The Effects of School Wide Bonuses on Student Achievement: Regression Discontinuity Evidence from North Carolina

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

Abstract Body

Background / Context

Educational accountability seeks to hold schools, and sometimes individual teachers and students, accountable for student performance. Accountability policy employs positive and negative incentives to induce teachers and principals to increase student performance. For its system based on positive incentives—school-wide bonuses for student test score growth—North Carolina has been a nationwide leader in educational accountability. Since 1998, North Carolina has awarded over one billion dollars in performance bonuses (about $100 million per year). Despite this relatively large investment, to my knowledge there has been no internal or external evaluation of this program.

This study conceives the bonus program as a type of accountability pressure. I posit that teachers and principals will view failing to get a bonus as an "accountability threat" that will change their work activity in ways that produce changes in student test scores. The North Carolina system of awarding bonuses based on test scores is an example of "high stakes" accountability. 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).

Scholars have questioned both the validity of these achievement gains and the unintended consequences of accountability responses. Achievement gains on high stakes assessments may produce test-specific skills, but not generalizable knowledge of academic content (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998). High stakes testing can narrow and fragment the curriculum, promote rote, teacher-directed instruction, and encourage schools to teach test-preparation skills rather than academic content, tendencies that may be stronger in schools with high minority and low income populations (Amrien & Berliner, 2002; Darling-Hammond, 2004; Linn, 2000; Nichols & Berliner, 2007; Orfield & Kornhaber, 2001; Valenzuela, 2005). Moreover, 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). In what is sometimes called "triage theory," 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). Therefore, studies of accountability threats must take into account both the intended and unintended effects of the policy.

Purpose / Objective / Research Question / Focus of Study

This study examines the incentive effects of North Carolina's practice of awarding performance bonuses on test score achievement on the state tests. Bonuses were awarded based solely on whether a school exceeds a threshold on a continuous performance metric. The study uses a

sharp regression discontinuity design, an approach with strong internal validity around the cutoff of the treatment assignment score, to examine three questions:

1) 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.

2) 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.

3) 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.

Setting

The study is set in North Carolina public schools elementary schools (statewide) in the spring of 2008.

Population / Participants / Subjects

The study examines the incentive effects of the bonus program on all public schools enrolling 5th grade students in North Carolina in the spring of 2008. The outcomes are spring 2008 student test scores. The treatment is whether or not the school received a bonus for 2007 test score growth. The unit of analysis of the study is, therefore, student outcomes nested within a school-level treatment. The full analysis sample size is about 78,000 5th grade students nested within about 1,250 schools. Within the bandwidth estimated by the RD models, the sample typically contains about 47,000 students in about 700 schools. Of the total number of schools in the full analysis sample, 877 (69%) were eligible for a performance bonus. To be included in the study, students must have been enrolled in the 4th grade in the spring of 2007 and the 5th grade in the spring of 2008. Students repeating 4th grade in 2007-2008 are dropped from the study.

The analysis sample is 58% white, 25% black, 10% Hispanic, 14% academically gifted, 14% with special education needs, 43% eligible for free or reduced priced lunch, and 29% with college educated parents.

Intervention / Program / Practice

The North Carolina accountability program, the ABCs of Public Education, first implemented in 1996 to 1997, awarded schools bonuses based on the annual achievement gains 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 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 actual 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 not achieving their expected growth were publicly identified and in some cases subject to intervention from the state. The program 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.

Research Design

The study employs a regression discontinuity design (RD), one of the only non-experimental research designs that if conducted appropriately can produce unbiased causal effects (Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007; Shadish, Cook, & Campbell, 2001). This design requires that the treatment assignment be either a deterministic or a probabilistic function of a continuous "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 was a continuous performance metric used by the state with a threshold that was the sole determinant of $750 bonus payments. Bonuses in 2007 were based solely on reading and math achievement. Treatment assignment was sharp in 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 complex formula by state assessment officials the summer following spring testing, there was little opportunity for manipulation around the cutoff. Consequently, there is no evidence of discontinuities in treatment assignment around the cutoff (see figure 2).

In 2008, science tests were administered to 5th graders for the first time and were not part of the 2008 growth calculations. They were therefore, "low stakes," unlike reading and math scores. Comparison of the incentive effects of the bonus program on reading and math with science is thus a comparison of incentive effects on high and low stakes tests.

Data Collection and Analysis

The study uses administrative records from the North Carolina Department of Public Instruction made available by the North Carolina Education Research Data Center at Duke University. 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. 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. The study reports effects at the optimal bandwidth, h^*, half optimal bandwidth and two times the optimal bandwidth. Rather than fitting a constant function, I fit local polynomial regression functions to the observations within a distance h^* on either side of the cutoff point (Imbens & Lemieux, 2008).

Bootstrapping the average treatment effect from nested data (students within schools) ignores the non-independence of clustered observations and would thus underestimate the size of the standard errors. I instead use the block bootstrap to produce standard errors, which preserves all within-cluster correlation. This approach randomly draws clusters and includes all within-cluster observations from each selected cluster in the bootstrap replication.

For the purpose of examining whether the bonus program has differential effects on students based on prior achievement, I estimate separate RD models for low, medium, and high achievers in each subject. The study defines students as low, medium, or high in reading prior achievement based on their position in the spring 2007 *within-school* reading test score distribution. Students defined as low fall below -.5 SD below the within-school mean, those defined as medium fall within the range -.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.

Results

Do bonuses induce incentive effects to increase math or reading test score gains?

Test score gains between 4th and 5th grade are higher in both reading and math for students in schools just missing the bonus threshold. Though math gains are of approximately the same size as reading gains, they are not statistically significant. Reading gains, which are statistically significant at the 95% confidence level, vary in size depending on bandwidth. At optimal bandwidth, the size of the discontinuity in reading gain is about a tenth of a standard deviation. At one-half this bandwidth, the size of the effect is one-fifth of a standard deviation (see table 1, panel A., and figures 3 and 4. Note: estimates in table are negative because effects are defined as the effect to the left of the cutoff minus the effect to the right of the cutoff).

Do bonuses promote "educational triage" based on the achievement level of the student?

Consistent with expectations, the bonus program produces relatively large, but imprecisely estimated, test score gains in math for low-achieving students in schools that just missed the threshold. For low achieving students, for example, gains are .16 SD at optimal bandwidth and .33 SD at one-half optimal bandwidth. By comparison, test score gains for high achieving students in schools barely missing the bonus threshold are .01 SD at optimal bandwidth and .02 SD at one-half optimal bandwidth. Because none of the math results are statistically significant, however, the evidence on differential effects on math gains is merely suggestive rather than conclusive (see table 1, panel B, and figures 7 and 8).

In contradiction to expectations, relatively large test score gains emerge for high achieving students in reading in schools barely missing the bonus threshold in the prior year. Estimates are estimated precisely with all results significant at the 95% confidence level. The size of the estimated effect varies by bandwidth. At optimal bandwidth, reading test scores increase by one-

fifth of a standard deviation for high achieving students; at one-half optimal bandwidth, test scores increase by almost one-third of a standard deviation. By contrast, test score gains for low achieving students in schools just missing the bonus threshold are .09 SD at optimal bandwidth and .16 SD (both non-significant) at one-half optimal bandwidth (see table 1, panel B., and figures 5 and 6).

Do bonuses promote a narrowing of the curriculum at the expense of science?
No evidence of discontinuities in science achievement scale scores emerge near the bonus threshold, which suggests that the bonus program does not narrow the curriculum at the expense of science. Complicating this conclusion somewhat is the fact that no evidence of discontinuities emerge in math or reading achievement scale scores either. Because it is not possible to compute *gains* in science scores, it is not possible to compare the incentive effects of the bonus programs on science, reading, and math gains (see table 1, panel C, and figure 9).

Conclusions
The study finds evidence consistent with the hypothesis that educators in North Carolina respond to incentives to increase test score gains in reading and math. Those students in schools that just missed the bonus threshold in 2007 have higher test score gains in 2008. This suggests that educators expend additional effort and may implement new practices in response to the failure to receive a bonus. I find suggestive, but not conclusive, evidence that math gains are primarily driven by low and average achieving students.

Contrary to expectations, reading gains are disproportionately driven by students with the highest within-school achievement. This suggests that either schools targeted high achieving students with reading interventions, which is unlikely, or that schools used whole-school interventions that had positive effects on high achievers and no effects on low achievers. This finding deserves future research into its generalizability across different time periods and investigation of the mechanisms through which this differential effect was produced.

I find no evidence of a narrowing of the curriculum at the expense of science. This is in contradiction to theory and prior research on a "narrowing of the curriculum" at the expense of low-stakes and non-tested subjects. The fact that the policy is focused on test score gains, rather than levels, however, raises questions about whether incentive effects on test score levels should be expected. That North Carolina's bonus policy had no effect on test score levels may be viewed as a shortcoming of the policy if absolute, rather than relative, levels of performance are also of interest.

This study takes place in the post-NCLB era and at a time when the bonus program was relatively mature. Further research will determine whether the incentive effects were stronger in the early years of the program and before NCLB. Moreover, I examine only the effects of the bonus program on test score achievement. A full evaluation, which is beyond the scope of this study, would take into account other important outcomes, such as teacher turnover, and weigh the costs of the program against the benefits.

Appendices

Appendix A. 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.

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. doi: DOI 10.1016/j.jpubeco.2009.06.002

Darling-Hammond, L. (2004). Standards, accountability, and school reform. *Teachers College Record, 106*(6), 1047-1085.

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. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics, 90*(1-2), 239.

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. A., & Raymond, M. A. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297-327.

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.

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., & 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.

Linn, R. L. (2000). Assessments and Accountability. *Educational Researcher, 29*(2), 4-16.

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, 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.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W., & Shavelson, R. (2007). *Estimating Causal Effects Using Experimental and Observational Designs*. Washington, DC: American Educational Research Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

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.

Appendix B. Tables and Figures

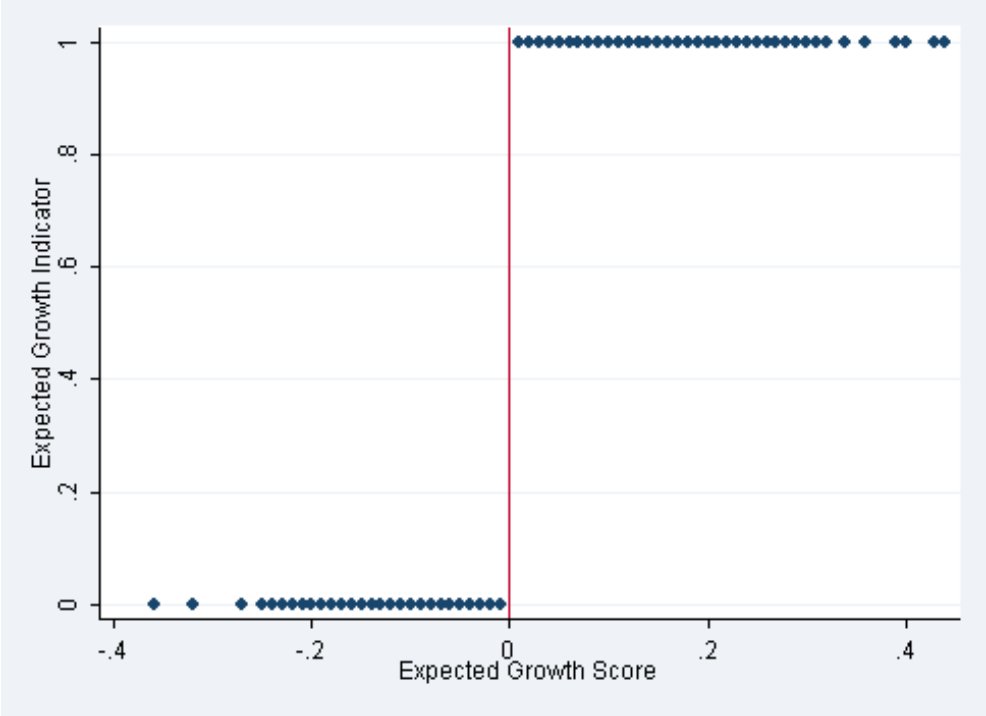

Figure 1.

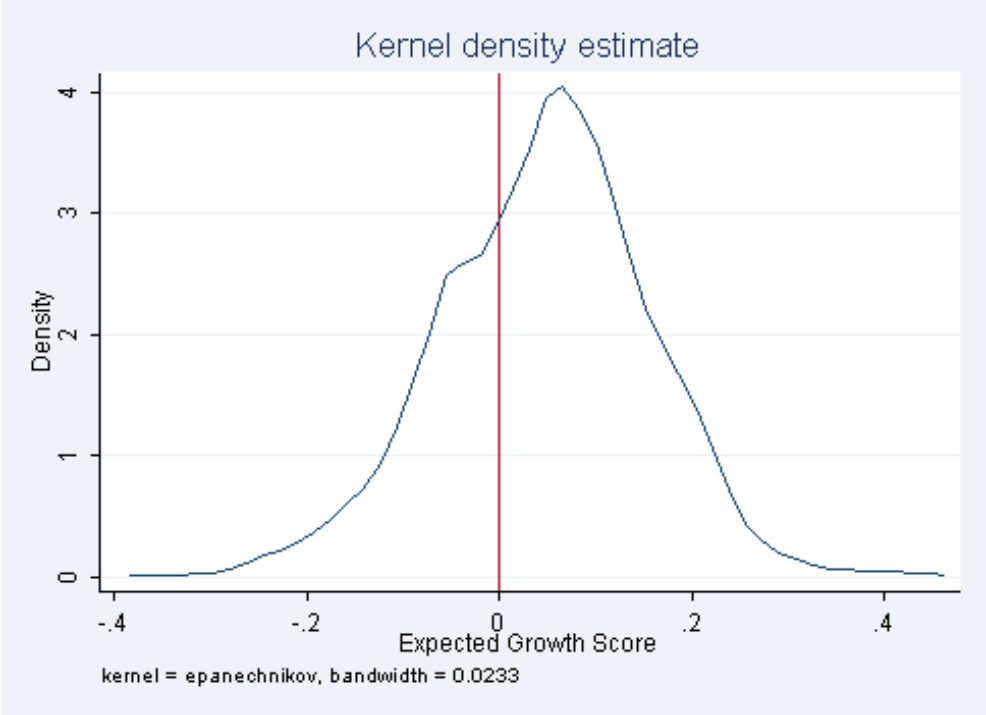

Figure 2.

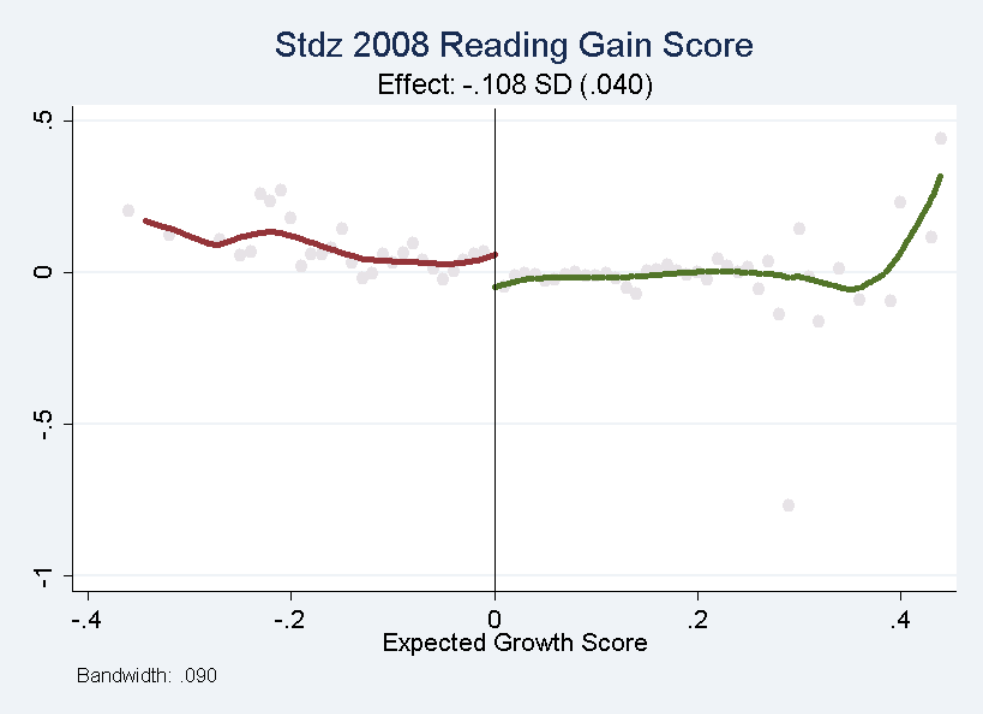

Figure 3.

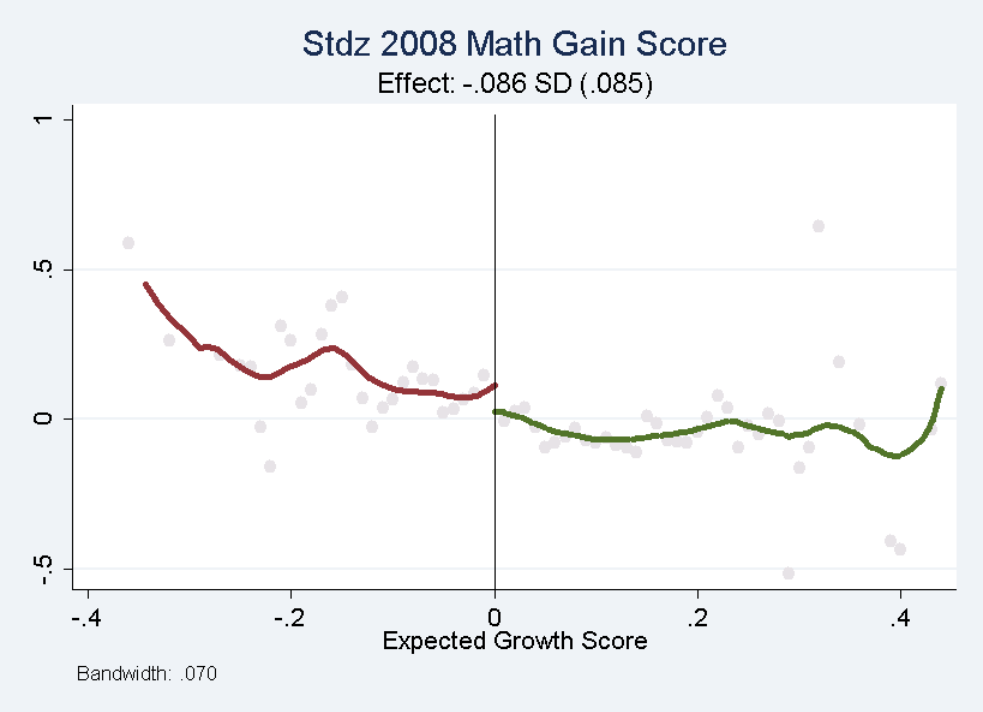

Figure 4.

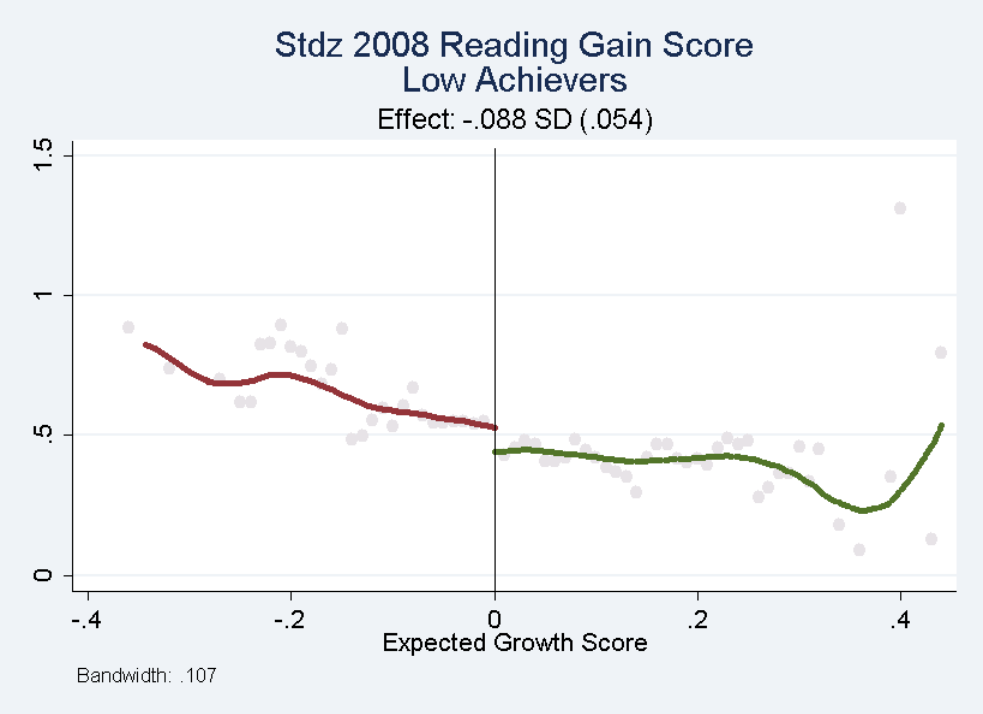

Figure 5.

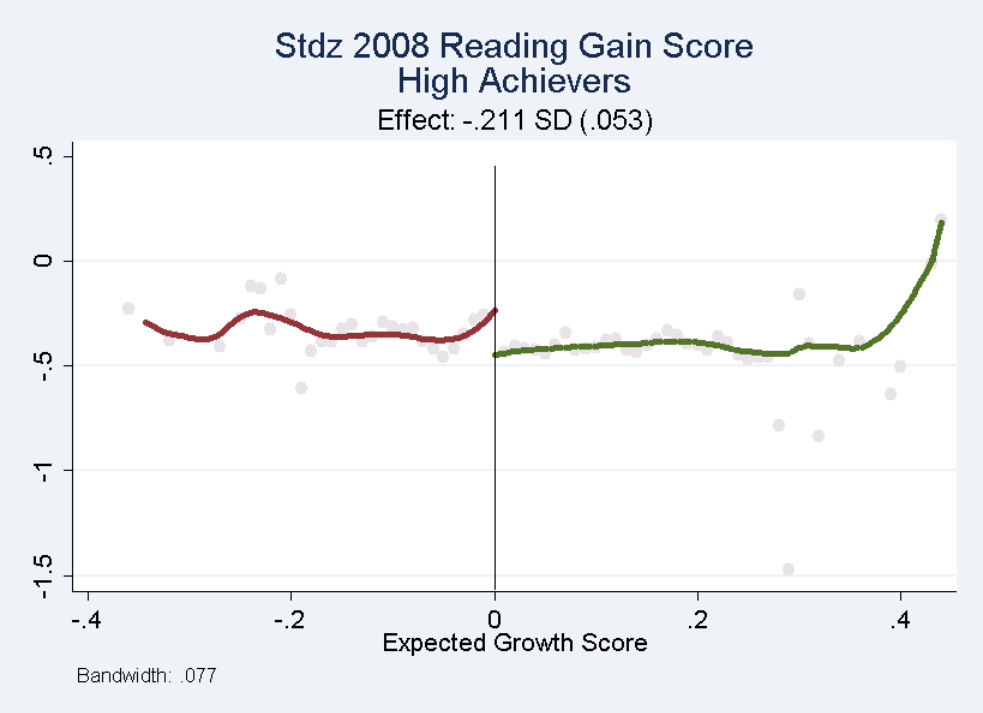

Figure 6.

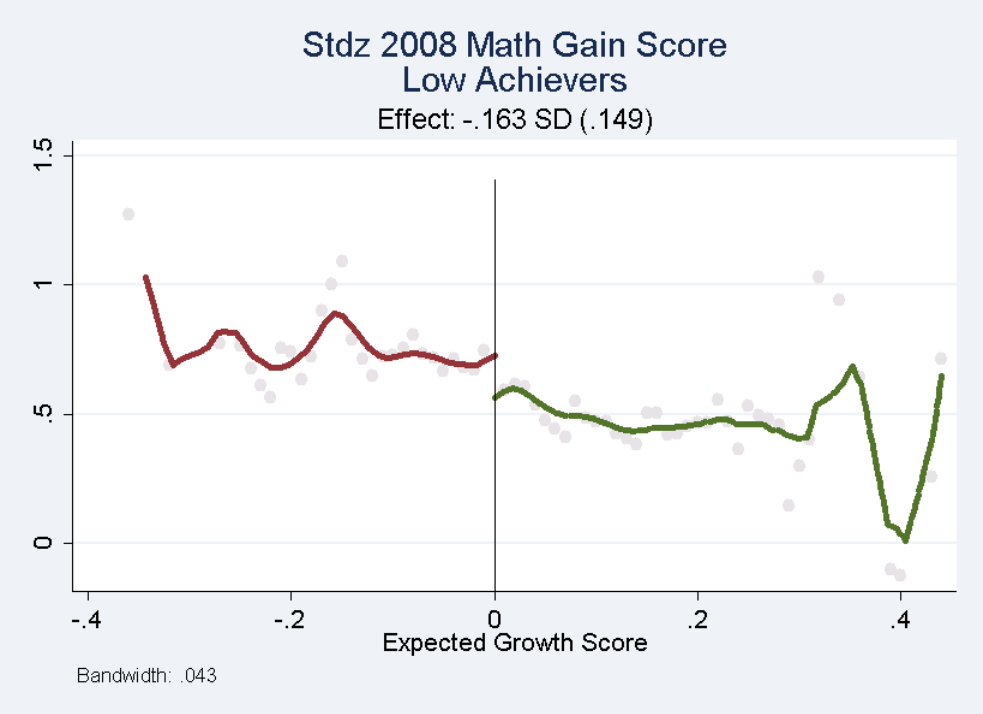

Figure 7.

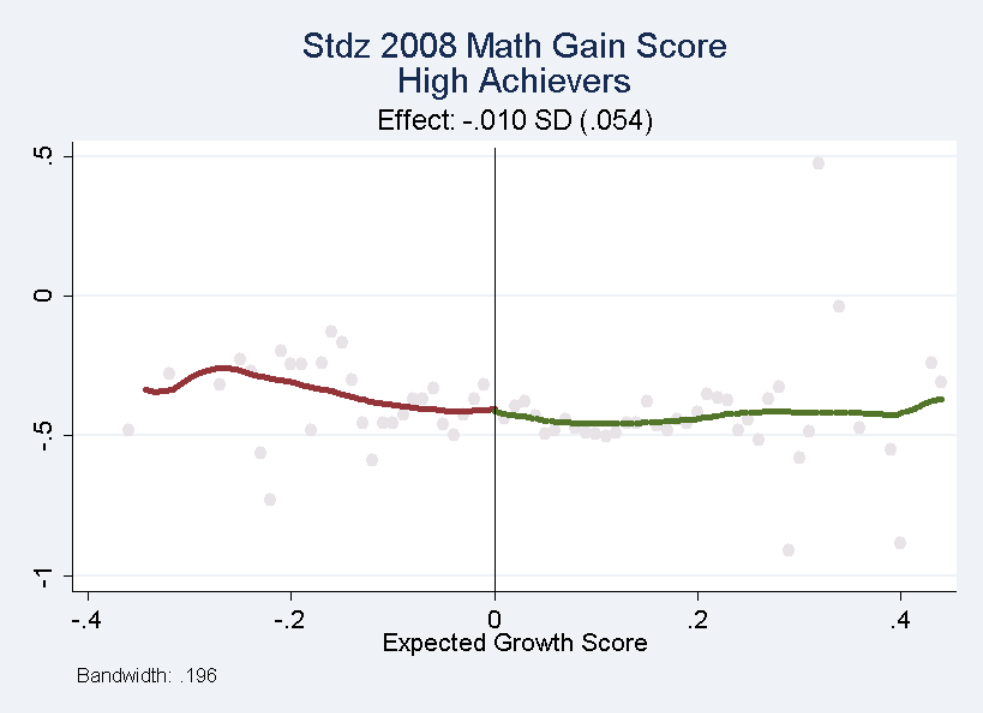

Figure 8.

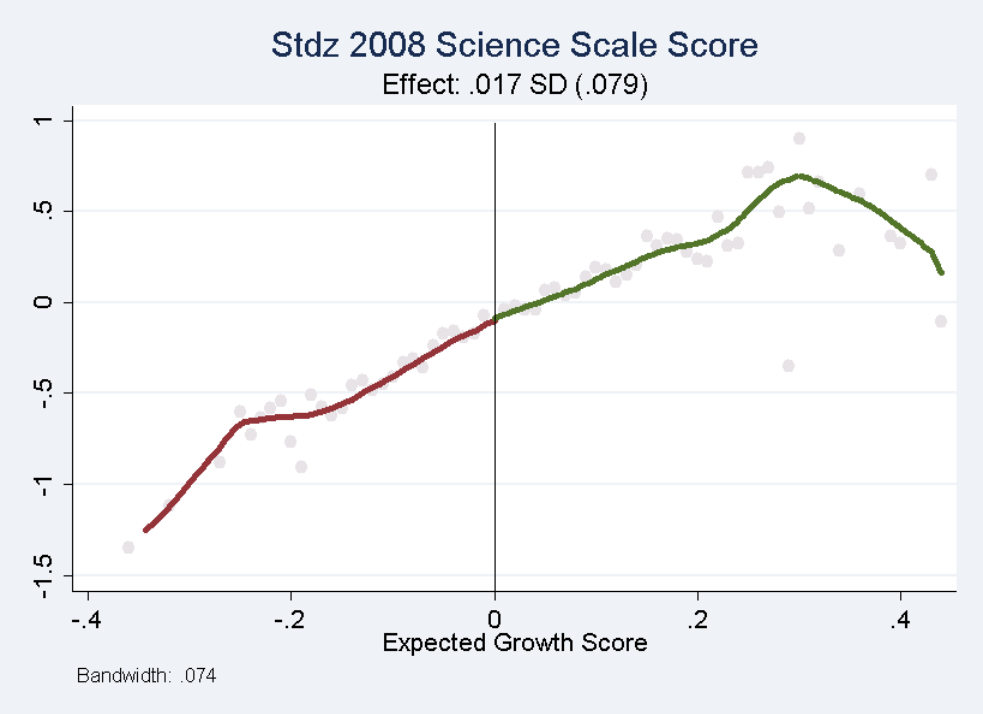
Stdz 2008 Science Scale Score
Effect: .017 SD (.079)
1
.5
0
-.5
-1
-1.5
-.4
-.2
0
.2
.4
Expected Growth Score
Bandwidth: .074

Figure 9.

Table 1. Incentive Effects of Bonus Program, North Carolina Fifth Graders

A. Effect of bonus on reading and math gains

	Bandwidth	Estimate	1/2 BW Est	2*BW Est
Reading gain	0.090	**-0.108**	**-0.196**	**-0.067**
		0.040	0.062	0.029
Math Gain	0.070	-0.086	-0.224	-0.055
		0.085	0.145	0.059

B. Effect of bonus on reading and math gains by prior achievement

	Bandwidth	Estimate	1/2 BW Est	2*BW Est
Reading Gain				
Low	0.107	-0.088	-0.155	-0.075
		0.054	0.082	0.041
Medium	0.109	-0.040	-0.108	-0.026
		0.044	0.068	0.031
High	0.077	**-0.211**	**-0.312**	**-0.110**
		0.053	0.082	0.037
Math Gain				
Low	0.043	-0.163	-0.333	-0.043
		0.149	0.240	0.088
Medium	0.059	-0.168	-0.256	-0.064
		0.118	0.236	0.072
High	0.196	-0.010	-0.020	-0.017
		0.054	0.071	0.044

C. Effect of bonus on math, reading, and science scale scores

	Bandwidth	Estimate	1/2 BW Est	2*BW Est
Reading scale score	0.067	-0.006	-0.123	-0.024
		0.071	0.116	0.046
Math scale score	0.148	-0.014	0.013	-0.012
		0.046	0.067	0.036
Science scale score	0.074	0.017	-0.055	0.001
		0.079	0.138	0.054

Estimates from regression discontinuity models specified with local polynomial regressions. Standard errors estimated with block bootstrap, 400 reps. Estimates in bold statistically significant at the 95% confidence level. Bandwidth estimated using the technique proposed by Imbens and Kalyanaraman (2009). Unit of analysis: students. N=78,000 to 79,000. 1,269 schools.