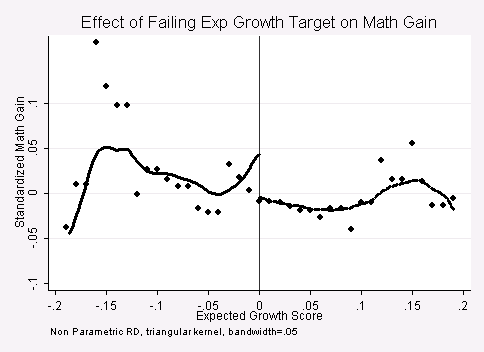


Figure 3. Treatment effect of failing expected growth on math gain (4th - 8th graders). Non-parametric RD model, triangular kernel. Cluster robust SE. Effect of failing to get a bonus: .048 (.043).

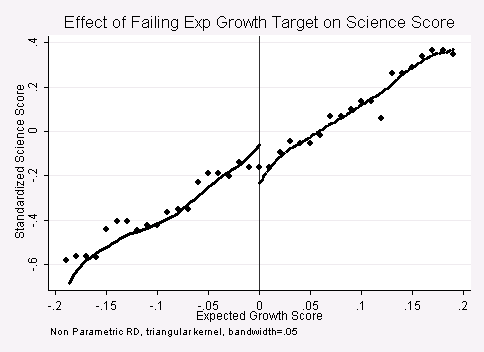


Figure 4. Treatment effect of failing expected growth on science scale score (5th and 8th graders). Non-parametric RD model, triangular kernel. Cluster robust SE. Effect of failing to get a bonus: .179 (.086)\*.

**Tables**

Table 1. Descriptive Statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Count | Mean | SD | Min | Max |
| **Dependent Variables** | | | | | | |
| zrgain | Standardized '07-'08 reading gain (4th-8th graders) | 472064 | 0 | 1 | -9.17 | 7.22 |
| zmgain | Standardized '07-'08 math gain (4th-8th graders) | 473706 | 0 | 1 | -5.72 | 6.86 |
| zsciscal | Standardized '08 science level (5th and 8th grade only) | 205455 | 0 | 1 | -3.18 | 3.22 |
| **Forcing variables and treatment indicators** | | | | | | |
| expgrow | Expected growth score (forcing variable) | 532822 | 0.05 | 0.092 | -0.36 | 0.44 |
| faileg | School failed expected growth cut score (treatment indicator) | 532994 | 0.25 | 0.433 | 0 | 1 |
| chratio | Change ratio (forcing variable) | 532890 | -0.21 | 0.378 | -1.5 | 4.29 |
| failhg | School failed high growth cut score (treatment indicator) | 400075 | 0.68 | 0.468 | 0 | 1 |
| **Student characteristics / control variables** | | | | | | |
| white | White student | 533391 | 0.55 | 0.497 | 0 | 1 |
| black | Black student | 533391 | 0.28 | 0.447 | 0 | 1 |
| hisp | Hispanic student | 533391 | 0.1 | 0.298 | 0 | 1 |
| gifted | Gifted student | 533391 | 0.13 | 0.337 | 0 | 1 |
| specialed | Student receives special education services | 531637 | 0.12 | 0.327 | 0 | 1 |
| frl | Student receives free or reduced priced lunch (poverty indicator) | 531842 | 0.46 | 0.499 | 0 | 1 |
| nonstrmv | School mobility indicator (non-grade-span-related move) | 533391 | 0.1 | 0.296 | 0 | 1 |
| **School characteristics / control variables** | | | | | | |
| zpblack07 | School percent black (standardized) | 533391 | -0.07 | 0.88 | -1.15 | 2.73 |
| zppoor07 | School percent poor (standardized) | 533391 | -0.19 | 0.917 | -2.17 | 2.31 |
| zptturn07 | School teacher turnover rate (standardized) | 511621 | 0.02 | 0.947 | -1.96 | 6.71 |
| zpnovice07 | School percent novice teachers (standardized) | 520286 | 0.03 | 0.958 | -2.16 | 3.74 |

Table 2. Count of students and schools at each level of the Expected Growth forcing variable

|  |  |  |
| --- | --- | --- |
| Exp Growth (forcing variable) | Count of Students | Count of Schools |
| -.10 | 3653 | 16 |
| -.09 | 4906 | 24 |
| -.08 | 11989 | 38 |
| -.07 | 6139 | 34 |
| -.06 | 7260 | 36 |
| -.05 | 13023 | 44 |
| -.04 | 12784 | 38 |
| -.03 | 14140 | 59 |
| -.02 | 16413 | 57 |
| -.01 | 20883 | 61 |
| 0 | 16541 | 52 |
| .01 | 18439 | 59 |
| .02 | 24664 | 73 |
| .03 | 28884 | 81 |
| .04 | 21798 | 68 |
| .05 | 21144 | 65 |
| .06 | 26482 | 78 |
| .07 | 23883 | 75 |
| .08 | 24196 | 71 |
| .09 | 17410 | 59 |
| .10 | 27445 | 79 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table 3. Effects of Failing the Expected Growth Bonus Target from Parametric and Non-Parametric RD Models | | | | | |
|  | Reading Gain, grades 4-8 | Math Gain, grades 4-8 | Science Scale, grades 5 & 8 |
| Non Parametric | 0.063\* | 0.048 | 0.179\* |
| Triangle | (0.023) | (0.043) | (0.086) |
| Non Parametric | 0.083\* | 0.071+ | 0.119 |
| Rectangle | (0.022) | (0.041) | (0.081) |
| Parametric | 0.083\* | 0.071+ | 0.119 |
|  | (0.022) | (0.041) | (0.081) |
| Obs | 153986 | 154529 | 66735 |
| Bandwidth / Interval | +/- 0.05 | +/- 0.05 | +/- 0.05 |
|  | Note: Non-parametric and parametric RD models with cluster robust standard errors. + p<=.10, \* p<=.05. | | | |

Table 4. Effects of Failing the Expected Growth Bonus Target Across Multiple Bandwidths, Non-Parametric RD Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) | (5) | (6) |
| Bandwidth | Reading Gain | Reading Gain | Math Gain | Math Gain | Science Scale | Science Scale |
| +/- 0.02 | 0.0145 | -0.0178 | 0.0300 | 0.0379 | 0.308\* | 0.146+ |
|  | (0.0345) | (0.0374) | (0.0689) | (0.0626) | (0.135) | (0.0866) |
| +/- 0.03 | 0.0142 | -0.0116 | 0.0371 | 0.0483 | 0.311\* | 0.133 |
|  | (0.0328) | (0.0351) | (0.0681) | (0.0641) | (0.135) | (0.0844) |
| +/- 0.04 | 0.0344 | 0.0159 | 0.0223 | 0.0481 | 0.244\* | 0.0970 |
|  | (0.0258) | (0.0273) | (0.0510) | (0.0488) | (0.101) | (0.0611) |
| +/- 0.05 | 0.0627\* | 0.0509\* | 0.0481 | 0.0693+ | 0.179\* | 0.0853+ |
|  | (0.0227) | (0.0236) | (0.0434) | (0.0413) | (0.0860) | (0.0517) |
| +/- 0.06 | 0.0644\* | 0.0554\* | 0.0484 | 0.0649+ | 0.142+ | 0.0762+ |
|  | (0.0202) | (0.0209) | (0.0385) | (0.0365) | (0.0757) | (0.0453) |
| +/- 0.07 | 0.0577\* | 0.0508\* | 0.0362 | 0.0497 | 0.133+ | 0.0730+ |
|  | (0.0187) | (0.0192) | (0.0355) | (0.0336) | (0.0701) | (0.0414) |
| +/- 0.08 | 0.0534\* | 0.0483\* | 0.0306 | 0.0437 | 0.125+ | 0.0678+ |
|  | (0.0175) | (0.0180) | (0.0332) | (0.0315) | (0.0658) | (0.0385) |
| +/- 0.09 | 0.0479\* | 0.0451\* | 0.0263 | 0.0404 | 0.123\* | 0.0567 |
|  | (0.0162) | (0.0166) | (0.0308) | (0.0293) | (0.0609) | (0.0356) |
| +/- 0.10 | 0.0438\* | 0.0424\* | 0.0224 | 0.0374 | 0.122\* | 0.0509 |
|  | (0.0155) | (0.0159) | (0.0293) | (0.0280) | (0.0579) | (0.0340) |
| Controls? | N | Y | N | Y | N | Y |

Nonparametric RD model with triangular kernel and cluster corrected standard errors; control variables are race/ethnicity, gifted, special education, family poverty (free/reduced priced lunch), student school mobility, grade level, school percent black, school percent poor, teacher turnover, school percent novice teachers; + *p* < 0.10, \* *p* < 0.05.

Table 5. Effects of Failing the High Growth Bonus Target Across Multiple Bandwidths, Non-Parametric RD Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) | (5) | (6) |
| Bandwidth | Reading Gain | Reading Gain | Math Gain | Math Gain | Science Scale | Science Scale |
| +/- 0.02 | 0.0509 | 0.101 | 0.122 | -0.0376 | 0.0390 | 0.0306 |
|  | (0.0881) | (0.118) | (0.152) | (0.200) | (0.280) | (0.175) |
| +/- 0.03 | 0.0524 | 0.110 | 0.115 | 0.00284 | 0.0187 | 0.0214 |
|  | (0.0876) | (0.105) | (0.149) | (0.173) | (0.272) | (0.145) |
| +/- 0.04 | 0.0655 | 0.101 | 0.118 | 0.0722 | 0.0768 | -0.0516 |
|  | (0.0676) | (0.0743) | (0.118) | (0.129) | (0.217) | (0.104) |
| +/- 0.05 | 0.0769 | 0.0877 | 0.102 | 0.107 | 0.145 | -0.0402 |
|  | (0.0570) | (0.0612) | (0.0986) | (0.0988) | (0.188) | (0.0912) |
| +/- 0.06 | 0.0776 | 0.0812 | 0.0913 | 0.103 | 0.137 | -0.0338 |
|  | (0.0512) | (0.0556) | (0.0864) | (0.0869) | (0.166) | (0.0855) |
| +/- 0.07 | 0.0741 | 0.0752 | 0.0697 | 0.0751 | 0.111 | -0.00192 |
|  | (0.0458) | (0.0498) | (0.0766) | (0.0765) | (0.148) | (0.0801) |
| +/- 0.08 | 0.0734+ | 0.0750+ | 0.0657 | 0.0674 | 0.0974 | 0.0291 |
|  | (0.0417) | (0.0446) | (0.0692) | (0.0684) | (0.136) | (0.0763) |
| +/- 0.09 | 0.0710+ | 0.0718+ | 0.0578 | 0.0577 | 0.0791 | 0.0275 |
|  | (0.0387) | (0.0409) | (0.0642) | (0.0633) | (0.126) | (0.0740) |
| +/- 0.10 | 0.0659+ | 0.0668+ | 0.0543 | 0.0560 | 0.0996 | 0.0483 |
|  | (0.0356) | (0.0375) | (0.0592) | (0.0581) | (0.118) | (0.0703) |
| +/- 0.11 | 0.0590+ | 0.0602+ | 0.0466 | 0.0508 | 0.106 | 0.0552 |
|  | (0.0332) | (0.0349) | (0.0553) | (0.0544) | (0.111) | (0.0671) |
| +/- 0.12 | 0.0544+ | 0.0563+ | 0.0437 | 0.0492 | 0.103 | 0.0595 |
|  | (0.0317) | (0.0333) | (0.0529) | (0.0520) | (0.106) | (0.0646) |
| +/- 0.13 | 0.0512+ | 0.0540+ | 0.0377 | 0.0429 | 0.0977 | 0.0616 |
|  | (0.0304) | (0.0319) | (0.0509) | (0.0502) | (0.102) | (0.0625) |
| +/- 0.14 | 0.0469 | 0.0500 | 0.0342 | 0.0395 | 0.0971 | 0.0619 |
|  | (0.0294) | (0.0307) | (0.0490) | (0.0484) | (0.0978) | (0.0605) |
| +/- 0.15 | 0.0419 | 0.0452 | 0.0292 | 0.0347 | 0.0959 | 0.0626 |
|  | (0.0284) | (0.0295) | (0.0471) | (0.0465) | (0.0935) | (0.0582) |
| +/- 0.16 | 0.0373 | 0.0405 | 0.0231 | 0.0294 | 0.0963 | 0.0615 |
|  | (0.0275) | (0.0286) | (0.0455) | (0.0450) | (0.0904) | (0.0564) |
| Controls? | N | Y | N | Y | N | Y |

Nonparametric RD model with triangular kernel and cluster corrected standard errors; control variables are race/ethnicity, gifted, special education, family poverty (free/reduced priced lunch), student school mobility, grade level, school percent black, school percent poor, teacher turnover, school percent novice teachers; + *p* < 0.10, \* *p* < 0.05.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 6. Differential Effects of Failing the Expected Growth Bonus Target Across Multiple Bandwidths, Parametric RD Models | | | | | | | | | | | | | | | | | |
|  | |  | |  | |  | | |  |  | |  | |  | | |  | |
|  | | | A. Reading Gain | | | | | B. Math gain | | | | | | | |
| Interval | | | Low | | Med | | High | Low | | | Med | | High | | |
| +/- 0.02 | | | **-.088** | | .031 | | .009 | .026 | | | .064 | | .097 | | |
| +/- 0.03 | | | -.009 | | .031 | | .048 | .012 | | | .054 | | .091+ | | |
| +/- 0.04 | | | .103\* | | .070\* | | .060\* | .075 | | | .108\* | | .082+ | | |
| +/- 0.05 | | | .077\* | | .054\* | | .050\* | .068 | | | .071+ | | .053 | | |
| +/- 0.06 | | | .063\* | | .040+ | | .030 | .022 | | | .039 | | .017 | | |
| +/- 0.07 | | | .051+ | | .044\* | | .033 | .030 | | | .046 | | .015 | | |
| +/- 0.08 | | | .050+ | | .045\* | | .032 | .030 | | | .050 | | .029 | | |
| +/- 0.09 | | | .052\* | | .039\* | | .025 | .027 | | | .045 | | .020 | | |
| +/- 0.10 | | | .053\* | | .039\* | | .022 | .030 | | | .050+ | | .019 | | |
|  | | |  | |  | |  |  | | |  | |  | | |
|  | | | C. Science with reading pre-test | | | | | D. Science with math pre-test | | | | | | | |
| Interval | | | Low | | Med | | High | Low | | | Med | | High | | |
| +/- 0.02 | | | .174\* | | .123 | | .094 | .114 | | | .118 | | .078 | | |
| +/- 0.03 | | | .106+ | | .092 | | .114+ | .055 | | | .120+ | | .094 | | |
| +/- 0.04 | | | .090+ | | .073 | | .098 | .042 | | | .101+ | | .078 | | |
| +/- 0.05 | | | .094\* | | .076 | | .077 | .046 | | | .092+ | | .063 | | |
| +/- 0.06 | | | .081\* | | .085+ | | .099\* | .053 | | | .096\* | | .086+ | | |
| +/- 0.07 | | | .058 | | .075+ | | .072 | .045 | | | .078+ | | .063 | | |
| +/- 0.08 | | | .028 | | .054 | | .032 | .024 | | | .056 | | .022 | | |
| +/- 0.09 | | | .030 | | .049 | | .028 | .028 | | | .054 | | .023 | | |
| +/- 0.10 | | | .036 | | .049 | | .019 | .032 | | | .062+ | | **.011** | | |
|  | | Note: Results from parametric RD models; significance tests obtained from equation 4; bolded coefficients denote a significant difference from medium; models include student and school controls: race/ethnicity, gifted, special education, family poverty (free/reduced priced lunch), student school mobility, grade level, school percent black, school percent poor, teacher turnover, school percent novice teachers; + *p* < 0.10, \* *p* < 0.05. | | | | | | | | | | | | | |

Table 7. Heterogeneity of Science Treatment Effect Estimates of Failing the Expected Growth Target

|  |  |  |  |
| --- | --- | --- | --- |
| Interval | High Poverty School | Low Poverty School | Sig Diff |
| +/- .02 | .070 | .223 |  |
| +/- .03 | .057 | .185+ |  |
| +/- .04 | .032 | .229\* | + |
| +/- .05 | .040 | .187\* |  |
| +/- .06 | .035 | .198\* | + |
| +/- .07 | .019 | .179\* | \* |
| +/- .08 | .002 | .167\* | \* |
| +/- .09 | .017 | .138\* |  |
| +/- .10 | .027 | .106+ |  |

Note: Results from parametric RD models; significance tests obtained from equation 4; a + or a \* in the sig diff column denotes whether the treatment effect in low poverty schools is statistically distinguishable from the treatment effect in high poverty schools; models include student and school controls: race/ethnicity, gifted, special education, family poverty (free/reduced priced lunch), student school mobility, grade level, school percent black, teacher turnover, school percent novice teachers; + *p* < 0.10, \* *p* < 0.05.

1. This paper was presented at the Society for Research on Educational Effectiveness Spring 2011 Conference and was supported by grant funding from the Spencer Foundation and University of North Carolina at Chapel Hill. I thank the North Carolina Education Research Data Center at Duke University for supplying the data for this project. [↑](#footnote-ref-1)
2. An important exception to this statement is the DC public schools recent policy of awarding annual bonuses of up to $26,000. [↑](#footnote-ref-2)
3. Personal email correspondence with Kristopher Nordstrom, North Carolina Fiscal Research, September 13, 2010. [↑](#footnote-ref-3)
4. 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 also 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. [↑](#footnote-ref-4)
5. Growth calculations for elementary and middle schools were based on vertically equated end of grade tests. For high schools they were based on non-vertically equated end-of-course tests. [↑](#footnote-ref-5)
6. In 2007, 30 percent of schools failed their expected growth target. The corresponding figures for 2004, 2005, and 2006 were 29, 35, and 45, respectively. [↑](#footnote-ref-6)
7. I begin with 547,310 students and drop 13,919 students (2.5%) who were either retained or attended a school with no "expected growth" rating, or both.  Included in this total of censored students are 132 dropped from 4 schools with very high expected growth scores and 24 students in 5 schools which had fewer than ten students.  Of the remaining 533,391 students, 472,064 (88.5%) had a reading gain score, 473,706 (88.8%) had a math gain score, and 205,455 of the 5th and 8th grade students (95.3%) had a science score.  [↑](#footnote-ref-7)
8. I use the term accountability incentives to mean the mix of positive and negative incentives embedded in the program: the potential for increased or reduced pay and the potential for public shaming or recognition. [↑](#footnote-ref-8)
9. 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. To estimate non-parametric RD models with cluster correction and to produce graphs, I use the free rd 2.0 Stata program (Nichols, 2011). [↑](#footnote-ref-9)
10. This result is obtained from a parametric regression of test score gain on the treatment indicator with cluster robust standard errors. I obtain very similar results from a non-parametric model with a triangular kernel and cluster robust standard errors. [↑](#footnote-ref-10)
11. Total N for the science models is lower due to the fact that only 5th and 8th graders took this test (total N of 63 schools and about 21,000 students). [↑](#footnote-ref-11)
12. Significance testing comes from adding a moderator, low poverty school (defined as 1 if the percent of students with free or reduced priced lunch is below 40%, 0 otherwise), and all the requisite interactions to equation 1. Different definitions of low poverty schools (bottom quartile, bottom quintile) affect significance, but not the qualitative conclusion that there is heterogeneity across the low and high poverty schools, with larger effects on science in low poverty schools than in higher poverty schools. [↑](#footnote-ref-12)