How salient are performance incentives in education?
Evidence from North Carolina

Thomas Ahn[*]

University of Kentucky

Jacob L. Vigdor

Duke University and NBER

Preliminary Draft

February 2012

Abstract

Since the mid-1990s, North Carolina has offered teachers performance bonuses of up to $1,500 when test score gains in their school exceed a predetermined threshold. We use regression discontinuity (RD) methods to show that schools posting gains just below the threshold in one year exhibit supernormal gains in the following year. This result is actually puzzling from a rational expectations perspective, as schools on both sides of the discontinuity should expect similar returns to effort. It instead supports a more behavioral model, which when generalized can potentially explain recent null results in experimental evaluations of performance incentives.

[*] Ahn: Thomas.ahn@uky.edu; Vigdor: Jacob.vigdor@duke.edu. The authors gratefully acknowledge financial support from the Institute for Education Sciences, award #R305A090019. Vigdor further acknowledges support via the Center for the Analysis of Longitudinal Data in Education Research (CALDER). Any opinions expressed herein are those of the authors and not any affiliated organization.

Introduction

Performance incentives, long a staple of employment contracts in certain industries, have risen to prominence in K-12 education. Critics of traditional seniority- and credential-based compensation systems argue that experience and credentials correlate poorly with on-the-job performance, and thereby encourage wasteful resource allocation (Hanushek 1989; Lazear 2003; Vigdor 2008; Grissom and Strunk 2011). Performance incentives typically focus on the outcomes of standardized test scores. While the focus on test scores has invoked criticism regarding potential unintended consequences (Jacob and Levitt 2003; Clotfelter, Ladd, Vigdor, and Aliaga 2004; Figlio and Winicki 2005), simple models of teacher behavior clearly predict increases in test scores (Lazear 2001). In light of these predictions, recent evaluations of performance pay initiatives have found surprising results – eligibility for bonus payments has not been associated with significant improvements in student test scores.

This paper contributes to the literature on teacher responses to, and educational effects of, performance incentives by analyzing the impact of North Carolina's ABC accountability program[1]. Since the mid-1990s, the ABC program has rewarded schools that achieve student test score gains above a discrete predetermined threshold, with bonus payments of up to $1,500 per teacher (Vigdor, 2009). Regression discontinuity methods reveal that schools with test score gains just below the bonus threshold exhibit significantly higher gains in the following year, relative to schools just above the threshold.

Ours is not the only paper to document the discontinuity effect: Vigdor (2009) and Jinnai (2012) document it as well. This central result is actually difficult to explain with a traditional

[1] The acronym stands for strong Accountability, teaching the Basics, and emphasis on local Control.

rational expectations model, however. Schools on either side of the threshold should anticipate roughly equal chances of being above the threshold in the subsequent year, and should therefore expect similar returns to effort. This pattern is instead more consistent with a form of rational ignorance model, where school personnel face significant costs associated with understanding how the incentive system works and how to respond to it, and incur those costs only if they receive a clear signal that the investment has the potential to pay off.

Further exploration of the results reveals more evidence that incentives are imperfectly understood and most salient in only a subset of the population. For math test scores, in spite of the fact that the North Carolina system rewards test score gains rather than proficiency, the discontinuity effect is strongest among students just above or below the state's proficiency threshold, and insignificant among students at both the high and low ends of the test score distribution. We conclude that teachers focus effort here because they believe that improving the performance of these students in particular will make it more likely to qualify for the bonus next year. For reading test scores, the impact is more concentrated at the high end of the score distribution, suggesting that parents of high achieving students are compensating for perceived academic deficiencies of the school.

For both reading and math scores, the discontinuous response is much stronger among schools with a recent history of failing to receive a bonus. Moreover, it is an order of magnitude larger in schools with a poor track record in the independent Federal No Child Left Behind accountability system, relative to schools that had consistently met the federal standard.[2]

[2] Again, it is worth noting that the NCLB criterion (which is based on absolute proficiency rate in the school) and the ABC criterion (which is based on test score growth) are not strongly correlated.

Our results offer an interpretation for the null results found in recent pay-for-performance evaluations in other parts of the country (Springer et al. 2010). Given the effort required to understand an incentive scheme and craft an optimal response to it – particularly given that the scheme offers no specific guidance as to how to improve test scores – teachers and other school personnel are unlikely to contemplate a change in behavior unless they receive a clear signal that such a change is quite likely to produce material gains. This implies that a period of learning must take place before any incentive system actually influences behavior, and that incentive systems that fail to provide useful signals may result in no behavioral change whatsoever.

Models of Employee Response to Incentives

Consider a standard principal-agent framework. Output y_{it} of employee i in period t, which in the context of educational production can be measured by test scores, is a function of an employee's ability, which we take to be a permanent characteristic a_i, time-varying effort level e_{it}, and an idiosyncratic shock η_{it}. The employee's utility is a function of their wage, w_{it}, and a cost function based on effort, $c_i(e_{it})$, which we take to be increasing and convex in its argument. We also allow for the possibility that $c_i(e_{it})$ may be less than zero for low levels of e_{it}, which would be the case if employees received some satisfaction from turning in a certain level of effort even in the absence of monetary reward. The subscript also indicates that there may be permanent differences across teachers in the valuation of effort.

The traditional teacher employment contract offers a salary w_{it} that does not vary with e_{it}, but rather with a set of credentials that can be considered crude proxies for a_i.[3] Under this contract, teacher effort remains at a corner solution at the point where $c_i(e_{it})$=0, traditionally at zero but possibly at some level $\underline{c}$ determined as the point where the cost of effort transitions from below to above zero.

To incentivize effort, the employer links salary to the observed indicator of output, $w_{it}(y_{it})$. In this scenario, the employee's optimal choice of effort equates the expected marginal cost and benefit. The anticipated effect of the incentive scheme on effort thus depends on the strength of the relationship between output and effort, and the strength of the relationship between output and the wage. In the case of teaching, a less stylized model would relax the assumption of a single-dimensioned effort input; the actions taken to educate a student most can in fact vary along many dimensions. Imperfect knowledge of the production function in this scenario might imply a weak relationship between a summary measure of effort and output.

Consider the special case when the incentive payment is binary: w_{it} is incremented by some positive amount when output rises above a critical threshold. This case corresponds to many incentive pay programs for teachers, including the North Carolina program studied here. The expected marginal benefit to effort then reduces to the marginal impact of effort on the probability of pushing the output indicator above the critical value.

Now, consider a pool of identical employees who have optimally chosen effort according to the same rules. Any variation in compensation across these teachers reflects

[3] Teachers earn pay increases by accumulating experience, earning advanced degrees, and getting certification. We abstract away from the dynamic nature of these human capital investment decisions.

variation in η_{it}. Were these error terms truly independently and identically distributed, and presuming no significant change in the production function, cost functions, or incentive program, we would expect each teacher to continue to follow the same decision rule in year t+1. If η_{it} were serially correlated, teachers experiencing a positive shock would adjust their behavior relative to those who experienced a negative shock. Among teachers whose output was arbitrarily close to the output threshold, though, we would expect similar values of η_{it}, which would yield minimal variation in adjustment behavior.

To apply the principal-agent model in this context, we must think of the agents not as individual teachers but rather entire schools. Presuming that the behavior of sub-agents within the school can be aggregated to the level of the school, the same logic applies.

Schools and evolution of the production process

Reformulating the notation above, schools employ teachers with predetermined stocks of ability who exert effort. Their effort and ability, coupled with student-level characteristics that can be considered a component of η_{it} (along with truly idiosyncratic factors), are transformed into knowledge according to the production process y(.).

Suppose that technological change and other factors lead to improvements in the production process over time. The production frontier defined by y(.) can be thought of as a subset of the optimal production process y*(.), which moves steadily outwards. Alternatively, y*(.) could be thought of as a production process obtained with an optimal stock of instructional capital, and y(.) the process that results when an originally optimal process degrades through depreciation of a capital stock.

At any given point in time, we presume that schools can adjust their production process to more closely approximate y*(.). The reoptimization is not costly, however, and we presume that the potential benefits of reoptimization are unknown to the school a priori. Under complete information, the school's optimal decision rule would be to pay the fixed cost of reoptimization when the expected benefits exceed the costs, but under incomplete information and uncertain rates of technological change (or depreciation), the optimal rule is harder to characterize.

The introduction of a bonus program with a discrete cutoff might induce schools to engage in reoptimization under certain circumstances. Schools who discover their performance is within a narrow range of the cutoff can expect larger returns to improving their production process, other things equal.

This logic by itself does not suggest that the effect of proximity to the threshold should be asymmetric. An asymmetry could be introduced under the following scenario: schools receive the binary signal of whether the bonus threshold has been reached, but cannot recognize their proximity to that threshold without incurring some supplemental cost. Presuming the cost is in some intermediate range, only those schools failing to receive the bonus will choose to incur it. Intuitively, schools that receive the bonus receive a signal that with high probability they will continue to receive the bonus in subsequent years without altering their production process. Schools that fail to receive the bonus elect to incur a modest initial cost to determine their proximity to the threshold – the expected gains from learning that one is close to the threshold must exceed the initial cost. Those who discover they are proximate to the cutoff then elect to incur the additional costs of reoptimization.

Figure 1 shows the basic prediction of the behavioral model under the assumptions outlined above. At the end of period *t*, teachers receive a binary signal based on their output in that period, which determines whether they pay the fixed cost of learning the optimal effort level. For teachers reasonably close to the bonus threshold, there are strong returns to increasing effort. For those at greater distance from the threshold, there is little chance of receiving the bonus payment regardless of effort level, so the optimal response is to continue providing the default level. Teachers who are at schools that qualify for the bonus are not incentivized to acquire information about 'how close' they were to the threshold, resulting in the default amount of effort next year.

The decision to learn about the incentive program and the optimal response to it need not be binary, and the signals triggering the learning process need not be as simple as one year's success or failure. A school with a track record of receiving the bonus may dismiss a negative outcome in a single year as a statistical fluke, for example. This model merely provides some intuition for why teachers' response to information might be discontinuous around a point.

Setting, Data, and Methods

Setting

Beginning in the 1996/97 school year, the state of North Carolina implemented the ABCs of Public Education accountability plan, which introduced a system of cash bonuses awarded to all teachers in schools meeting test score-based performance goals. Initially, the bonus amount was set to $1,000 per teacher, but after one year the state switched to a two-tiered bonus structure, with payment amounts of $750 and $1,500. The performance measure used to assess schools was based on year-over-year changes in test scores for enrolled students, which makes the program distinct from the Federal No Child Left Behind (NCLB) program or other incentive schemes based purely on proficiency rates. The formula for computing the performance measure changed after the 2004/05 school year; our analysis below focuses on the measure in place during the more recent period.

Details regarding the computation of the performance measure can be found in Vigdor (2009). Importantly, a bonus of $750 per teacher was awarded if the school's measure exceeded a predetermined threshold, and a $1,500 bonus awarded in schools where the measure exceeded a second, higher threshold. This implies that the effect of being awarded a bonus (or of failing to receive a bonus) can be estimated with a regression discontinuity design.

From the 2002/03 school year forward, the NCLB program imposed a simultaneous but distinct set of requirements and sanctions upon public schools in North Carolina. Because these sanctions were based on student proficiency rates, and not test score growth, the correlation between qualifying for positive sanctions – bonus receipt in the state system, Adequate Yearly Progress (AYP) in NCLB – is modest. Table 1 shows a cross-tabulation of AYP status and bonus receipt for school years2005/06 and 2006/07. Over 40% of schools qualify for some bonus payment even though they have failed to make AYP, and about 30% receive no bonus in spite of the fact they have made AYP.

Data

We use individual-level test score data provided by the North Carolina Education Research Data Center (NCERDC) to analyze the differences in student performance on either side of the bonus discontinuity.[4] The NCERDC data provide longitudinal links for students in grades 3-8, based on standardized test score records. We use these records to compute individual-level gain scores. We also observe a range of demographic and socioeconomic indicators at the individual level, including race, gender, free/reduced price lunch participation, and parental education.

Table 2 presents summary statistics for our analysis sample, which consists of students enrolled schools with grades 3-5 in the 2005/06 and 2006/07 school years. [5] North Carolina is a racially and socioeconomically heterogeneous state, with a rapidly growing immigrant population and a mix of prosperous metropolitan areas and poorer rural and inner-city regions.

The math and reading gain scores are computed by subtracting a student's prior year standardized math or reading score from his or her prior year's standardized score in the same subject. For both math and reading, the average test score gain is close to zero – as expected, since it is impossible for all students to simultaneously improve their relative standing in the test score distribution. The standard deviation of the gain score is 0.46 for math and 0.64 for reading.

[4] The NCERDC data are available to researchers with an approved IRB protocol from their home institution, conditional on registration to use the data.

[5] The set of schools that are considered are schools with grades capped at 5. Schools that contain both middle school grades (Gr. 6, 7, and/or 8) and elementary school grades are excluded from the analysis. Because students in these upper grades may move classes and teachers from subject to subject, the teacher utility maximization problem is significantly complicated.

We couple these individual-level data with official school-by-year records from the state's Department of Public Instruction. These record the official value of the composite growth index used to determine bonus eligibility, along with a few other school-level summary statistics. This growth score ranges from -0.45 to 0.66, with the school qualifying for the $750 bonus if it scores above 0.0.[6]

Methodology

Our basic goal is to determine whether students on opposite sides of the bonus eligibility threshold experience different test score gains in the following academic year, using regression discontinuity (RD) analysis. There are two basic forms of RD, parametric and non-parametric. In both varieties, the outcome is modeled as a smooth function of the assignment variable, with the possibility of a discrete jump at the threshold point. We focus here on the non-parametric variety (Imbens and Lemieux 2008), based on local linear regression. The local linear regression provides a slope coefficient that is unique to each data point, and is based on the OLS regression coefficient derived from data points within a certain bandwidth. While it is not necessary to specify a functional form using this method, a bandwidth must be selected. As the bandwidth increases, the local linear regression approaches a simple linear model; small bandwidths permit a greater number of inflection points in model fit. We use the general

[6] Lending credence to our assertion that it takes effort to understand the incentive scheme and construct a best response, we were unable to perfectly duplicate the state's growth scores using the individual-level data. In addition, while North Carolina has been making statistical information available on the web since before the ABC program was in place, the growth scores were only made public for the 2005/06 and 2006/07 school years.

convention of specifying multiple bandwidths to gauge sensitivity of our results. Standard errors for the RD estimates are generated using bootstrapping.[7]

To attach a causal interpretation to RD estimates of the difference in test score growth on either side of the bonus discontinuity, we must verify a series of assumptions that underlie the method. First, we need to check for evidence that schools are able to manipulate their assignment variable so as to place themselves on the more beneficial side of the discontinuity. Schools clearly have an incentive to qualify for bonus payments, but it is not clear that this can translate into ex post manipulation, as schools are generally not aware of their test score results until the outcome of the bonus determination process is announced. Second, we need to check for balance in covariates on both sides of the discontinuity. Third, we need to verify that there is in fact a discontinuity – that schools on either side of the eligibility threshold were in fact differentially likely to receive a bonus.

Figure 2 shows the distribution of the average growth performance measure across all school-year observations in school years 2005/06 and 2006/07. The bonus threshold is at zero, implying that students' test score improvements were in line with expectations. The peak of the distribution falls just to the right of the bonus threshold. There is no evidence of bunching just above or below the bonus threshold.

Figures 3 and 4 show results from 'placebo' regression discontinuity analysis with school minority percentage and free/reduced lunch percentage as the 'outcome' variables. As expected, there is no treatment effect of the discontinuity on the demographic distribution of students. This lends support to the assertion that the impact on test score growth at the

[7] See Nichols (2009) for details.

discontinuity is driven by the policy itself, and not sharp differences in student characteristics at schools that either just fail or just succeed in qualifying for the bonus.

In addition, we note that there is a negative relationship between the ‘outcome variables’ and average growth score, indicating that school that perform better have lower proportions of minority and free/reduced price lunch students, as expected.

Figure 5 shows teachers’ bonus receipt as a function of the average growth score we are using as the assignment variable. It is clear that there is a sharp discontinuity in probability of bonus receipt (from zero to one) at zero average growth. Teachers to the right of the discontinuity receive a bonus of at least $750. There is an additional fuzzy discontinuity around 0.1 to 0.2 in average growth, above which teachers receive $1,500.[8]

Results

Figure 6 presents a graphical representation of our most basic RD estimates, and table 3 reports the associated effects and bootstrapped standard errors. In the case of math scores, our estimates indicate that students in schools just below the bonus eligibility threshold exhibit higher test score gains relative to students in barely-eligible schools. The estimated effect is fairly robust to bandwidth choice, ranging from 0.0175 to 0.02 with higher point estimates in models with narrower bandwidths. In itself, this is not a very dramatic improvement in test scores from one year to the next – previous studies have, for example, attributed much larger

[8] We attempted to incorporate this second discontinuity in a previous version of the paper. However, the results at the $1,500 discontinuity were insignificant for the most part. The fuzziness around the second bonus threshold is attributable to the fact that eligibility for the higher bonus amount is the function of two variables, not just the composite growth score.

test score gains to the first years' worth of teacher experience. Figure 6 shows, however, that this improvement is quite meaningful for schools in close proximity to the bonus threshold.

For reading scores, point estimates are more sensitive to bandwidth choice, and once again display the pattern that larger point estimates associate with narrower bandwidths. The point estimates shown range from 0.005 and insignificant to 0.032 and significant beyond the 1% level. The association of larger effects with narrower bandwidth is consistent with an incentivization effect that is highly localized to the area immediately adjacent to the discontinuity. In light of the model above, this proves to be a rational interpretation of the results. Schools to the left of the border derive information from their failure to receive the bonus, and they invest effort in learning about the program and optimizing their behavior. Behavioral optimization leads to relatively modest gains, so only schools in very close proximity to the point of discontinuity engage in reoptimization.

This explanation is admittedly a post-hoc rationalization of what would otherwise appear to be a puzzling result. To gain further insight, we examine variability in the estimated treatment effect across varying types of schools and students.

Effect heterogeneity across students

Previous literature has established that schools under accountability pressure place particular emphasis on certain students. In a system that rewards proficiency, teachers have a clear incentive to focus on students who are very close to the proficiency threshold, and prior work has verified that such an effect exists (Neal and Schanzenbach, 2010). In tables 4, 5, 6, and 7 we test whether the boost to test scores created by failure to receive a bonus accrues disproportionately to certain types of students. Note that since the North Carolina system rewards growth, rather than proficiency, the incentive to focus on students just below the threshold is rational only if those students can be expected to produce the strongest test score gains, which is not clear *a priori*.

We find clear evidence of heterogeneous effects across students in different categories. Table 4 shows that the effects on math scores are concentrated among minority and low-income students. By contrast, Table 5 indicates that reading score effects are if anything concentrated among more advantaged students. Prior research suggests that substitution between school and family inputs is easier in the domain of reading. Thus one interpretation of the evidence is that schools reoptimize by devoting more attention to rudimentary math at the expense of reading (and higher order math). Some parents respond by substituting for the withdrawn inputs with their own, but find that they are effective only in the domain of reading.

Table 6 shows that math scores among students categorized as near proficiency level on the state's four-tiered scoring scale in the preceding year increase at the discontinuity.[9]. In the narrow-bandwidth specification, the estimated effect is quite large for students just below proficiency: relative to students in schools just above the bonus threshold, these students gain one-tenth of a standard deviation. There are more modest, but significant, effects among those students classified as "proficient," but little to no evidence of discontinuity effects among students classified as "advanced" or "below basic."

[9] Students are considered proficient in the subject if they score level III and above. Again, it bears emphasizing that the cut off defining proficiency is unrelated to the discontinuity at zero average growth score.

While it is clear that teachers from schools that just failed to qualify for the bonus respond substantively, the apparent focus on students near the proficiency level suggests that teachers may not fully understand the North Carolina bonus program. Thus to interpret the behavior of schools just below the bonus threshold as “optimizing” is in some sense generous; evidence suggests that these schools identify ways of improving their performance, but do not specifically tailor their efforts to the structure of the state’s incentive program. These results amplify concerns that a nationwide focus on proficiency has come at the expense of instructional attention to highly advantaged and disadvantaged students (Neal and Schanzenbach 2010).

The results in Table 7 echo those in Table 5: for reading scores, the strongest discontinuity effects are among the most advanced students. This is consistent with parent substitution compensating for a shift towards math instruction. The math and reading results are also consistent, however, with a scenario in which the easiest test score gains occur in the middle of the math test score distribution and the high end of the reading test score distribution.

Effect heterogeneity across schools

The model outlined above suggests that school personnel act to assess and reoptimize their behavior only in the presence of a signal that such activity will yield dividends. Results to this point suggest that failure to receive a bonus might serve as such a signal, and that schools within a narrow band short of the bonus threshold discover that reoptimization may be sufficient to push them into the eligible category. In this section, we consider whether this signal might be interpreted differently across schools.

Tables 8 and 9 show RD estimates for subsets of schools, divided according to their past performance in both the North Carolina and NCLB incentive systems. For both math and reading scores, the first panel divides schools according to their bonus receipt patterns over the five years prior to the year of assessment. One might expect that schools with a strong track record of bonus receipt would attach less weight to the signal – inferring that their failure to receive a bonus was an aberration not requiring corrective action.

In fact, this is exactly what we find. Looking at math score results in table 8, the discontinuity effect is eight times larger in schools that received no bonus in the majority of years than it is in schools that tended to receive the bonus consistently. If we follow a literal interpretation of the model, infrequently rewarded schools had prior opportunities to assess their standing and potential for re-optimization. As suggested above, however, they may have had little reason to re-optimize if their prior assessments had informed them that there would be little reward to doing so. By contrast, following a year in which the school just missed the threshold, the optimal response might well involve a change in behavior.

A nearly identical pattern emerges when we stratify schools by their prior frequency of meeting the AYP provisions under NCLB. The estimated discontinuity effect is nearly eight times larger among schools that had failed to make AYP more than half of the prior four or five year period. The AYP result is interesting in part because schools facing to make AYP multiple times over a period of several years are likely to be experiencing negative sanctions ranging from offering students transfers to school reorganization – whether they qualify for bonus payments or not. These results suggest that the NCLB sanctions are in general insufficient to generate the type of re-optimization witnessed among schools in close proximity to the bonus threshold.

Although not as strong as math results observed in table 8, reading results in table 9 follow much the same pattern, where consistent failure under either NCLB or ABC is taken more seriously as a signal by the school and teachers, resulting in a stronger response.

Implications for school incentive schemes

Complicated incentive schemes are unlikely to provoke immediate responses from teachers or schools. In equilibrium, response is triggered by an event signaling potential gains from reoptimization, and even so, many schools may conclude that no change is warranted if the expected gains are slight.

For incentive policies such as the North Carolina ABC bonus program, this is a real concern, as funding and administering of the program cost the state over $90 million in the 2006/07 school year. If the majority of schools and teachers do not change their behavior in response, the incentive policy is an inefficient use of funds.

From our analysis above, it is clear that incentive schemes must be designed with the typical behavioral response of teachers in mind. That is, to get the most bang for buck, we must either make the incentive scheme easier to understand and simpler to evaluate one’s own performance (so that there are fewer costs associated with assessment and crafting an improvement strategy), or structure it so that more schools will fall just below the threshold and hence expect tangible benefits from reoptimization.

Another reasonable response to these results would be to favor performance assessments that incorporate some form of specific feedback regarding how a school might improve. Our results suggest that principals can typically only identify strategies for realizing modest improvements. An assessment scheme that incorporated actual feedback regarding strategies for improvements could yield substantial benefits at a cost comparable to that of awarding cash bonuses to roughly half the state's instructional personnel (Taylor and Tyler 2011).

References

Figlio, D.N. and J. Winicki (2005) "Food for thought: The effects of school accountability plans on school nutrition." *Journal of Public Economics* v.89 pp.381-94.

Grissom, J.A. and K.O. Strunk (2011) "How Should School Distircts Shape Teacher Salary Schedules? Linking School Performance to Pay Structure in Traditional Compensation Schemes." *Educational Policy* preprint.

Hanushek, E.A. (1989) "The Impact of Differential Expenditures on School Performance." *Educational Researcher* v.18 n.4 pp.45-62.

Imbens, G.W. and T. Lemieux (2008) "Regression discontinuity designs: A guide to practice." *Journal of Econometrics* v.142 pp.615-635.

Jacob, B. and S. Levitt (2003) "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* v.118 pp.843-877.

Lazear, E.P. (2001) "Paying Teachers for Performance: Incentives and Selection." Stanford University mimeo.

Lazear, E.P. (2003) “Teacher Incentives.” Swedish Economic Policy Review v.10 pp.179-214.

Neal, D. and D. Schanzenbach (2010) “Left Behind By Design: Proficiency Counts and Test-Based Accountability.” *Review of Economics and Statistics* v.92 pp.263-283.

Nichols (2009) “Causal Inference with Observational Data: Regression Discontinuity and Other Methods in Stata.” Stata Users Group.

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCafrey, D., Pepper, M., and Stecher, B. (2010). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives at Vanderbilt University

Taylor, E.S. and J.H. Tyler (2011) “The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers.” NBER Working Paper #16877.

Vigdor, J.L. (2008) “Scrap the Sacrosanct Salary Schedule.” Education Next.

Vigdor, J.L. (2009) “Teacher Salary Bonuses in North Carolina.” In M.G. Springer, ed., *Performance Incentives: Their Growing Impact on American K-12 Education*. Washington: Brookings Institution Press.

Figures:

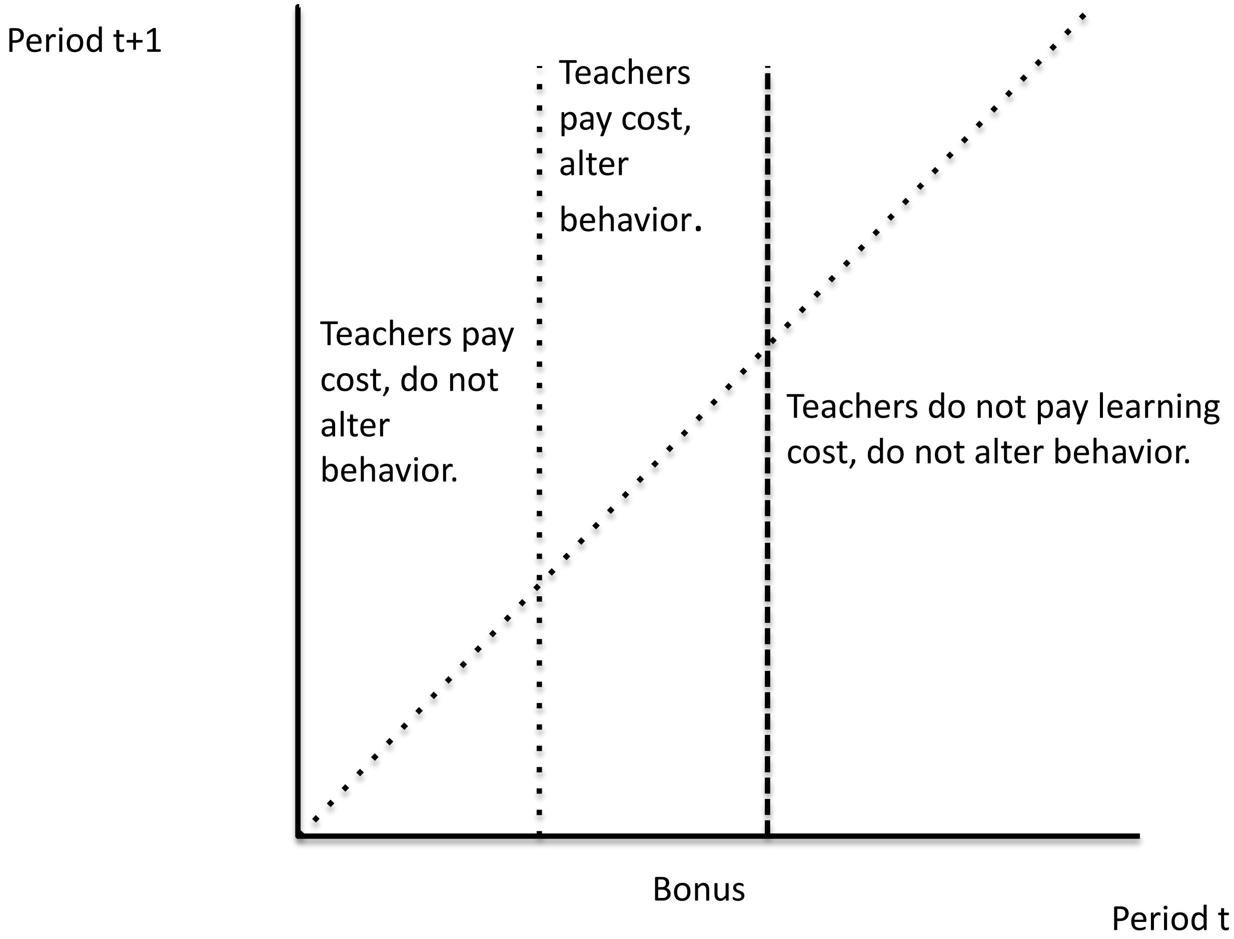
Period t+1
Teachers pay cost, alter behavior.
Teachers pay cost, do not alter behavior.
Teachers do not pay learning cost, do not alter behavior.
Bonus
Period t

Figure 1: responses to signal with costly learning

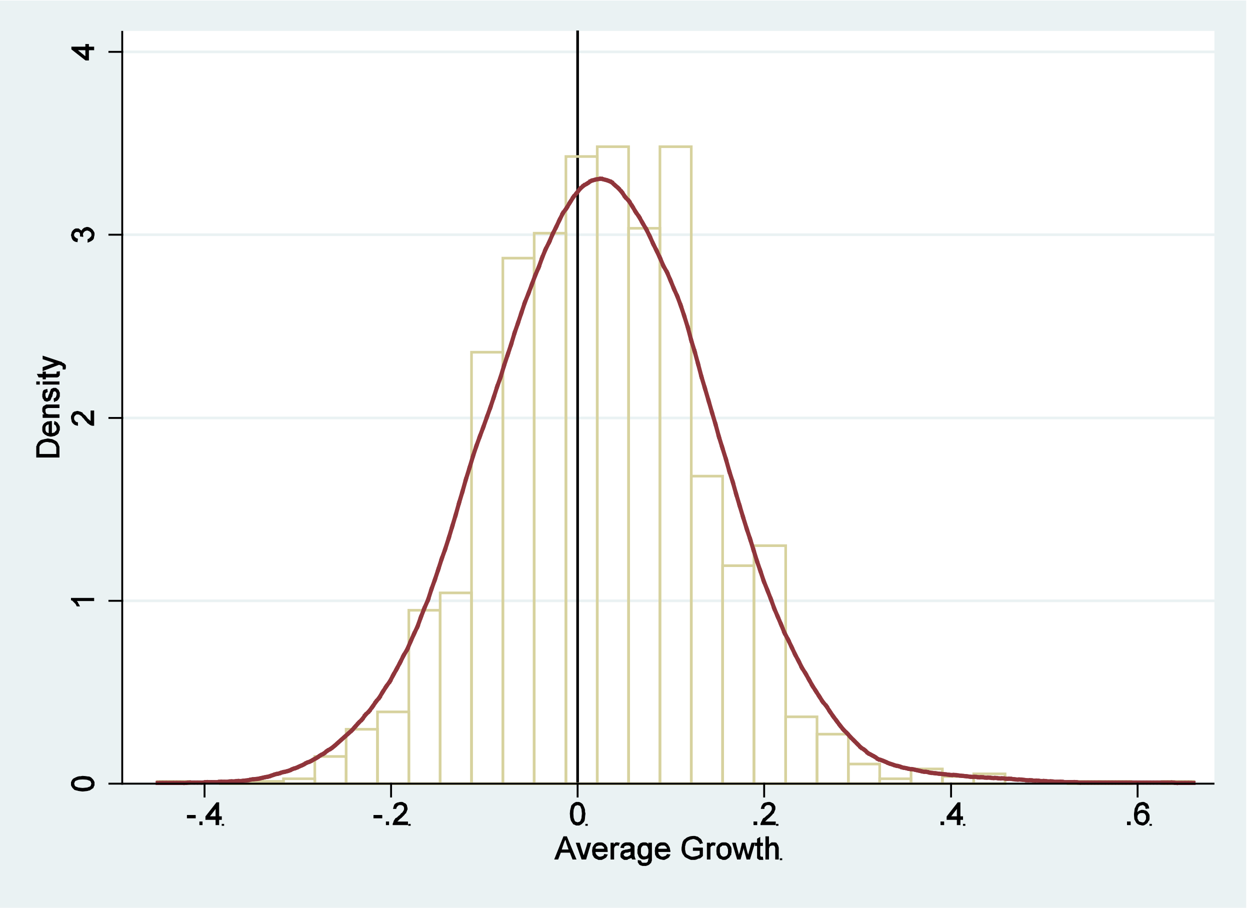

Figure 2: Density of observations across assignment variable.

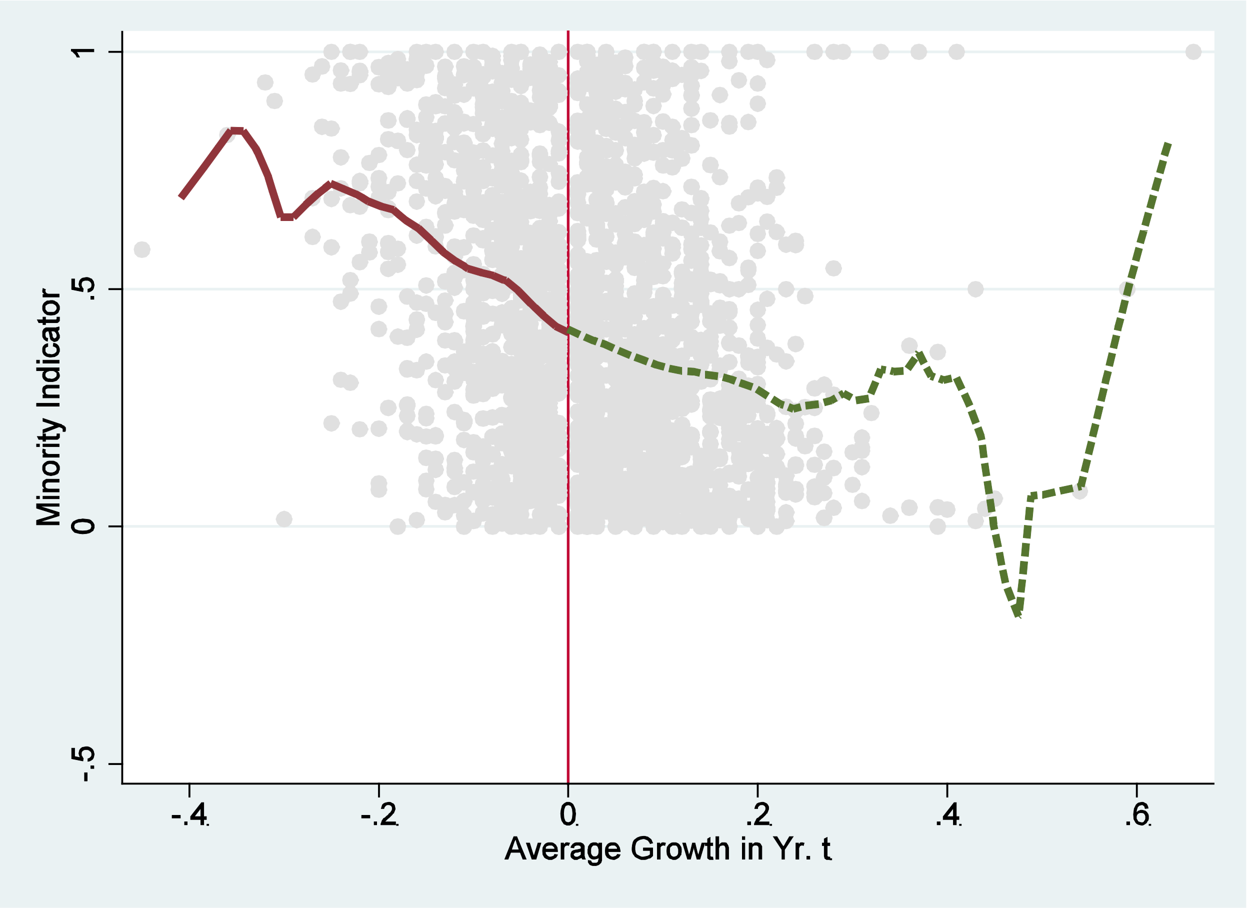

Figure 3: 'Placebo' RD of minority percentage

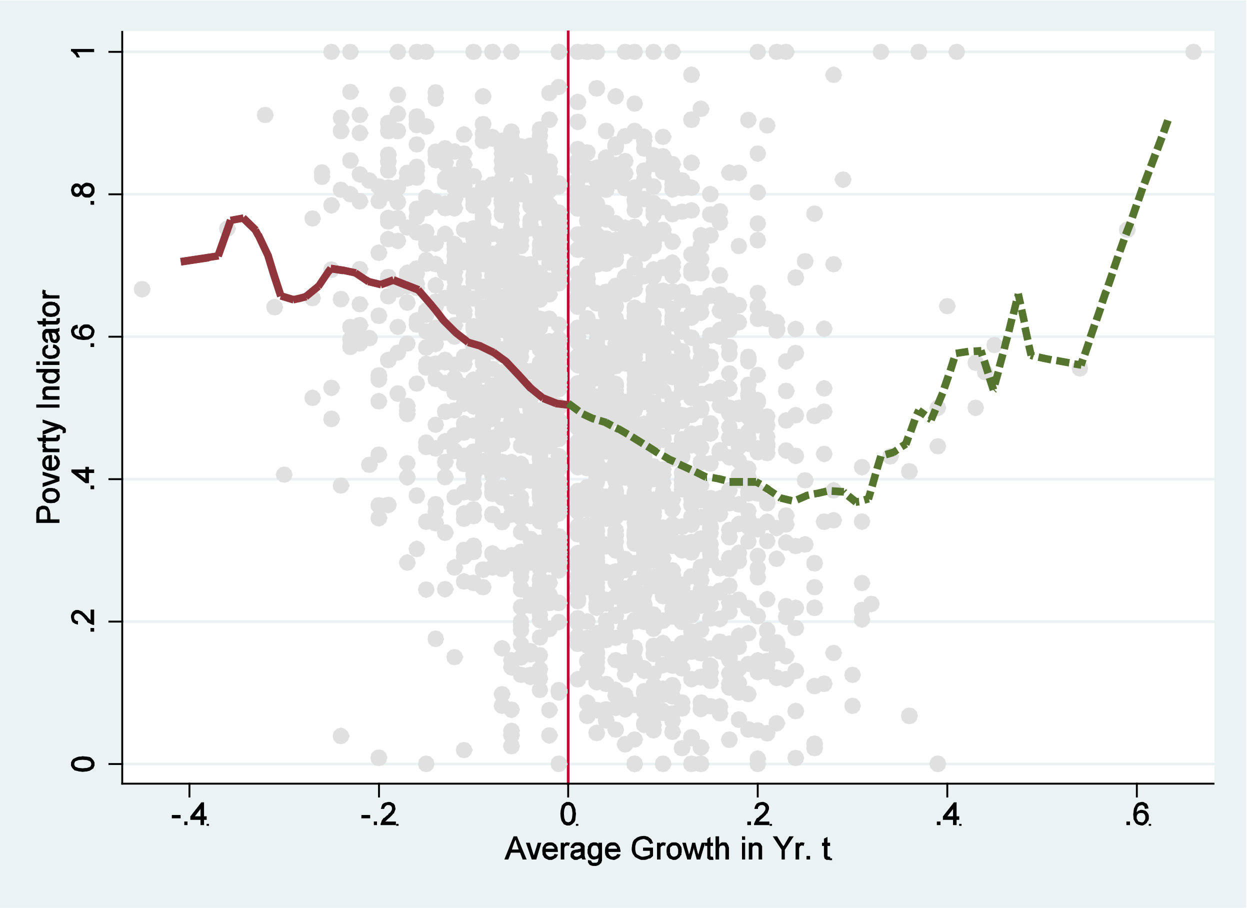

Figure 4: ‘Placebo’ RD of poverty (free/reduced price lunch) percentage

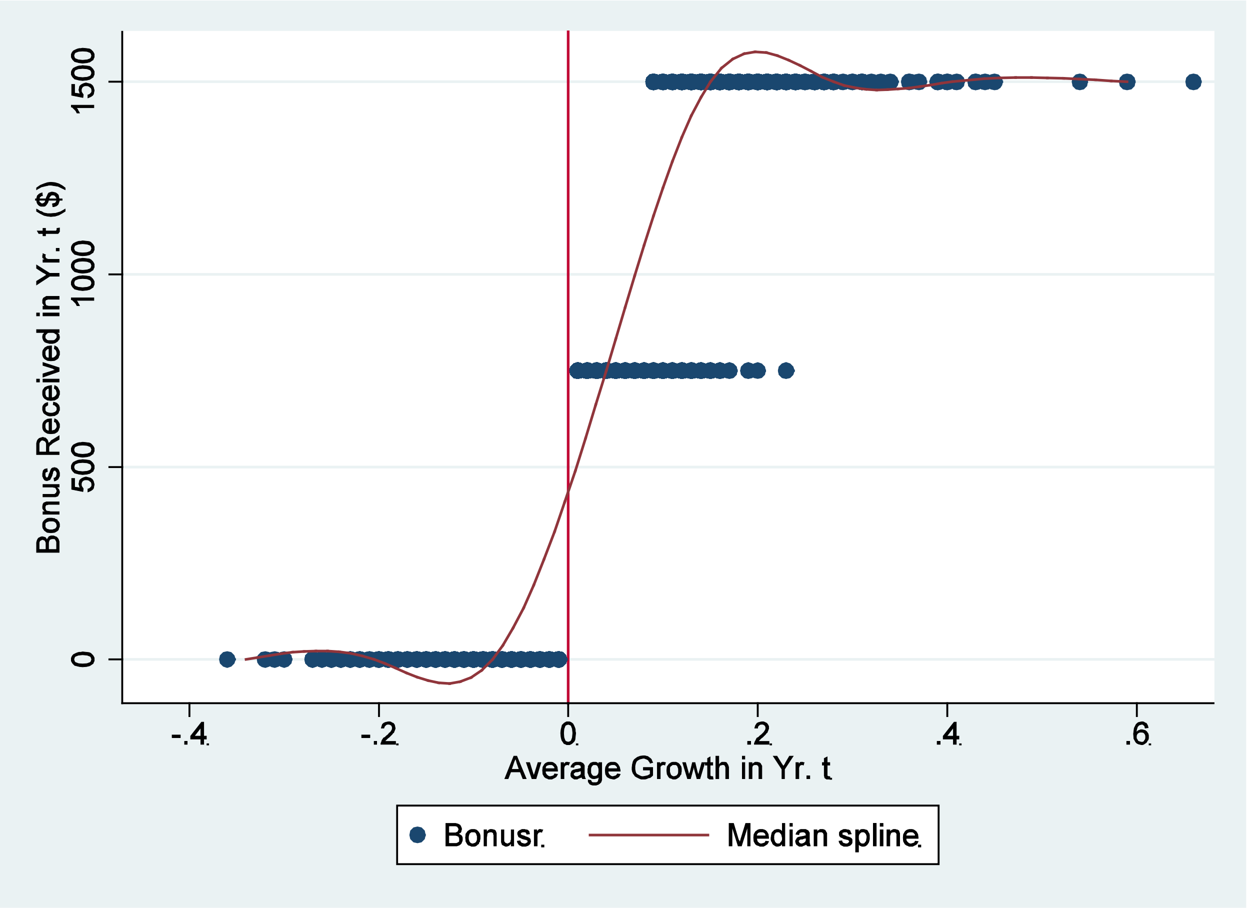

Figure 5: Existence of discontinuity in probability of bonus receipt at policy change

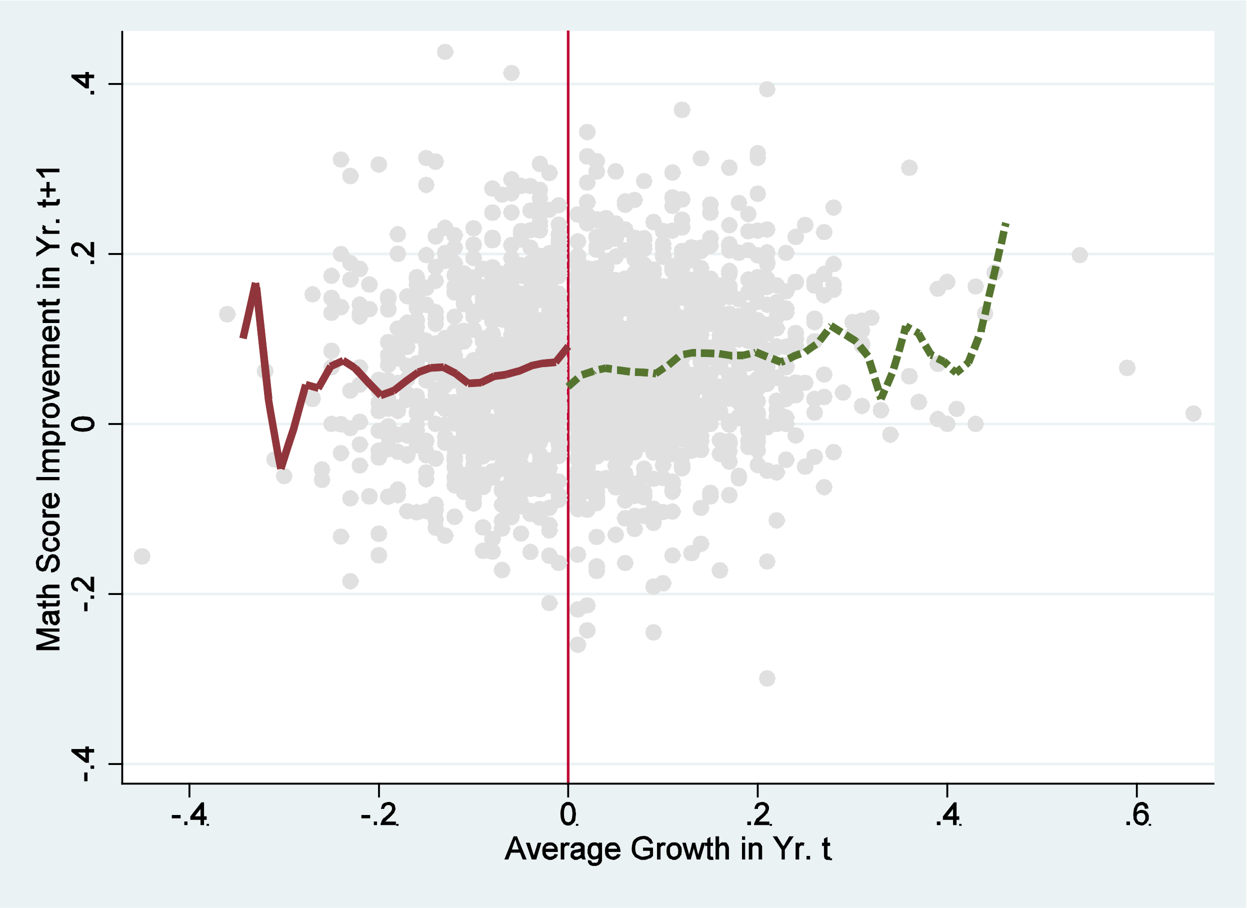

Figure 6: Simple RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t.

Tables

Table 1: AYP and ABC status

		ABC	
		Yes	No
AYP	Yes	956	284
	No	423	635

Table 2: Summary Statistics

Variable	Mean (Std. Dev.)
∆ math score	0.0617 (0.4555)
∆ reading score	-0.0348 0.6362)
math proficiency level	2.8408 (0.8439)
reading proficiency level	3.2976 0.7838)
∆ math proficiency level	0.0454 (0.6251)
∆ reading proficiency level	-0.0327 (0.7516)
% minority	0.3959 (0.4891)
% poverty	0.4582 (0.4982)
Years since last bonus	0.6524 (0.5001)
Number of no bonus years in last 5 years	1.2267 (1.2530)
Years since AYP made	0.5558 (0.9146)
Number of AYP failed since 2002-03	1.0547 (1.0776)
Observations	569,808

NCERDC data of elementary school and students from 2005-06 to 2006-07. Math and reading scores are c-scores.(See text for description) A student is proficient in a subject with a level 3 or 4. Students receiving free/reduced price lunch are poverty status. Minority students are blacks, Hispanics, and American Indians.

Table 3: Regression Discontinuity Results for Bonus Receipt: Entire Sample

Outcome Measure	RD Effect (Std. Err.)	Bandwidth
Δ math score	-0.0188 (0.0032)***	0.1195
	-0.0200 (0.0049)***	0.0597
	-0.0175 (0.0024)***	0.2390
Δ reading score	-0.0114 (0.0064)*	0.0829
	-0.0325 (0.0104)***	0.0415
	-0.0050 (0.0045)	0.1659

Table 4: Regression Discontinuity Results for Bonus Receipt: Math Score Only, By Demographic Subsamples

Subsample	RD Effect (Std. Err.)	Bandwidth
Minority	-0.0783 (0.0123)***	0.0612
	-0.0821 (0.0223)***	0.0306
	-0.0612 (0.0081)***	0.1224
Non-minority	-0.0019 (0.0073)	0.0977
	0.0038 (0.0120)	0.0489
	-0.0156 (0.0054)***	0.1955
Poverty	-0.0476 (0.0124)***	0.0571
	-0.0787 (0.0244)***	0.0285
	-0.0352 (0.0081)***	0.1151
Non-poverty	-0.0271 (0.0062)***	0.1419
	-0.0169 (0.0091)*	0.0710
	-0.0275 (0.0049)***	0.2839

Table 5: Regression Discontinuity Results for Bonus Receipt: Reading Score Only, By Demographic Subsamples

Subsample	RD Effect (Std. Err.)	Bandwidth
Minority	-0.0089 (0.0083)	0.1528
	-0.0258 (0.0121)**	0.0764
	-0.0027 (0.0066)	0.3057
Non-minority	-0.0085 (0.0085)	0.0902
	-0.0238 (0.0136)*	0.0451
	-0.0110 (0.0060)*	0.1804
Poverty	-0.0004 (0.0103)	0.0878
	-0.0289 (0.0165)*	0.0439
	0.0059 (0.0072)	0.1756
Non-poverty	-0.0297 (0.0088)***	0.0929
	-0.0415 (0.0146)***	0.0464
	-0.0244 (0.0063)***	0.1858

Table 6: Regression Discontinuity Results for Bonus Receipt: Math Score Only, By Proficiency Level

Level	RD Effect (Std. Err.)	Bandwidth
I	-0.0052(0.0228)	0.0979
	-0.0020 (0.0372)	0.0490
	-0.0037 (0.0168)	0.1958
II	-0.0338 (0.0205)*	0.0469
	-0.1001 (0.0342)***	0.0234
	-0.0389 (0.0127)***	0.0937
III	-0.0405 (0.0064)***	0.1523
	-0.0467 (0.0092)***	0.0762
	-0.0317 (0.0051)***	0.3047
IV	-0.0150 (0.0116)	0.0996
	-0.0025 (0.0192)	0.0498
	-0.0216 (0.0085)**	0.1992

Table 7: Regression Discontinuity Results for Bonus Receipt: Reading Score Only, By Proficiency Level

Level	RD Effect (Std. Err.)	Bandwidth
I	0.0110 (0.0420)	0.1017
	-0.0394 (0.0695)	0.0508
	0.0400 (0.307)	0.2033
II	0.0078 (0.0211)	0.0866
	-0.0273 (0.0342)	0.0433
	-0.0032 (0.0149)	0.1732
III	-0.0192 (0.0105)*	0.0834
	-0.0537 (0.0172)***	0.0417
	-0.0118 (0.0074)	0.1667
IV	-0.0252 (0.0085)***	0.1012
	-0.0263 (0.0137)*	0.0506
	-0.0170 (0.0062)***	0.2024

Table 8: Regression Discontinuity Results for Bonus Receipt: Math Score Only, By Accountability History

Accountability History	RD Effect (Std. Err.)	Bandwidth
No bonus more than 2 out of last 5 years	-0.0813(0.0107)***	0.0735
	-0.0495 (0.0168)***	0.0367
	-0.0673 (0.0075)***	0.1469
Bonus in 3 or more of the last 5 years	-0.0093 (0.0053)*	0.0593
	-0.0252 (0.0106)	0.0296
	-0.0092 (0.0035)***	0.1186
Failed to make AYP for the last 2 or more years	-0.1282 (0.0186)***	0.0374
	-0.1036 (0.0116)***	0.0187
	-0.1037 (0.0107)***	0.0748
Made AYP every year since 2003	-0.0033 (0.0056)	0.0540
	-0.0146 (0.0106)	0.0270
	-0.0074 (0.0036)**	0.1081
Failed to make AYP more than 2 years	-0.0834 (0.0141)***	0.0460
	-0.0813 (0.0229)***	0.0230
	-0.0555 (0.0089)***	0.0921
Made AYP two years or more	-0.0110 (0.0033)***	0.1334
	-0.0067 (0.0049)	0.0667
	-0.0108 (0.0025)***	0.2668

Table 9: Regression Discontinuity Results for Bonus Receipt: Reading Score Only, By Accountability History

Accountability History	RD Effect (Std. Err.)	Bandwidth
No bonus more than 2 out of last 5 years	-0.0339(0.0153)**	0.0962
	-0.0368 (0.0258)	0.0481
	-0.0291 (0.0113)**	0.1923
Bonus in 3 or more of the last 5 years	-0.0071 (0.0066)	0.0895
	-0.0268 (0.0104)**	0.0448
	-0.0002 (0.0047)	0.1791
Failed to make AYP for the last 2 or more years	-0.0802 (0.0214)***	0.0582
	-0.1021 (0.0419)**	0.0291
	-0.0384 (0.0136)***	0.1163
Made AYP every year since 2003	-0.0067 (0.0071)	0.0793
	-0.0288 (0.0118)**	0.0396
	-0.0026 (0.0049)	0.1586
Failed to make AYP more than 2 years	-0.0341 (0.0131)***	0.0587
	-0.0727 (0.0257)***	0.0293
	-0.0103 (0.0085)	0.1173
Made AYP two years or more	-0.0108 (0.0078)	0.0888
	-0.0265 (0.0124)**	0.0444
	-0.0060 (0.0054)	0.1775