NBER WORKING PAPER SERIES

WHEN INCENTIVES MATTER TOO MUCH: EXPLAINING SIGNIFICANT RESPONSES TO IRRELEVANT INFORMATION

Thomas Ahn
Jacob L. Vigdor

Working Paper 20321
http://www.nber.org/papers/w20321

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2014

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). The authors are grateful to John Holbein for research assistance and to 2013 SEA conference, 2014 AEA conference, 2014 AEFP conference, and 2014 IWAEE conference attendees and seminar participants at the University of Colorado, University of Virginia, and University of Kentucky for helpful comments on earlier drafts. Any opinions expressed herein are those of the authors and not any affiliated organization. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

When Incentives Matter Too Much: Explaining Significant Responses to Irrelevant Information
Thomas Ahn and Jacob L. Vigdor
NBER Working Paper No. 20321
July 2014
JEL No. D03,I2,J33

ABSTRACT

When economic agents make decisions on the basis of an information set containing both a continuous variable and a discrete signal based on that variable, theory suggests that the signal should have no bearing on behavior conditional on the variable itself. Numerous empirical studies, many based on the regression discontinuity design, have contradicted this basic prediction. We propose two models of behavior capable of rationalizing this observed behavior, one based on information acquisition costs and a second on learning and imperfect information. Using data on school responses to discrete signals embedded in North Carolina's school accountability system, we find patterns of results inconsistent with the first model but consistent with the second. These results imply that rational responses to policy interventions may take time to emerge; consequently evaluations based on short-term data may understate treatment effects.

Thomas Ahn
BE 335X
Lexington, K 40506
thomas.ahn@uky.edu

Jacob L. Vigdor
Sanford School of Public Policy
Box 90245
Duke University
Durham, NC 27708
and NBER
jacob.vigdor@duke.edu

1. Introduction

Traditional economic theory suggests that agents should exhibit no reaction to the introduction of irrelevant information. In scenarios where agents have access to a quantitative information set, supplementing the data with discrete signals derived directly from the data should have no impact. For example, when a publicly available, continuous rating of product quality exists, the introduction of discrete rating categories based solely on the continuous measure should not alter consumer decisions.

A growing number of empirical analyses identify scenarios where this basic principle fails in practice. The phenomenon is particularly salient among papers employing the Regression Discontinuity (RD) analysis, applied in scenarios where subjects are assigned to discrete categories on the basis of an underlying continuous measure.[1] The mounting empirical evidence motivates a theoretical question: is there a simple model of behavior that can explain significant responses to signals that should be irrelevant?

There are several important theoretical models of seemingly irrational behavior (Mullainathan 2002; Sarver 2008; Gabaix 2014; see Gigerenzer and Goldstein 1996 for a review of models in the psychology literature). Informed by this prior literature, this paper presents two candidate explanations for significant responses to irrelevant information in a principal-agent style production framework with output incentives, and uses data on school responses to an educational accountability program in North Carolina to evaluate their empirical implications. The first model focuses on information costs, presuming that agents are "sparse maximizers" in

[1] As originally conceived, the RD design is intended to reveal the local average treatment effect of an intervention provided to individuals in one category but not the other. In some cases, there is no intervention *per se*, but rather subjects are exposed to differing values of a signal that is a simple function of some readily available continuous variable. This is the type of empirical exercise considered here.

the sense of Gabaix (2014) using the discrete signal as a convenient proxy for the underlying continuous variable. The second posits that agents are Bayesian updaters who exhibit rational behavior when fully informed. When imperfectly informed, however, their incomplete understanding of the production process may lead them to react significantly on the basis of the discrete signal (Sims 2003). These two models yield divergent empirical predictions.

In the empirical case examined here, personnel at public schools in North Carolina were provided with annual reports of school effectiveness, with effectiveness measured as a continuous variable. When this continuous variable exceeded certain predefined thresholds, the personnel were awarded salary bonuses. From a rational perspective, presuming that the effectiveness measure is stationary or subject to a smooth drift process, receipt of the bonus yields no additional information regarding the likelihood of being awarded a bonus in the future.

We demonstrate that schools falling below the threshold in fact exhibited significant improvements in performance in following years relative to schools just above the threshold.[2] The discontinuity is greatest among schools with inexperienced principals, and among schools with a track record of inconsistent performance. These patterns are more consistent with the learning model than with the information cost model.

In addition to offering insight into the nature of seemingly irrational responses to irrelevant information, this paper makes a contribution to the large and growing literature on educational accountability. Accountability systems are designed to improve student performance by introducing incentives or market pressures that may be absent in public education systems. While the literature has so far found both encouraging and cautionary evidence regarding the effects of accountability systems, the implicit or explicit model underlying the empirical analysis

[2] One other study has confirmed these basic results, that the NC bonus system creates a discontinuity in test score growth. See Jinnai 2013.

has usually been static, one-shot game with all agents having full information (Hanushek and Raymond 2005, West and Peterson 2006, Chiang 2009, Chakrabarti (forthcoming), Springer et al. 2012, Yuan et al. 2013 among many others). This study adds to the evidence that accountability systems can be useful in inducing performance gains while emphasizing that agents in a noisy production environment will exhibit dynamic learning and respond rationally using the limited (and sometimes erroneous) information at hand.

This evidence carries important implications for the design of incentive systems in education and other domains. The theory underlying the principal-agent model presumes that agents fully understand the incentives applied by the principal and the process by which their own efforts translate into outcomes. Our results suggest that agents may require several iterations to attain this mastery. Among other things, this implies that short-term evaluations of incentive programs may vastly understate the potential long-run implications of incentive regimes. Our results also suggest that efforts to "tweak" incentive programs may interrupt the learning process, implying that the possible benefits of better aligning incentives need to be weighed against the costs of interrupting agents' learning process.

2. Conceptual Framework

The set of scenarios where economic agents base decisions on an information set that includes both continuous measures and discrete codings of those measures is large. Diners make use of restaurant sanitation and online (crowd-sourced) review scores that may be summarized as a letter grade (Jin and Leslie 2003, Anderson and Magruder 2011). Home owners make discretely different electricity consumption decisions based on "grades" assigned by the utility company (Allcott 2011). Consumers in the used car and diamond markets pay discrete higher

prices for marginally different goods, based on the left-most numerical digit of the product descriptor (odometer reading for used cars, and carat size for diamonds) (Lacetera, Pope, and Sydnor 2012, Scott and Yelowitz 2010). Even in cases where there is no explicit information being transmitted due to the treatment, researchers have found discrete behavioral differences. Basketball teams react differently to being down or up a point at half time (Berger and Pope 2011), and racial segregation (white flight) may be instigated by the racial mix of a neighborhood reaching a "tipping point" (Card, Mas, and Rothstein 2008). School performance is commonly measured using continuous variables such as proficiency rates, but summarized with letter grades or binary designations. Evidence indicates that home buyers respond to the summary categorization even when the underlying continuous measure is publicly available (Figlio and Lucas 2004, Martinez 2010). Contributions to parent-teacher organizations display a similar sensitivity to discrete grade information (Figlio and Kenny, 2009). Student human capital investment decisions are also affected by discrete grade designations (Ahn 2014, Papay, Murnane, and Willett 2011).

In all these scenarios, a traditional model of rational behavior would predict no sensitivity to the discrete indicator conditional on knowledge of the underlying continuous variable. In many cases, the failure of some agents to react rationally to irrelevant information creates profit opportunities for other agents. Previous literature has identified numerous cases of overreaction to irrelevant information, but explanations for this behavior and empirical tests of these explanations are still uncommon.

2.1 Traditional model

As noted above, there have been numerous empirical exercises refuting traditional rational models. Nonetheless, to frame our discussion of alternative models we begin by

outlining a basic rational framework, in this case based on a principal-agent model.

Suppose output y_{it} of employee *i* in period *t*, which in the context of educational production might be measured by improvements in student test scores, is a known function of a vector of inputs chosen by the employee, x_{it}, which might generalize to effort in a simple model but might be more realistically thought of as an array of possible uses of time, and an idiosyncratic shock ϵ_{it}. The employee's utility is a function of their wage, w_{it}, and a cost function based on input choices, $c_i(x_{it})$, which we take to be increasing and convex. We also allow for the possibility that $c_i(x_{it})$ may be less than zero for certain values of x_{it}, which would be the case if employees received some satisfaction or pride 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 employee observes y_{it}, and can thus determine the value of ϵ_{it} ex post.

To incentivize effort, the employer links compensation 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.[3]

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.

[3] 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.

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 variation in ϵ_{it}. Under a variety of assumptions regarding the evolution of ϵ_{it}, we should expect no change in the optimal effort choice as a function of incentive receipt. Were ϵ_{it} entirely uncorrelated across years, rational employees would clearly behave exactly the same in the subsequent period. Even under non-random evolution of ϵ_{it}, so long as $E(\epsilon_{it+1} \mid \epsilon_{it})$ does not exhibit a discontinuity precisely at the threshold distinguishing incentive recipients from non-recipients there is no reason to expect discontinuous behavior changes conditional on the value of ϵ_{it}. The most plausible scenario involving a discontinuous behavior change at the point of discontinuity would occur if there were no year-to-year variation whatsoever in ϵ_{it}. In that scenario, the idiosyncratic determinant of output would be better described as an element of an employee's ability that is uncertain until the first period of employment.[4] In the context of educational production, it is unrealistic to think of an outcome such as a class-level average change in standardized test scores as perfectly predictable conditional on information regarding agent effort choices. As such, we dismiss this scenario.

2.2 Introducing information costs and selective re-optimization

To rationalize the existence of discontinuous responses to signals based on available information, we must introduce some additional complication to the basic model. The most obvious extensions would involve departures from complete information. A simple extension

[4] This basic scenario can be straightforwardly translated to simple models of consumer choice. When selecting a product, such as a used car or diamond, the quality of the product may vary monotonically as a function of a continuous measure such as mileage or weight, but in most cases there is no reason for quality inferences to vary discontinuously at any arbitrary point in the distribution. The exception would be in a "lemons"-type scenario, where potential suppliers of a good selectively choose to sell because of a known proclivity among buyers. For example, potential sellers of cars with an odometer reading of 10,000 miles might rationally withhold their cars from the market given consumer discounting.

would introduce costs to the agent of observing the realization of ϵ_{it}. The notion of information costs underlies other theoretical models of boundedly rational behavior, e.g. Gabaix (2014).

To generate a prediction of asymmetric response at the incentive threshold, the model would need to yield a decision rule that involved paying the cost to realize the information only conditional on receiving a positive (or negative) signal. Plausibly, if there is significant serial correlation in the value of ϵ_{it} agents might find it optimal to pay the information cost only in the event of an unexpected shock, such as the failure to receive a discrete bonus payment after a steady period of receiving it, as the investment in the information cost would yield an expected return in the form of resolving uncertainty regarding the potential return to altering effort.[5] If the agent were to learn that their performance lay significantly close to the threshold they might rationally choose to increase it.

In this scenario, one would predict a significant increase in effort among agents that barely missed the performance threshold, relative to those who barely made it, because only the former group would engage in the reoptimization made possible by incurring the information cost. One would further expect that the discontinuous response would be particularly noteworthy among agents that had a track record of being above the threshold consistently in earlier periods.

There is a companion scenario where agents with a track record of falling below the threshold find themselves above it, and incur the information cost to determine whether they can safely reduce their effort in subsequent periods. In such a scenario, one would expect the agents who barely made the threshold to exert more effort than their counterparts on the other side.

2.3 Introducing incomplete information regarding the production process

[5] The intuition here is similar to that of the finite adjustment cost or [S,s] model (Bertola and Caballero 1990).

A second variation on the imperfect information theme regards uncertainty in the nature of the production process itself. Here we consider a model of uncertainty and Bayesian updating closely related to a category of problems in the bandit framework (Rothchild 1974, Berry and Friestedt 1985, Easley and Kiefer 1988, Kiefer 1989, El-Gamal and Sundaram 1993, and Bala and Goyal 1995, among many others). Generally, the problem involves a long-lived agent with a set of prior beliefs regarding the state of the world whose actions results in some observed outcome that is used to derive a new posterior set of beliefs, which in turn becomes the new prior guiding the next round of action. We consider agents that begin their careers, or their experiences under the incentive regime, uncertain exactly how their effort allocations translate into output and uncertain of the role of exogenous factors relative to effort in production. This arguably characterizes the status of new entrants to a profession such as teaching.

After the first production period, agents receive information on a continuous measure of their output and a binary indicator of whether they met a production threshold. This single data point is insufficient for agents to disentangle the effects of luck vis-à-vis effort allocations. It is clear, however, that agents receiving different signals face differing incentives to alter their effort allocation in the next period, so long as their outcome in the first period can be relied upon as the expected value of their output with the same effort choices. Particularly for agents whose prior beliefs emphasize effort rather than luck, the impetus to alter effort allocations is strongest among teachers who receive the initial signal that their output was insufficient.

In subsequent periods, agents who maintain the same effort allocation can rapidly update their prior beliefs regarding the role of exogenous factors in production, while only minimally updating prior beliefs regarding how output varies with their efforts. Agents who alter their effort allocation – because they have been instructed to improve – learn about the links between

effort and output more rapidly, but about the role of exogenous factors more slowly. At intermediate experience levels, then, the model predicts significantly different patterns of discontinuous response to the discrete signal. Agents with a strong track record correctly impute that their likelihood of retaining the incentive payment in the next period is effectively identical on either side of the discrete threshold – and in any event, know little about how to improve their performance were it to be indicated. Agents with a weaker track record are more empowered to take actions to improve performance when they are asked to do so, and may be more likely to underestimate the role of chance in determining whether they stay above or below the threshold in the following period.

In the long run, agents converge to full information regarding the nature of the production process, at which point the model predicts that no significant discontinuities will persist.

Describing the model more formally, the agent attempts to maximize the probability of bonus receipt:

$$\max_{x_{it}} Pr(y_{it} \geq y^*)$$

$$s.t. y_{it} = f(x_{it}, \theta) + \epsilon_{it}$$

Output in year t (y_{it}) which must be greater than the predetermined cut-off (y^*), is determined by some input(s) x_{it}, parameters governing the translation of inputs to output, represented by θ, and some idiosyncratic noise ϵ_{it}, with variance σ_ϵ^2. The noise term may also exhibit autocorrelation.

We assume, but do not explicitly specify, that agent choices of x_{it} are limited by a set of budget and time constraints. In some cases, elements of x_{it} may be qualitative rather than quantitative in nature, for example an instructor's choice of textbook or the decision to instruct

students as an entire class or in smaller groups. Certain of these choice alternatives may be mutually exclusive. The agent initially has some set of prior beliefs about the parameters governing the translation of input choices into the deterministic component of output and the variance of the noise: $p(\theta, \sigma_\epsilon^2)$.[6]

As an illustrative case, suppose agents have prior beliefs about the distribution of variance of noise, with some positive mass of agents who believe $\sigma_\epsilon^2 = 0$ – that is, they believe that output is a deterministic function of their input choices.[7] These agents are referred to below as "naïve" agents. If we further assume that these agents have perfectly diffuse priors on θ are perfectly diffuse, it follows that their vector of initial inputs x_{i0}, are randomly distributed.[8]

In period 0, some fraction α of agents (which contains both naïve and non-naïve agents) observe outcome $y_{i0} > y^*$, and $(1-\alpha)$ fraction of agents observe outcome $y_{i0} < y^*$. Naïve agents who observe $y_{i0} > y^*$ have no incentive to change their inputs in the next period, because given their diffuse priors, they can identify no change in inputs with a positive expected effect on output, regardless of their prior beliefs on σ_ϵ^2. Therefore, $x_{i0} = x_{i1} | y_{i0} > y^*$ and $\mathrm{E}(y_{i1} | y_{i0} > y^*) = y_{i0}$.

Naïve agents who observe $y_{i0} < y^*$ select a new, *random* vector of x_{i1}. They will change inputs because they believe $\sigma_\epsilon^2 = 0$, implying that their probability of receiving the bonus in period 1 will be zero if they do not alter input choices; they can do no worse by altering inputs and have some chance of doing better. Because they lack information on which inputs improve

[6] The prior on the two parameters may be joint, but we assume they are uncorrelated for simplicity here.
[7] This is essentially equivalent to all agents having the same prior about the distribution of the variance with a positive probability mass at $\sigma_\epsilon^2 = 0$.
[8] For example, assume initially that $\theta \sim \mathrm{N}(0, \infty)$. Qualitative results are unchanged if we relax our assumptions on noise and tuning parameters, as long as some positive mass of agents are less informed about the tuning parameter and mistakenly believe that the noise component is smaller than the true value; that is, some naïve principals are overly optimistic about their degree of control over the education production process.

rather than worsen output, roughly half of the naïve agents will see *y* increase in the next period, and the other half will observe it decrease.

At early stages, then, there is little reason to expect a discontinuity in period *t* outcomes among naïve agents on differing sides of the incentive threshold in period *t*-1. Average outcomes for naïve agents above the threshold do not change because inputs do not change ($\mathrm{E}(y_{i1}|y_{i0} > y^*) = y_0$). Among naïve agents below the threshold, $\mathrm{E}(y_{i1}|y_{i0} < y^*) = y_0$ because their efforts to change inputs are as likely to do harm as good.[9] Among more sophisticated agents, in the limit rational and perfectly informed ones, we also expect no difference in outcomes on either side of the threshold.

In period 1, the α fraction of naïve agents from period 0 who did not change their inputs observe $y_{i1} \neq y_{i0}$. Even if some of these agents believed initially that $\sigma_\epsilon^2 = 0$, they now learn with probability approaching one that $\sigma_\epsilon^2 > 0$ and update their posterior on the variance of exogenous shocks accordingly. Agents with outcome y_{i1} sufficiently below y^* will choose to draw a new x_{i2} because their ϵ_{i1} draw has changed their expected value of future outcome to be below the threshold. Importantly, agents immediately to the left and right of the threshold will not alter their input choices because of their revised posterior on σ_ϵ^2. As such, these agents will not display discontinuous behavior at y^*.

The naïve agents among the $(1 - \alpha)$ fraction, who have the benefit of observing outcomes under two separate draws of x_{it}, will observe y_{i1} and update their beliefs on θ. It is critically important to note here that if agents initially believed that $\sigma_\epsilon^2 = 0$, they fail to receive any information that would contradict their priors. In essence, they now possess better information on how to translate inputs into output, but they are learning with error (of which they

[9] Note the implication that $\mathrm{V}(y_{i1} - y_{i0}|y_{i0} < y^*) > V(y_{i1} - y_{i0}|y_{i0} > y^*)$

are unaware). As described above, half of these agents have $y_{i1} > y_{i0}$ and the other half have $y_{i1} < y_{i0}$.

Naïve agents with outcome $y_{i1} > y^*$ will choose to not change their input allocation in the next period because their beliefs about σ_ϵ^2, coupled with continued uncertainty regarding θ, dictate that inaction maximizes their expected utility. On the other hand, naïve agents with outcome $y_{i1} < y^*$ will choose a third set of inputs x_{i2} informed by their new posterior on θ. Because these agents are now choosing x_{i2} under partial information, the net average effect is *positive*. That is , $\mathrm{E}(y_{i2}|y_{i1} < y^*) > y_{i1}$. In period 2, then, a discontinuous response is expected among naïve agents who elect to reoptimize only upon receiving the negative signal, and can do so with at least some degree of efficacy.

As time progresses, agent behavior continues following the basic heuristics described above. Initially naïve agents continue to select new combinations of inputs as long as their output remains below the threshold, continuously improving their knowledge of how inputs can be chosen to improve output but remaining stunted in their understanding of exogenous determinants of output. Agents who reach the threshold stabilize their input choices, which permit them to acquire better information on the role of chance.

In the long run, agents become fully informed, at which point there is no longer any reason to expect discontinuous behavior changes around the incentive threshold. It is at intermediate stages, where naïve agents have learned enough to be efficacious in their reoptimization choices but not enough to realize that output is not entirely deterministic, where discontinuous responses are anticipated.

3. Empirical Application: The North Carolina ABC Program

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, the *composite growth index*, 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.

Figure 1, reprinted from Vigdor (2009), shows the proportion of schools eligible for bonus payments from the inception of the program through 2006/07. Between half and 90% of schools were eligible for at least some bonus payment in every year, while the proportion eligible for the full $1,500 bonus varied between 10% and 70%.

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 years 2005/06 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.

It is important to emphasize that there is no direct connection between a school's performance in year *t*-1 and the stakes for making or missing the bonus threshold in year *t*. The substantial fluctuation in the proportion of schools receiving bonus payments from year to year underscores the importance of noise in the evolution of school performance measures over time. Schools on either side of the bonus threshold in year *t*-1 should have derived little or no information regarding their prospects for receiving a bonus in year *t*, particularly after conditioning on the composite growth index.

4. Data and Methods

4.1 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.[10] 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, which when aggregated to the school level yield our measure of output. 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

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

in schools serving grades 3-5 in the 2005/06 and 2006/07 school years.[11] 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 dataset contains roughly 340,000 student/year observations in 2,248 elementary schools/years.

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. Besides a continuous score, elementary school students in North Carolina are placed into one of four proficiency levels for reading and mathematics, with level three indicating "sufficient mastery" of the subject, which equates to grade-level proficiency. On average, students in North Carolina are slightly below grade-level proficiency for math and slightly above for reading.

To proxy for the number of opportunities available to the school leadership to update its priors on noise in the education production process and/or the efficacy of inputs, we use years of experience of the school principal. Student-level data is linked to the payroll data for teachers and administrators to track the experience level of the principals. Payroll data is available from 1992, which allows us to count up to 13 years of experience for a principal at a school in the 2005-06 academic year. Therefore, years of experience for principals suffers from slight right censoring. However, principals with 13 or more years of experience comprise less than 10 percent of the data.

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

[11] 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.

index, along with a few other school-level summary statistics. The composite growth index ranges from -0.45 to 0.66; schools with values above zero qualifying for the $750 bonus.[12]

4.2 Methodology

Our basic goal is to examine the impact of bonus receipt on student performance in the following academic year, using regression discontinuity (RD) analysis. Regression discontinuity analysis can be performed either parametrically or nonparametrically. 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 use the Hahn, Todd, and van der Klaauw (2001) nonparametric specification in this study by estimating a local linear regression to fit a smooth function to either side of the discontinuity.[13] While it is not necessary to specify a functional form using this method, a bandwidth – effectively, the number of data points incorporated into the local linear regression at any point – 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 report results for a variety of bandwidths centered around the "optimal" bandwidth as defined by Imbens and Kalyanaraman (2009).

Our estimation is based on student-level records, yet assignment to the treatment is at the school level. To estimate the impact of the treatment correctly, we collapse the individual-level

[12] 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.
[13] We also estimate parametric RD to lend support to our findings.

data to the school level averages. We then weight observations by the number of student observations used in the school-specific means of the dependent variable and covariates.[14]

4.3 Standard tests of RD Validity

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 (Imbens and Lemieux 2008; Lee and Lemieux 2010).

First, we check for evidence of manipulation: that schools employ strategies to place their composite growth index score just above the critical value. Such manipulation would be expected to create a discontinuity in the density of school-year observations at the bonus threshold (McCrary 2008). Figure 2 shows the distribution of the average growth performance measure across all school-year observations in school years 2005/06 and 2006/07, with a normal density overlaid. The bonus threshold occurs at the composite growth index score of zero.

The density is relatively smooth, and importantly does not exhibit any sharp changes near the cut-off value. Coupled with an insignificant McCrary test statistic (McCrary 2008), the smooth density gives us confidence that manipulation of test scores is not a significant area of concern.[15] Intuitively, as the composite growth index is a function of the performance of dozens to hundreds of students, calculated using a complex formula, and revealed *ex post*, it would seem impossible to engage in *ex ante* manipulation.

[14] We also estimate models using student-level data, using a bootstrapping procedure to approximate clustered standard errors at the school level. Point estimates are comparable, but the clustered standard errors are large. This is consistent with the notion that clustering is a conservative solution to the problem of grouped data.

[15] Because our analysis divides the full sample into groups of schools with different accountability outcome histories and differently experienced principals, the density of the assignment variable is plotted for each subsample. Figures are available on the online appendix at http://sites.google.com/site/tomsyahn/.

We next check for balance in covariates on both sides of the bonus threshold. Figures 3-6 plot the school-level percent female students, percent minority students, percent limited English proficient students, and percent of students receiving free or reduced price lunch against the composite growth index. None of these covariates exhibit a discontinuity near the zero threshold.

Next, we 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 7 shows teachers' bonus receipt as a function of the composite growth index. It is clear that there is a sharp discontinuity in probability of bonus receipt (from zero to one) at zero. 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.[16]

These results support the validity of the North Carolina ABCs program for causal estimation of a local average treatment effect. As noted above, a traditional rational behavioral model would predict a LATE of zero as the bonus payments are applied *ex post* and schools on either side of the threshold should have nearly identical expectations regarding the potential returns to effort in the following year.

A rigorous RD investigation incorporates not only checks of validity but also checks regarding the robustness of results and falsification tests using alternate thresholds. We review these additional tests after introducing the basic results.

5. Results

[16] The "fuzziness" reflects the fact that eligibility for the $1,500 bonus was determined by additional factors beyond the composite growth index. In certain years, for example, test score gains needed to be distributed sufficiently broadly across students in order for schools to receive the $1,500 bonus.

5.1 Documenting the basic effect

Figure 8 presents a graphical representation of our most basic RD estimates, and Table 3 reports the associated effects and standard errors. In the case of math scores, our estimates indicate – as promised – that schools just below the bonus eligibility threshold exhibit higher test score gains relative to barely-eligible schools. The estimated effect is fairly robust to bandwidth choice, ranging from 0.0260 to 0.0458 with higher point estimates in models with narrower bandwidths. Figure 9 represents a model using a wider bandwidth and Figure 10 a narrower one.

These are substantial local average treatments effects, about one and a half times larger in magnitude compared to discontinuities estimated with the same dataset for the impact of failing to make adequate yearly progress under No Child Left Behind (Ahn and Vigdor 2013). In addition, Figure 8 shows that this improvement is quite meaningful for schools in close proximity to the bonus threshold. The association of larger effects with narrower bandwidth – and hence more flexible functional form – 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. Schools closest to the border may well perceive the greatest expected gains from increases in effort.[17]

For reading scores, the pattern of discontinuity estimates across bandwidths is similar to math, with point estimates larger at narrower bandwidths. However, most of the estimates are

[17] One possible reason for the observed test score gains may be teachers moving en-masse in response to the bonus outcome. If *less* effective teachers in failing schools are more likely to transfer, this may generate the observed discontinuity. However, the relatively modest bonus amount ($750 or $1,500) is most likely not enough to induce a large scale exodus. Similarly, while poorly performing students transferring in response to the bonus failure might produce the discontinuity, student transfer rates in NC have been traditionally very low. See Ahn and Vigdor 2013.

statistically insignificant.[18] Point estimates from these specifications are reported in Table 3. From this point forward, our discussion will focus on math results as reading results are consistently statistically insignificant in all specifications.

5.2 Traditional RD robustness checks

Figures 8-10 show some evidence of robustness of the math treatment effect estimate to bandwidth choice.[19] Figure 12 goes further, representing effect estimates and 95% confidence intervals at 20 different bandwidths. Nineteen of the twenty estimates lie within a narrow band indicating an effect size between 0.04 and 0.06. Fifteen of the twenty estimates have confidence intervals that exclude zero. The contradictory effect estimates correspond to extremely narrow bandwidth values, where the effect is estimated using a smaller set of data points leading to progressively less precise estimates. The estimate with narrowest bandwidth is the only one outside the 0.04-0.06 range; this estimate is so imprecise that we fail to reject effect sizes three times the value of other estimates in either direction.

Figure 13 shows the results of standard falsification tests using thresholds of no policy relevance. Regression discontinuity estimates using placebo composite growth index values from -0.2 to 0.3 yield imprecise estimates that are never significantly different from zero.

In summary, the strong estimated reaction to bonus receipt appears robust. It does not, however, make sense in the context of a traditional rational model. From this point forward, we assess alternate explanations for the effect.

---

[18] This is in line with most of the literature that finds teachers and schools less able to impact reading scores compared to math scores, most likely due to the fact that reading achievement may require more home and non-school inputs, compared to math achievement. See Jacob 2005, Reback 2008, and Rouse et al. 2013, among many others. We note that a small number of studies have found the opposite, with reading scores more sensitive to incentives. See Muralidharan and Sundararaman 2011.

[19] We also graph the RD using different degree polynomials (degree 0 to degree 3) to show that the discontinuity is not sensitive to the order of the polynomial (See Figure 11).

5.2 Testing the first alternate model: bonus as signal

The model outlined above suggests that schools act to assess and potentially re-optimize their behavior only in the presence of a signal that such activity may 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 believe that re-optimization is necessary to push them into the eligible category. As discussed above, this line of reasoning suggests that the signal value of bonus receipt (or non-receipt) is stronger when it comes as a surprise. For this reason, we now turn to a study of how reactions to bonus receipt vary across schools with differing histories, and therefore differing expectations, regarding the bonus.

Table 4 shows RD estimates for subsets of schools divided according to their past performance in the North Carolina bonus system. Schools are divided into those that have continuously qualified for the bonus, and those that have had at least one failure in the last five years. If indeed the failure to qualify for the bonus serves as an easy to interpret signal, we would expect to see strong reaction from high performing schools upon their first failure. Additionally, we may expect lower performing schools exerting maximal effort to stay above the bar once they qualify.[20] That is to say, we might expect an opposite-signed effect among schools with a history of infrequent bonus attainment.

Results from Table 4 rule out both hypotheses. Schools that have never failed do not react to their first failure. The discontinuity is not observed across any bandwidth when schools have qualified for the bonus in all years prior. Schools that have had previous failures in the recent past register substantial extra gains after a near-miss, relative to a near-make. For math scores,

[20] One may argue that these schools may incrementally increase academic growth, instituting more costly reforms as required. However, this may be assuming too much sophistication from these schools. One would assume that schools this savvy would not repeatedly fail to qualify for the bonus.

academic performance increases by approximately 0.04 of a standard deviation after the next failure.[21] Clearly, the data fail to support a simple story of bonus receipt as a cheap-to-acquire signal.

5.3 Testing the second alternate model: uncertain production technology and learning

As argued above, school administrators' asymmetric responses at the point of bonus discontinuity may reflect incomplete knowledge about the nature of incentivization and the production process more generally. Whereas the bonus-as-signal model predicts more significant responses to the bonus over time – because it has very little signal value at the beginning of time – the learning model suggests that agents will adopt rational behavior in the long run, while suggesting that behavioral asymmetries might not translate effectively into output asymmetries in the very short run. Indeed, our results above identifying stronger asymmetries among schools with a track record of poor performance is entirely consistent with the predictions generated in section 2.3. In this section, we present further tests based on principal experience levels – under the presumption that principals are analogous to the "agents" modeled in section 2.3 above.

Table 5 reproduces the basic results from table 3 by splitting schools into those headed by principals with low (less than 5 years as a principal), medium (5 to 10 years), and high (more than 10 years) experience.[22] Estimates indicate that school with principals of mid-level experience just below the bonus eligibility threshold exhibit significantly higher test score gains

[21] As seen in Table 3, reading results follow the pattern of estimates seen for math scores. However, most results are statistically insignificant, and reading results are suppressed from subsequent tables. Full tables with reading score outcomes are available online at sites.google.com/site/tomsyahn/.

[22] The principal experience coding is based on total number of years holding the title of "Principal," rather than tenure at an individual school. Splitting the sample by tenure at a given school yields similar results, consistent with the notion that information about the production process at one school may not translate directly to another.

relative to barely-eligible schools. The estimated effect ranges from 0.0442 to 0.0536. These are large improvements, comparable to the impact of *replacing* an ineffective principal as part of the NCLB sanctions (Ahn and Vigdor 2013). Similar effects are not observed for principals with low or high levels of experience. In fact, not only are the discontinuities statistically insignificant from zero, the estimated magnitudes are also 30 to 70 percent smaller than those estimated for mid-level experience principals. This pattern of discontinuity estimates across principal experience is consistent with our model of imperfect information and learning.[23]

Splitting schools by accountability history and principal experience further supports our model. Table 6 presents these results. While all schools with spotless records do not have statistically significant discontinuities in response to their first failure, the exceptionally small estimates for schools with highly experienced principals indicates that these schools do not respond at all to the first failure, supporting our hypothesis that they rationally attribute the aberrant result to chance and do not implement wholesale changes in response. The over-reaction around the bonus threshold is concentrated among schools with histories of poor performance headed by principals with mid-level experience, with a response of about 0.06 of a standard deviation.[24] Once again, this is consistent with a model where principals learn on the job systematically over-emphasize the role of effort relative to luck in scenarios where they have been repeatedly exhorted to exert more effort.

To buttress our results from the non-parametric RD analysis, we run an alternate set of parametric RD estimations. The benefits of this analysis is two-fold: the preferred specification

[23] Some of our insignificant results have large standard errors, due to the low-power of regression discontinuity. We use parametric RD below to add to the evidence that we are estimating a real "zero" impact, and not a true impact drowned out by noise from RD.

[24] It is interesting to note that although there seems to be some response by highly experienced principals upon additional failures, once principals with very short tenures (less than 3 years) are eliminated, the (still insignificant) discontinuity drops to similar magnitude as schools with no failures. We hypothesize that some of the short-tenure, yet highly experienced principals may be principals tasked with resuscitating chronically underachieving schools under the NCLB sanction regime.

(labeled (3) below) allows us to use the entire sample of schools in one regression, increasing the power, and it allows us to "test" whether the non-linear effects of experience we observe in the non-parametric RD actually exists.

To start, the simplest base-line parametric regression discontinuity framework is defined as follows:

$$\Delta math_{it} = \beta_0 + X_{it}\beta_1 + \beta_2 I_{it} + \beta_3 LEV_{it} + \beta_4 I_{it} \cdot LEV_{it} + f(\Delta y_{i,t-1}, LEV_{it}) + f(I_{it} \cdot \Delta y_{i,t-1}, LEV_{it}) + \epsilon_{it} \quad (1)$$

As in the non-parametric version, the dependent variable, $\Delta math_{it}$, is the change in math score in year t. The assignment variable $\Delta y_{i,t-1}$ is school *i*'s normalized academic growth rate in year (t-1). LEV_{it} is the experience level for the principal in school *i*. I_{it} is an indicator variable which equals one if the school qualified for the bonus in year (t-1), with $\Delta y_{i,t-1} \geq 0$. X_{it} is a vector of control variables, such as percent female, minority, limited English proficient, and free and reduced price lunch-eligible students, as well as year and school dummy variables. The $f(\cdot)$ term is a flexible function. Each argument inside the function has polynomial controls in its own and all possible interaction terms.[25] The inclusion of the second $f(\cdot)$ with $I_{it} \cdot \Delta y_{i,t-1}$ allows the conditional mean function on the other side of the discontinuity to have a different shape. The idiosyncratic error term is represented by ϵ_{it}. All regressions are weighted by the number of students in the school.

The parameter of interest here is β_4, which captures the discontinuity in test score growth at the state defined bonus threshold. Following the results of the non-parametric RD results, we would expect β_4 to be negative.

[25] For example, if there are only two experience levels, $f(y, LEV) = \sum_{k=1}^{n} \gamma_k y^k + \theta_k LEV^k + \delta_k (y \cdot LEV)^k$.

In order to ensure that the results are not being driven by the artificial cut-off values for experience, the discontinuity is re-estimated with a continuous measure of experience:

$$(2) \quad \Delta math_{it} = \beta_0 + X_{it}\beta_1 + \beta_2 I_{it} + \beta_3\, exp_{it} + \beta_4\, exp_{it}^2 + \beta_5 I_{it} \cdot exp_{it} + \beta_6 I_{it} \cdot exp_{it}^2 + f\left(\Delta y_{i,t-1}, exp_{it}, exp_{it}^2\right) + f\left(I_{it} \cdot \Delta y_{i,t-1}, exp_{it}, exp_{it}^2\right) + \epsilon_{it}$$

Now, the parameters of interest are $\beta_5$and β_6. If experience is non-linear as portrayed in the non-parametric RD results, we would expect β_5 to be negative and β_6 to be positive.

Finally, to examine the impact of accountability history of the school, we include a dummy variable which equals one for schools that have consistently qualified for the bonus, Z_{it}, interacted with the experience measure. This represents our preferred specification for the parametric RD regression.

$$(3) \quad \Delta math_{it} = \beta_0 + X_{it}\beta_1 + \beta_2 I_{it} + \beta_3\, exp_{it} + \beta_4\, exp_{it}^2 + \beta_5 I_{it} \cdot exp_{it} \cdot Z_{it} + \beta_6 I_{it} \cdot exp_{it}^2 \cdot Z_{it} + \beta_7 I_{it}\, exp_{it}(1 - Z_{it}) + \beta_8 I_{it}\, exp_{it}^2(1 - Z_{it}) + f\left(\Delta y_{i,t-1}, exp_{it}, exp_{it}^2, Z_{it}\right) + f\left(I_{it} \cdot \Delta y_{i,t-1}, exp_{it},\ exp_{it}^2\ , Z_{it}\right) + \epsilon_{it}$$

We are interested in β_5, β_6, β_7, and β_8. We would expect $\beta_5$and $\beta_6$to be statistically insignificant, β_7 to be negative, and β_8 to be positive. The results for the relevant parameters for the three specifications above are presented in Table 7.[26]

As demonstrated, the relevant parameter estimates in all three specifications are consistent with the non-parametric RD results. In particular, in specification (3), the shape of the experience response shows that the peak of response is between 6 and 7 years of experience.

5.4 Assessing the role of confusion: mixing up accountability incentives

As described in section 3 above, North Carolina's accountability program paid cash bonuses on the basis of year-over-year test score gains. There is no reason to believe that test

[26] Further robustness checks with a dummy variable for the maximum observable value of experience (13 years) yielded no qualitative differences.

score gains are easier to produce among students in close proximity to the proficiency threshold. Indeed, depending on the scale properties of the test, large gains may be easiest to produce in the tails of the distribution. By contrast, the Federal No Child Left Behind system incentivizes proficiency rates, which quite clearly gives schools an incentive to target instructional resources on those students in close proximity to the state-defined proficiency threshold (Neal and Schanzenbach, 2010). Particularly given the low degree of correlation between bonus receipt and NCLB sanction status shown in Table 1, evidence that principals focus on students near the proficiency threshold when the state system has given them strong reason to focus on generating gains would be consistent with a fundamental confusion regarding the nature of the incentive system.

Table 8 presents an analysis of math score improvements for students stratified by initial achievement level. The unit of observation continues to be the school/year, but only data on students in a given performance category is used to compute the average test score growth statistic. We see that statistically significant discontinuities exist for students at achievement levels II and III. The border between these levels is the bar for grade-level proficiency as defined by the state. While it is clear that schools that just failed to qualify for the bonus respond substantively, the apparent focus on students near the proficiency level suggests that they may not fully understand the North Carolina bonus program.

In all, then, evidence does suggest some basic confusion regarding the operation of accountability incentives, and thus it is difficult to rule out a basic behavioral hypothesis of sheer confusion in explaining why irrational responses to irrelevant information occur. Our earlier results, however, suggest a possible framework for understanding both confusion and the learning process which might prove useful in the design of future incentive schemes.

6. Conclusion

Across many domains, economic agents – acting as consumers or producers – exhibit a tendency to be excessively sensitive to discrete signals based on continuous variables that are available in their information sets. These tendencies are prone to be exposed empirically in studies using the regression discontinuity design, which in some applications compares individuals receiving nearly identical continuous information but differing discrete signals based on that continuous indicator.

This paper, beyond identifying another scenario where such a behavioral quirk can be observed, offers some insight into the nature of the behavior. While it is difficult to empirically exclude the basic hypothesis that agents act irrationally, we show evidence that this behavior is consistent with a rational learning model, where agents are at first unsure how to react to the information conveyed to them and act on the basis of prior beliefs that yield to experience over time. In our case, the implication is that school administrators obtaining information about their school's performance learn over time to rationally ignore discrete signals and pay greater attention to the underlying continuous information. Such a model is also consistent with, for example, the empirical observation that wholesale prices for used cars, which are established in transactions involving experienced buyers and sellers, do not exhibit the same reactivity to the leftmost odometer digit as retail prices, which reflect transactions involving a typically inexperienced buyer.

Beyond exploring possible explanations for seemingly irrational behavior, this paper sheds light on issues in mechanism design. Agents operating in an incentive system may require repeated iterations to gain a picture of how their efforts, as well as factors beyond their control,

map into outcomes. While this picture is incomplete, agents may exhibit behaviors that appear to be irrational, as was found in this case. This basic insight might help to explain why, for example, experimental evaluations of one-shot educational incentive schemes sometimes find no significant effects, even while more systematic evaluations of schemes implemented over multiple years suggest the existence of important relationships (Springer et al. 2010). The behaviors described in this paper suggest that it may be inappropriate to evaluate incentive systems on the basis of short-term implementation experiments.

As the administrators of incentive schemes collect information on their effectiveness, they often face a temptation to "tweak" the system in order to increase the amount of effort incented per dollar spent. The results shown here suggest a countervailing cost to the potential benefits of such tweaking: upon changing the system, principals may force agents to engage in extended and potentially unproductive experimentation to determine their optimal response to the new regime. In some cases, these costs may outweigh the long-term benefits of better calibrating the incentive system.

References

Ahn, T. (2014) "A Regression Discontinuity Analysis of Graduation Standards and Their Impact on Students' Academic Trajectories." *Economics of Education Review*, 37, Pages 64-75.

Ahn, T. and J. L. Vigdor (2013) "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina, " *Working Paper*.

Allcott H. (2011) "Social norms and energy conservation," *Journal of Public Economics, Volume 95, Issues 9–10,* Pages 1082-1095.

Anderson, M., and J. Magruder. (2011) "Learning From The Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database." *The Economic Journal 122.563,* Pages 957-989.

Bala, V. and S. Goyal (1995) "A Theory of Learning with Heterogeneous Agents," *International Economic Review*, v. 36(2), Pages 303-323.

Berger J. and D. Pope, (2011) "Can Losing Lead to Winning?" *Management Science 57:5*, Pages 817-827.

Berry, D. and B. Friestedt (1985) *Bandit Problems: Sequential Allocation of Experiments* (Chapman and Hall, London)

Bertola, G. and R. Caballero (1990) "Kinked Adjustment Costs and Aggregate Dynamics" *NBER Macroeconomics Annual 1990, Volume 5.*

Card D., A. Mas, and J. Rothstein, (2008) "Tipping and the dynamics of segregation," *Quarterly Journal of Economics, Volume 123 (1)*, Pages 178 – 218.

Chakrabarti, R. (forthcoming) "Incentives and Responses under No Child Left Behind: Credible Threats and the Role of Competition." *Journal of Public Economics.*

Chiang, H. (2009) "How Accountability Pressure on Failing Schools Affects Student Achievement." *Journal of Public Economics* v.93, Pages1045-1057.

Easley, D. and N. Kiefer (1988) "Controlling a Stochastic Process with Unknown Parameters" *Econometrica* 56, Pages 1045-1064.

El-Gamal M., and R. Sundaram, (1993) "Bayesian Economists, Bayesian Agents: An Alternative Approach to Optimal Learning," *Journal of Economic Dynamics and Control*, v. 17(3), Pages 355-383.

Figlio D. N., and M. E. Lucas (2004) "What's in a Grade? School Report Cards and the Housing Market." *American Economic Review, 94(3),* Pages 591-604.

Figlio D. N., and L. W. Kenny (2009) "Public sector performance measurement and stakeholder support," *Journal of Public Economics, Volume 93, Issues 9–10*, Pages 1069-1077.

Gabaix, X. (2014) "A Sparsity-based Model of Bounded Rationality." *Working Paper*.

Gigerenzer, G. and D. Goldstein (1996). "Reasoning the fast and frugal way: Models of bounded rationality," Psychological Review 103 (4), Pages 650–669.

Hahn, J., P. Todd, and W. van der Klaauw (2001) "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1), Pages 201-209.

Hanushek, E.A. and M.E. Raymond (2005) "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24(2), Pages 297-327.

Imbens, G.W. and K. Kalyanaraman (2009) “Optimal Bandwidth Choice for the Regression Discontinuity Estimator.” *NBER Working Paper #14726*.

Imbens, G.W. and T. Lemieux (2008) “Regression Discontinuity Designs: A Guide to Practice.” *Journal of Econometrics* 142(2), Pages 615-635.

Jacob, B.A. (2005) “Accountability, Incentives, and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools.” *Journal of Public Economics* v.89, Pages 761-796.

Jin G., and P. Leslie. (2003) “The Effect Of Information On Product Quality: Evidence From Restaurant Hygiene Grade Cards.*” The Quarterly Journal of Economics 118.2*, Pages 409-451.

Jinnai, Y (2013) “The Effects of a Teacher Performance-Pay Program on Student Achievement: A Regression Discontinuity Approach” *Working Paper*.

Kiefer, N. (1989) "A Value Function Arising in the Economics of Information," *Journal of Economic Dynamics and Control* 13, Pages 201-223.

Lacetera N., D. G. Pope, and J. R. Sydnor (2012) "Heuristic Thinking and Limited Attention in the Car Market," *American Economic Review, vol. 102(5),* Pages 2206-36.

Martinez, E. (2010) “Do Housing Prices Account for School Accountability?” Working Paper.

Mullainathan, S. (2002) "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economic, vol. 117(3*), Pages 735-774.

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

Neal, D. and D. Schanzenbach (2010) “Left Behind By Design: Proficiency Counts and Test-Based Accountability.” *Review of Economics and Statistics* 92(2), Pages 263-283.

Papay J., R. J. Murnane, and J. B. Willett (2011) "How Performance Information Affects Human-Capital Investment Decisions: The Impact of Test-Score Labels on Educational Outcomes" *NBER Working Paper No. 17120*

Reback, R. (2008) “Teaching to the Rating: School Accountability and the Distribution of Student Achievement.” *Journal of Public Economics* 92(5-6), Pages 1394-1415.

Rothschild, M. 1974 "A Two-armed Bandit Theory of Market Pricing," *Journal of Economic Theory* 9, Pages 185-202.

Rouse, C.E., J. Hannaway, D. Goldhaber, and D. Figlio (2013) “Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure.” *American Economic Journal: Economic Policy* v.5, Pages 251-281.

Scott F. and A. Yelowitz. (2010) "Pricing Anomalies in the Market for Diamonds: Evidence of Conformist Behavior" *Economic Inquiry 48.2,* Pages 353-368.

Sarver, T. (2008) "Anticipating Regret: Why Fewer Options May Be Better," *Econometrica, 76*, Pages 263–305.

Sims, C. (2003) "Implications of Rational Inattention," *Journal of Monetary Economics*, 50, Pages 665-690.

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

Springer, M.G., J. Pane, V. Le, D. McCaffrey, S.F. Burns, L. Hamilton, and B. Stecher (2012) "Team Pay for Performance: Experimental Evidence from the Round Rock Pilot Project on Team Incentives." *Educational Evaluation and Policy Analysis* v.34, Pages 367-390.

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.

West, M.R. and P.E. Peterson (2006) "The Efficacy of Choice Threats Within School Accountability Systems: Results from Legislatively Induced Experiments." *Economic Journal* 116(510), Pages C46-C62.

Yuan, K., V. Le, D.F. McCaffrey, J.A. Marsh, L. Hamilton, B. Stecher and M.G. Springer (2013) "Incentive Pay Programs Do Not Affect Teacher Motivation or Reported Practices: Results from Three Randomized Studies." *Educational Evaluation and Policy Analysis* v.35, Pages 3-22.

Figures and Tables

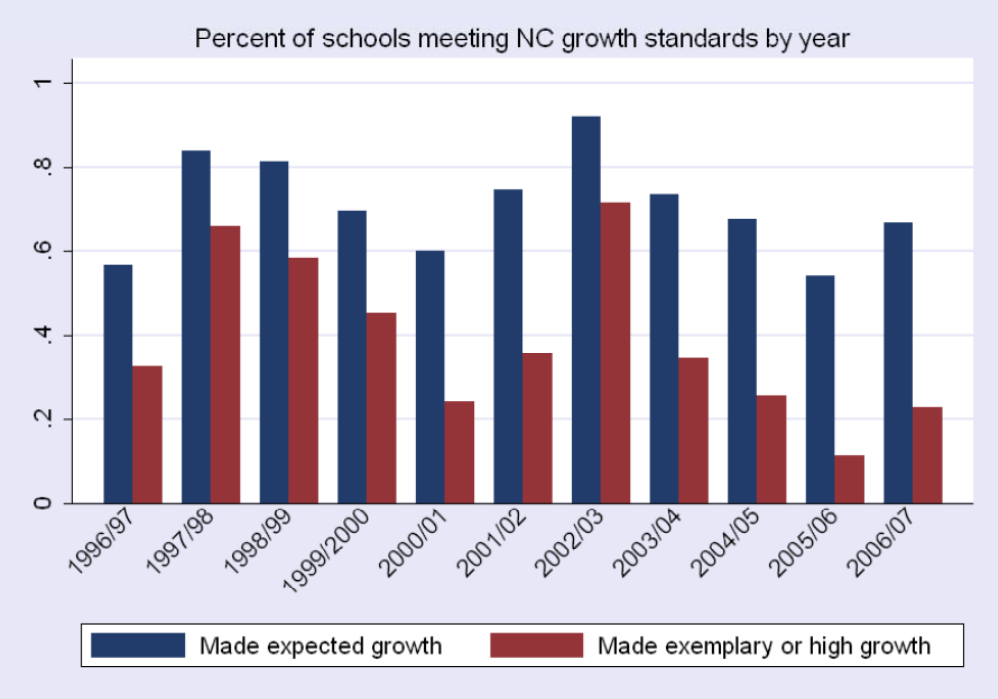

Figure 1: Proportion of schools qualifying for NC bonus. (From Vigdor 2009)

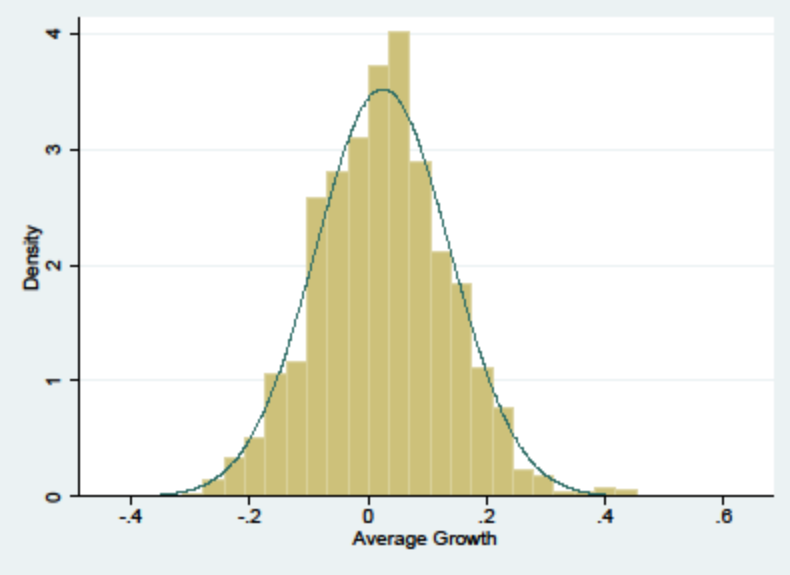

Figure 2: Density of observations across assignment variable.

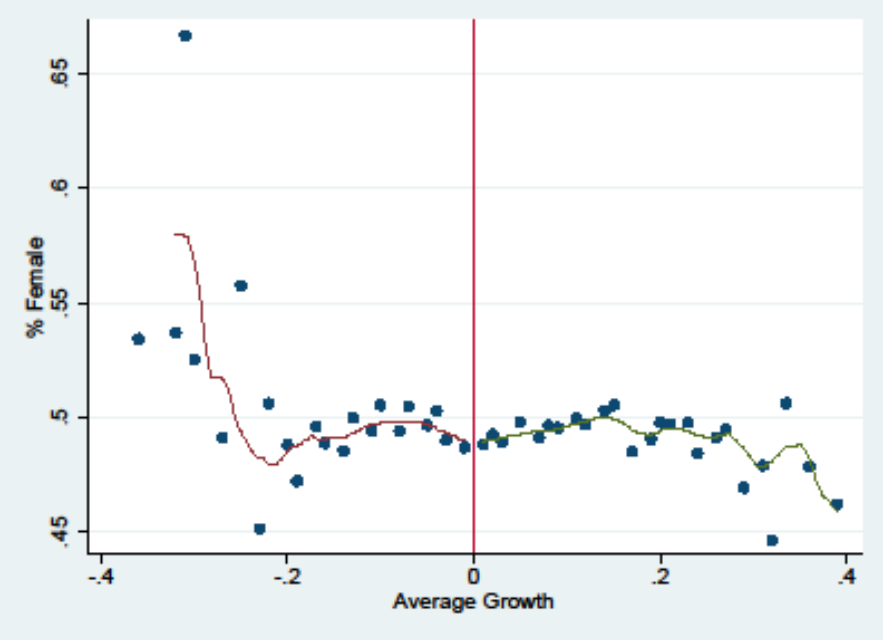

Figure 3: Placebo female percent. Figure is generated with local polynomial of degree zero. Local averages presented with 50 bins. Covariates are excluded. Observations for running variable less than -0.4 or greater than 0.4 (which comprise approximately 3.7 % of observations) are dropped for presentation purposes.

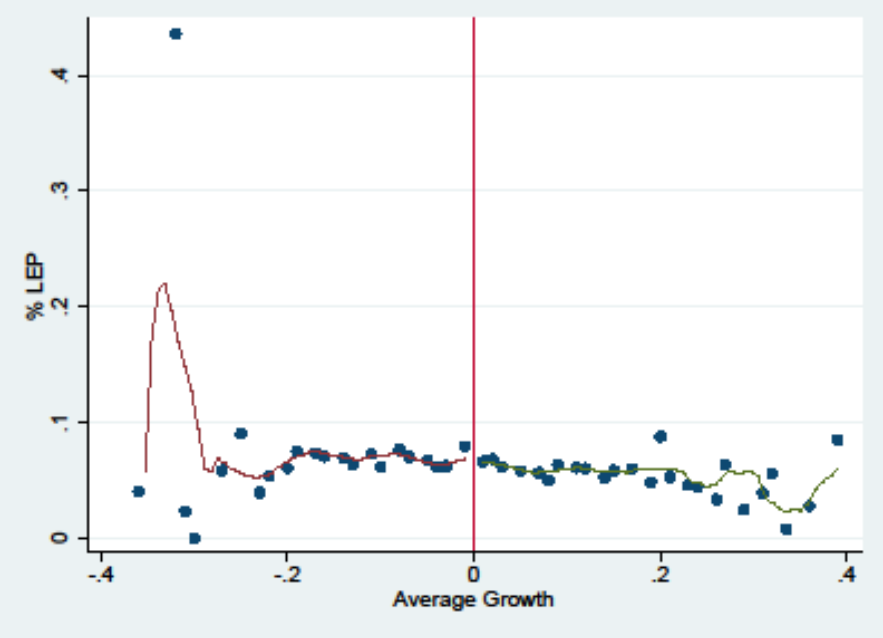

Figure 4: Placebo of LEP percent. Figure is generated with local polynomial of degree zero. Local averages presented with 50 bins. Covariates are excluded. Observations for running variable less than -0.4 or greater than 0.4 (which comprise approximately 3.7 % of observations) are dropped for presentation purposes.

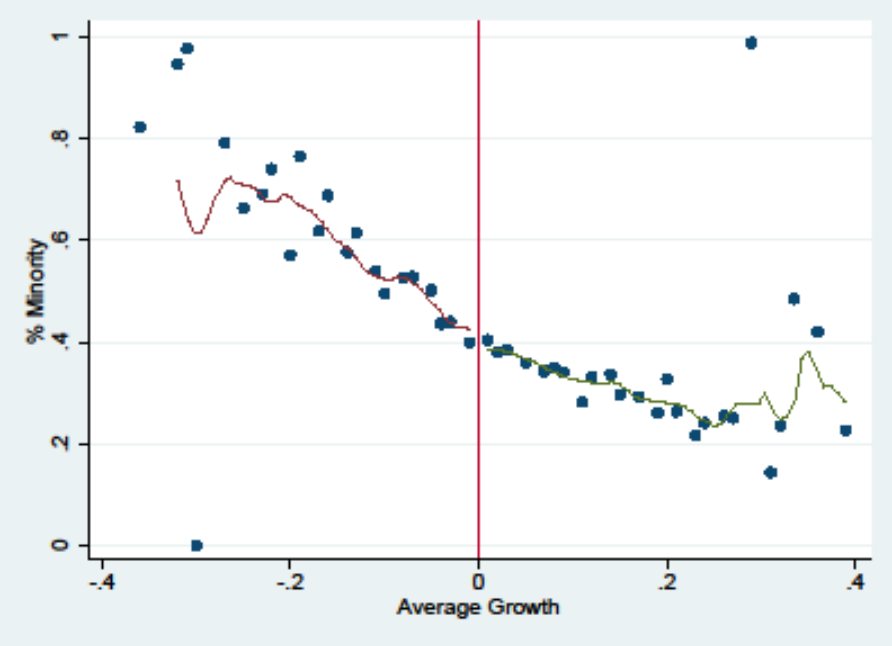

Figure 5: Placebo of minority percent. Figure is generated with local polynomial of degree zero. Local averages presented with 50 bins. Covariates are excluded. Observations for running variable less than -0.4 or greater than 0.4 (which comprise approximately 3.7 % of observations) are dropped for presentation purposes.

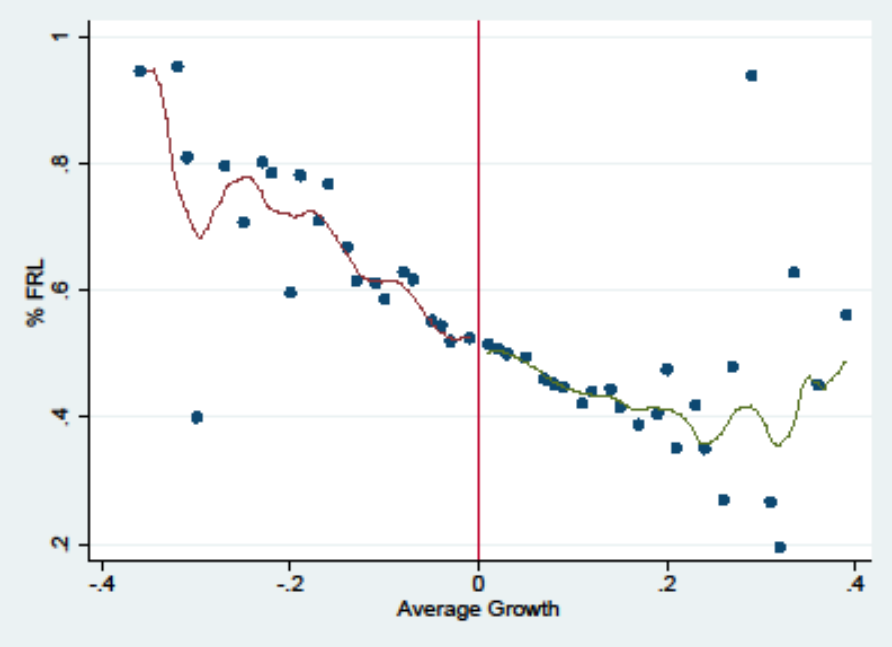

Figure 6: Placebo of percent free and reduced price lunch students. Figure is generated with local polynomial of degree zero. Local averages presented with 50 bins. Covariates are excluded. Observations for running variable less than -0.4 or greater than 0.4 (which comprise approximately 3.7 % of observations) are dropped for presentation purposes.

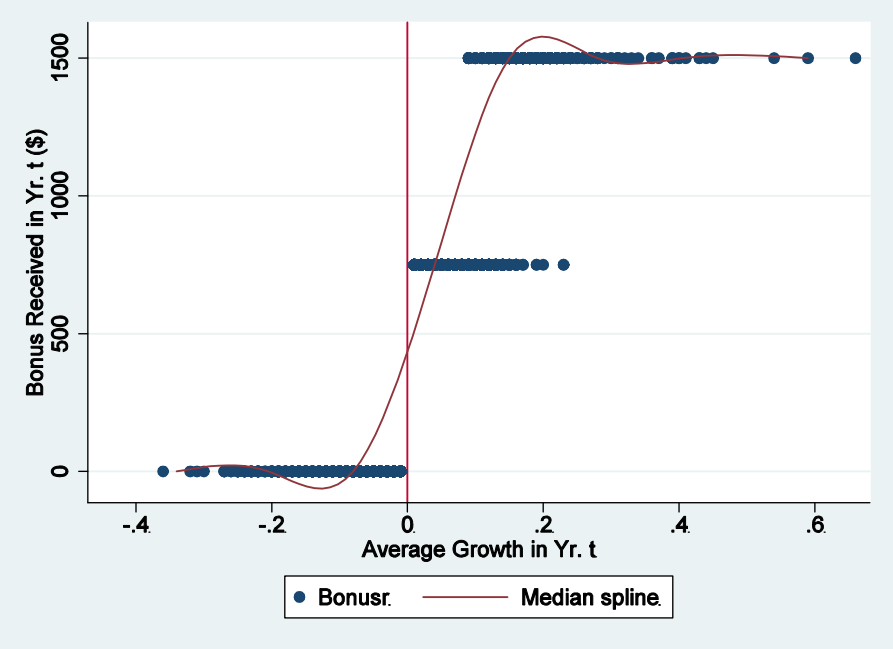

Figure 7: Existence of discontinuity in probability of bonus receipt at policy change.

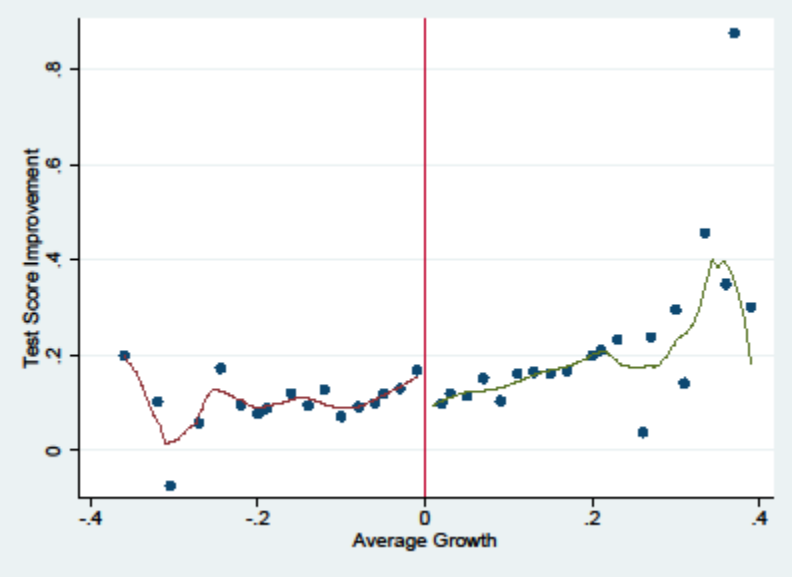

Figure 8: Simple RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t. Figure is generated with local polynomial of degree zero. Local averages presented with 50 bins. Covariates are excluded. Observations for running variable less than -0.3 or greater than 0.4 (which comprise approximately 3.7 % of observations) are dropped for presentation purposes.

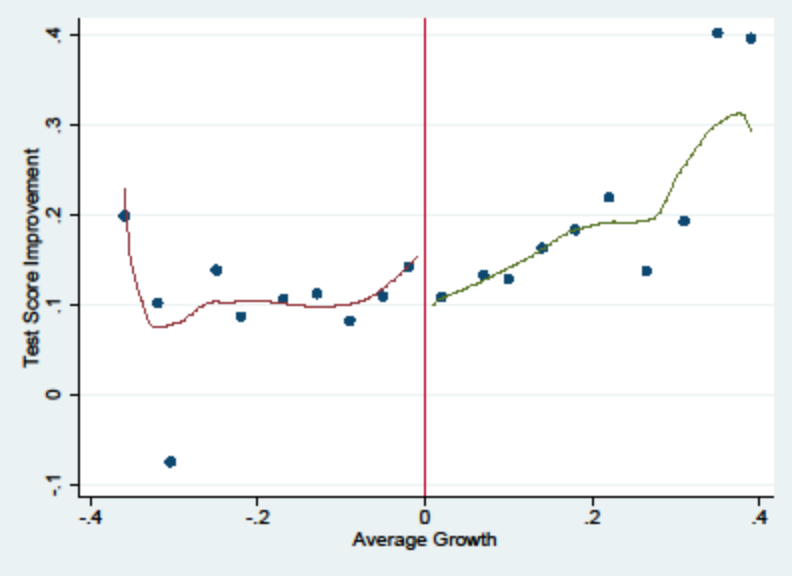

Figure 9: Simple RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t. Figure is generated with local polynomial of degree zero. Local averages presented with 25 bins. Covariates are excluded. Observations for running variable less than -0.3 or greater than 0.4 (which comprise approximately 3.7 % of observations) are dropped for presentation purposes.

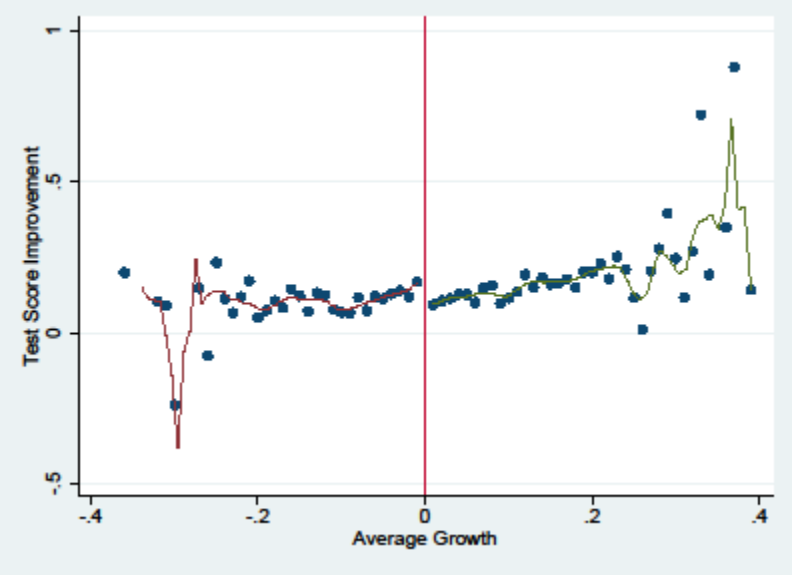

Figure 10: Simple RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t. Figure is generated with local polynomial of degree zero. Local averages presented with 100 bins. Covariates are excluded. Observations for running variable less than -0.3 or greater than 0.4 (which comprise approximately 3.7 % of observations) are dropped for presentation purposes.

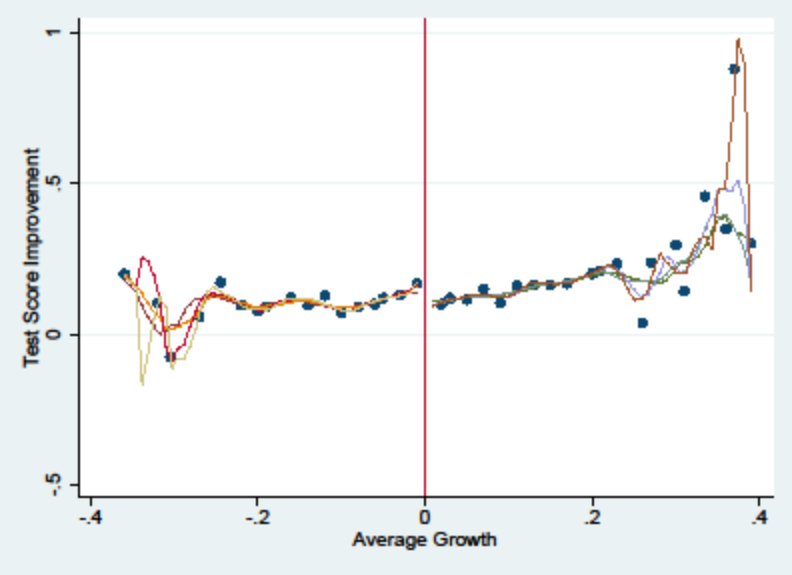

Figure 11: Simple RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t. Figure is generated with local polynomials of degree zero to degree three. Covariates are excluded. Observations for running variable less than -0.3 or greater than 0.4 (which comprise approximately 3.7 % of observations) are dropped for presentation purposes.

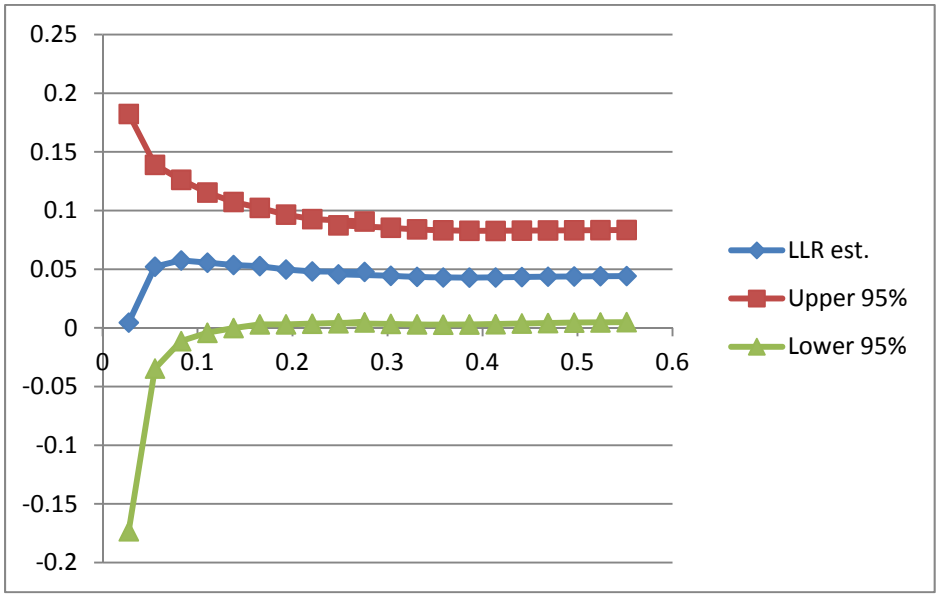

Figure 12: Local Linear Regression Estimates with Varying Bandwidths. Math score improvement in year t+1 conditional on just being below qualification for the bonus in year t in schools with mid-level experience principals.

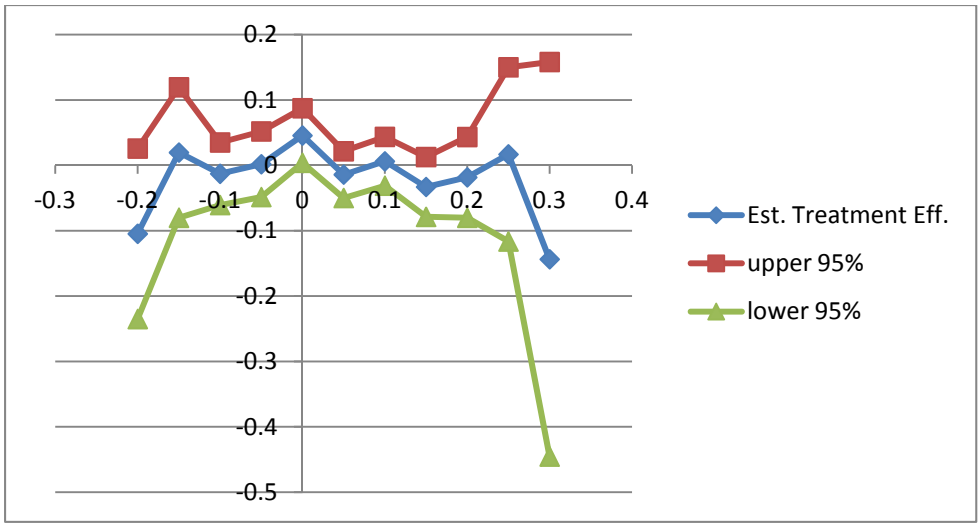

Figure 13: Local Linear Regression Estimates with Artificial Cut-Off Points. Math score improvement in year t+1 conditional on just being below qualification for the bonus in year t in schools with mid-level experience principals.

Table 1: AYP and ABC Status

		ABC	
		Yes	No
AYP	Yes	956	284
	No	423	635

Table 2: Summary Statistics

Variable	Mean (Std. Dev.)
Δ math score	0.1385 (0.6007)
Δ reading score	-0.0544 (0.6205)
math proficiency level	2.8763 (0.8399)
reading proficiency level	3.326 (0.7676)
% minority	0.3803 (0.4855)
% FRL eligible	0.4677 (0.4990)
% female	0.4948 (0.5000)
% LEP	0.0642 (0.2450)
Years since last bonus	0.6663 (0.9780)
Number of no bonus years in last 5 years	1.1577 (1.1944)
Years since AYP made	0.6692 (1.0012)
Number of AYP failed since 2002-03	1.2075 (1.1467)
School size	211.5 (113.2)
Principal years of experience	6.1873 (3.8569)
Principal years of tenure at current school	4.4226 (3.1135)
Observations	338,240

Note: 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. Minority students are blacks, Hispanics, and American Indians.

Table 3: RD estimates of the impact of failing to receive the bonus

	Reading (n=2,276)	Math (n=2,276)
At optimal bandwidth	-0.0098	-0.0351**
	(0.0111)	(0.0140)
Half optimal bandwidth	-0.0226	-0.0458**
	(0.0155)	(0.0192)
Twice optimal bandwidth	-0.0057	-0.0260**
	(0.0096)	(0.0116)

Note: Standard errors in parentheses. Dependent variables defined in the text. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications control for minority percent, free/reduced price lunch eligible percent, and school size. *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 4: RD estimates by ABC bonus history

Qualified for bonus every year	Reading (n=817)	Math (n=817)
At optimal bandwidth	-0.0077	-0.0381
	(0.0194)	(0.0361)
Half optimal bandwidth	-0.0246	-0.0384
	(0.0252)	(0.0535)
Twice optimal bandwidth	-0.0045	-0.0440
	(0.0179)	(0.0304)
Failed to qualify for bonus in at least 1 year	**Reading (n=1,431)**	**Math (n=1,431)**
At optimal bandwidth	-0.0139	-0.0415***
	(0.0137)	(0.0160)
Half optimal bandwidth	-0.0324	-0.0485**
	(0.0204)	(0.0221)
Twice optimal bandwidth	-0.0118	-0.0292**
	(0.0117)	(0.0132)

Note: Standard errors in parentheses. Dependent variables defined in the text. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications control for minority status, free/reduced price lunch eligible, and school size.
*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 5: RD estimates (math, by experience level)

Years of Experience	Exp. > 10 (n=456)	5<=Exp.<=10 (n=824)	Exp.<5 (n=968)
At optimal bandwidth	-0.0302	-0.0456**	-0.0209
	(0.0310)	(0.0213)	(0.217)
Half optimal bandwidth	-0.0373	-0.0536**	-0.0507
	(0.0417)	(0.0274)	(0.0302)
Twice optimal bandwidth	-0.0161	-0.0442**	-0.0157
	(0.0253)	(0.0200)	(0.0175)

Note: Standard errors in parentheses. Dependent variables defined in the text. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications control for minority percent, free/reduced price lunch eligible percent, and female percent. Regression is weighted by school size.
*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 6: RD estimates by ABC bonus history (by experience level)

Qualified for bonus every year	Exp. > 10 (n=180)	5<=Exp.<=10 (n=285)	Exp.<5 (n=352)
At optimal bandwidth	0.0179	-0.0823	-0.0510
	(0.0665)	(0.0544)	(0.0557)
Half optimal bandwidth	-0.0095	-0.0746	-0.0618
	(0.0897)	(0.0685)	(0.0883)
Twice optimal bandwidth	0.0050	-0.0908*	-0.0549
	(0.0588)	(0.0522)	(0.0453)
Failed to qualify for bonus at least once	Exp. > 10 (n=276)	5<=Exp.<=10 (n=539)	Exp.<5 (n=616)
At optimal bandwidth	-0.0420	-0.0597**	-0.0109
	(0.0334)	(0.0273)	(0.0215)
Half optimal bandwidth	-0.0550	-0.0660*	-0.0317
	(0.0441)	(0.0376)	(0.0278)
Twice optimal bandwidth	-0.0314	-0.438**	-0.0168
	(0.0295)	(0.0229)	(0.0184)

Note: Standard errors in parentheses. Dependent variables defined in the text. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications control for minority percent (where appropriate), free/reduced price lunch eligible percent (where appropriate), and female percent. Regression is weighted by school size.

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 7: Parametric Regression Discontinuity Parameter Estimates

Specification (1)	Coefficient (Std. Error)
β_4: Low Experience	-0.0260 (0.0216)
β_4: Mid Experience	-0.0467** (0.0203)
β_4: High Experience	-0.0134 (0.0260)
Specification (2)	
β_5: Linear Experience	-0.0155** (0.0066)
β_6: Quadratic Experience	0.0010** (0.0005)
Specification (3)	
β_5: Linear Experience (0 Fails)	-0.0377 (0.0400)
β_6: Quadratic Experience (0 Fails)	0.0040 (0.0029)
β_7: Linear Experience (> 0 Fails)	-0.0671** (0.0339)
β_8: Quadratic Experience (> 0 Fails)	0.0054** (0.0024)
Observations	2,248

Note: Standard errors in parentheses. Dependent variables defined in the text. Specifications control for minority percent, limited English proficient percent, free/reduced price lunch eligible percent, female percent, and year and school dummies. Regression is weighted by school size. Cubic polynomial controls for assignment variables presented. Different degrees polynomial results and parameter estimates for all control variables available at: http://sites.google.com/site/tomsyahn/

*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level

Table 8: RD estimates by proficiency level

Level I: insufficient mastery	Math (n=2,078)
At optimal bandwidth	0.0121
	(0.0348)
Half optimal bandwidth	0.0092
	(0.0502)
Twice optimal bandwidth	0.0046
	(0.0283)
Level II: inconsistent mastery	Math (n=2,212)
At optimal bandwidth	-0.0342*
	(0.0182)
Half optimal bandwidth	-0.0324
	(0.0218)
Twice optimal bandwidth	-0.0336*
	(0.0175)
Level III: sufficient mastery	Math (n=2,195)
At optimal bandwidth	-0.0399**
	(0.0185)
Half optimal bandwidth	-0.0498*
	(0.0263)
Twice optimal bandwidth	-0.0317**
	(0.0152)
Level IV: superior mastery	Math (n=2,161)
At optimal bandwidth	-0.0261
	(0.0168)
Half optimal bandwidth	-0.0255
	(0.0221)
Twice optimal bandwidth	-0.0236
	(0.0152)

Note: Standard errors in parentheses. Dependent variables defined in the text. Bandwidth determination is by the Imbens-Kalyanaraman algorithm. Specifications control for minority status, free/reduced price lunch eligible, and school size.
*** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.