When Incentives Matter Too Much: Explaining Significant Responses to Irrelevant Information

Thomas Ahn[*]
University of Kentucky

Jacob L. Vigdor
Duke University

Draft
November, 2013

Abstract

When economic agents have access to 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 studies, many based on the regression discontinuity design, have contradicted this basic prediction. We examine one such scenario, involving the educational accountability system implemented in North Carolina public schools until 2009. Results are consistent with a model of learning and imperfect information. Among other findings, irrational responses to the discrete signal are concentrated in schools where principals have intermediate experience levels.

[*] 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). The authors are grateful to John Holbein for research assistance. Any opinions expressed herein are those of the authors and not any affiliated organization.

1. Introduction

Basic 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, including the one described in this paper, 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. 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 in differing categories are provided different information. In scenarios where subjects are aware of the continuous underlying measure, informing them of their assignment to a category on the basis of that measure should be irrelevant.

In the 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.

Confirming earlier work on North Carolina's bonus program, we demonstrate that personnel in schools falling below the threshold in fact exhibited significant improvements in performance in following years relative to personnel in schools just above the threshold. We describe and evaluate two possible explanations for this excessive response. The first posits that agents incur some cost to acquiring and processing a continuous measure, and thus rationally use the discrete signal as a cue that the marginal benefit of examining the data will exceed the marginal cost. The second introduces possible uncertainty regarding the data-generating process, and suggests that agents may exhibit inappropriate responses to the discrete information when they possess only a partial understanding of the process.

Empirical patterns fail to support the simple information cost story; asymmetric responses around the bonus threshold are not consistent with agents ignoring the signal unless it presents a surprise. The data are more consistent with a process of uncertainty about the data-generating process which abates over time. Bonus threshold asymmetry is most pronounced in schools served by principals with intermediate experience levels. We argue below that inexperienced principals show no response because they either do not know how to respond or do not pay sufficient attention to the signal. As principals gain experience, they better understand the nature of the signal and how to respond, but misinterpret the data-generating process. We present other results indicating that moderately experienced principals misunderstand the nature of the bonus program and incentives embedded therein.

Beyond offering insight into the nature of seemingly irrational responses to irrelevant information, 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 2013, Papay, Murnane, and Willett 2011).

In all these scenarios, a standard prediction is that a rational actor armed with the continuous measure should not exhibit any sensitivity to information regarding discrete codings based solely on that measure. 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 Basic scenario

There are numerous economic models of behavior that could be used to frame a discussion of overreactions to irrelevant information, but a standard principal-agent framework suits our empirical exercise below.

Suppose output y_{it} of employee i in period t, which in the context of educational production can be measured by improvements in student 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 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 η_{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.[1]

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 variation in η_{it}. Under a variety of assumptions regarding the evolution of η_{it}, we should expect no change in the optimal effort choice as a function of incentive receipt. Were η_{it} entirely uncorrelated across years, rational employees would clearly behave exactly the same in the subsequent period. Even under non-random evolution of η_{it}, so long as $E(\eta_{it+1} \mid \eta_{it})$ 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 η_{it}. The most plausible scenario involving a discontinuous behavior change at the point of discontinuity would

[1] 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.

occur if there were no year-to-year variation whatsoever in η_{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.[2]

2.2 Alternate models yielding predictions of discontinuous change

In the context of educational production, it is unrealistic to think of an outcome such as a classroom's change in standardized test scores as purely deterministic. Thus 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. Even simple incomplete information models, however, seem unlikely to generate predictions of rational asymmetry in response. Consider, for example, a scenario where agents have diffuse priors over their own ability and the relative importance of idiosyncratic shocks relative to heterogeneity in ability. Even in this case, agents on either side of a discrete threshold should anticipate a nearly identical distribution of potential outcomes in the next period and thus respond nearly identically.

A variation on this theme would invoke both ability uncertainty and costs to the agent of observing the realization of η_{it}. 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 η_{it} agents might find it optimal to pay the information

[2] 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 might rationally withhold their cars from the market given consumer discounting.

cost only in the event of an unexpected shock, such as the failure to receive the bonus payment after a steady period of receiving it, as the investment in the information cost would yield a return in the form of resolving uncertainty regarding the potential return to altering effort.[3] If the agent were to learn that their performance lay significantly below the threshold, they might choose to reduce their effort; if close to the threshold they might instead 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. 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.

A second variation on the imperfect information theme regards uncertainty in the nature of the production process itself. Suppose agents begin their careers with prior beliefs along multiple dimensions. They are uncertain of their innate ability and uncertain of the importance of idiosyncratic factors and their own effort relative to ability in the production process. 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, effort, and innate ability. For a particular type of agent, however – those whose prior beliefs emphasize effort and downplay luck – there is reason to expect an asymmetric response around the bonus threshold. Agents with

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

these beliefs will be motivated to exert greater effort if just below the threshold, but will expect low returns to further effort above the threshold.

It is important to emphasize that not all agents may begin with the expectation that effort matters and luck does not. At the other extreme, agents who begin with the belief that luck matters to the exclusion of effort will have little reason to exert additional effort on either side of the bonus threshold.

As agents gather more information, in particular observing the imperfect correlation between their own effort levels and output, one expects asymmetries in response to dissipate. Fully-informed agents would rationally expect heightened returns to effort when their current effort level has left them close to the boundary. The slowest rate of convergence to fully-informed, rational behavior should be among agents who vary their effort levels considerably and observe concomitant output changes. An agent who begins with output far below the threshold and believes that effort matters, for example, might continually increase effort until they reach it, at which point they may display asymmetric response even at an advanced stage.

To summarize, then, one alternate explanation for asymmetric responses at the informational discontinuity emphasizes the use of the information as a shortcut and suggests that the agents exhibiting the discontinuity should be those witnessing a greater "surprise." This explanation suggests that if anything asymmetries should occur after agents accrue more information, because the potential for surprise is greater once prior beliefs have been solidified. A second explanation emphasizes uncertainty regarding the roles of effort and luck in the production process, and predicts asymmetric responses specifically among agents who believe effort matters much more than luck at a stage when they are relatively uninformed.

Beyond these fundamentally rational explanations for the patterns observed in the data, there is the simpler but less classically economic hypothesis that agents are simply confused: that they actually believe that the incentive system somehow inoculates them against needing to exert effort in subsequent periods. Given the rapid growth of behavioral economics research demonstrating cognitive limitations in deciphering incentive schemes, we do not mean to rule out such explanations. By offering rational incomplete information arguments, however, we hope to derive more theoretically justified implications for improving the design of incentive schemes.

The discussion to this point has analogized agents to teachers in an educational setting. Given the nature of the North Carolina bonus program, described below, it is more accurate to think of the "principal" as the state government and the "agent" as a school-level administrator with the authority to direct a set of sub-agents – teachers – regarding expected effort levels.

As a final caveat, the characterization of the education process as a simple translation of one input – effort – into a unidimensional output clearly oversimplifies reality. In reality, educators engage in a myriad of behaviors and efforts to improve amount to re-adjustments along multiple dimensions. It may be unreasonable to expect school administrators to hit on the perfect formula for improving performance on the first try. Indeed, evidence below suggests that there is a "learning curve" for improving the education process that is traveled somewhat more rapidly than the "learning curve" for understanding the relative importance of these efforts and luck.

3. 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 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 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.

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 underscore 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 continuous measure of test score growth.

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.[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 in schools serving 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

[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

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.

Student-level data is linked to the payroll data for teachers and administrators to track the experience level of the principals at the schools. 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 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]

---

upper grades may move classes and teachers from subject to subject, the teacher utility maximization problem is significantly complicated.

[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.

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. 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 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.[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.

[7] 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.

Schools clearly have an incentive to qualify for bonus payments, but it is not clear that the assignment variable is directly manipulable. 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. Consistent with Figure 1, which shows just over half of eligible schools qualifying for a bonus in these school years, 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 - 6 show results from 'placebo' regression discontinuity analysis with school minority percentage, free/reduced lunch percentage, female percentage, and limited English proficient percentage as the 'outcome' variables, respectively. 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 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 these covariates and average growth score, indicating that school that perform better have lower proportions of minority and free/reduced price lunch students, consistent with expectations.

Figure 7 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]

5. Results

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.

These are substantial improvements, 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.

[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.

For reading scores, the pattern of discontinuity estimates across bandwidths is similar to math, with point estimates larger at narrower bandwidths. However, the estimates are statistically insignificant. This is in line with most of the literature that finds teachers and schools less able to impact reading scores compared to math scores. From this point forward, our discussion will focus on math results as reading results are consistently statistically insignificant in all specifications.

5.2 Testing the first alternate model: bonus as signal

The model outlined above suggests that schools act to assess and potentially reoptimize 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 reoptimization 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, such that any subsequent failures cannot result in additional gains.[9] 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, academic performance increases by approximately 0.04 of a standard deviation after the next failure.[10] 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 "rational" responses will emerge only over time. The learning model does not necessarily suggest a monotonic pattern, however. Indeed, one interpretation of the results above is that chronically failing schools continuously increase their "effort" and thus never accurately parse the roles of effort and luck. Their over-attribution

[9] 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.

[10] 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/.

of outcomes to effort leads them to react asymmetrically once they reach the vicinity of the bonus threshold. Schools that have consistently earned the bonus, however, never face the same incentives to increase effort and thus form fairly accurate beliefs regarding the role of luck quickly. In this section, we present further tests based on principal experience levels.

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.[11] Estimates indicate that school with principals of mid-level experience just below the bonus eligibility threshold exhibit significantly higher test score gains 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.

Splitting schools by accountability history and principal experience further buttresses our argument. 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 experience 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

[11] 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.

headed principals with mid-level experience, with a response of about 0.06 of a standard deviation.[12] Once again, this is consistent with a model where principals learning 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.

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 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 7 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

[12] 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.

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 about the relevance of

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 (cite Springer et al. pay-for-performance paper, literature on bonuses more generally). 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. (2013) "A Regression Discontinuity Analysis of Graduation Standards and Their Impact on Students' Academic Trajectories." Forthcoming in *Economics of Education Review.*

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:* 957-989.

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

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

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):* 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.

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

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): 615-635.

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*: 409-451.

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.

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

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

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 and Tables

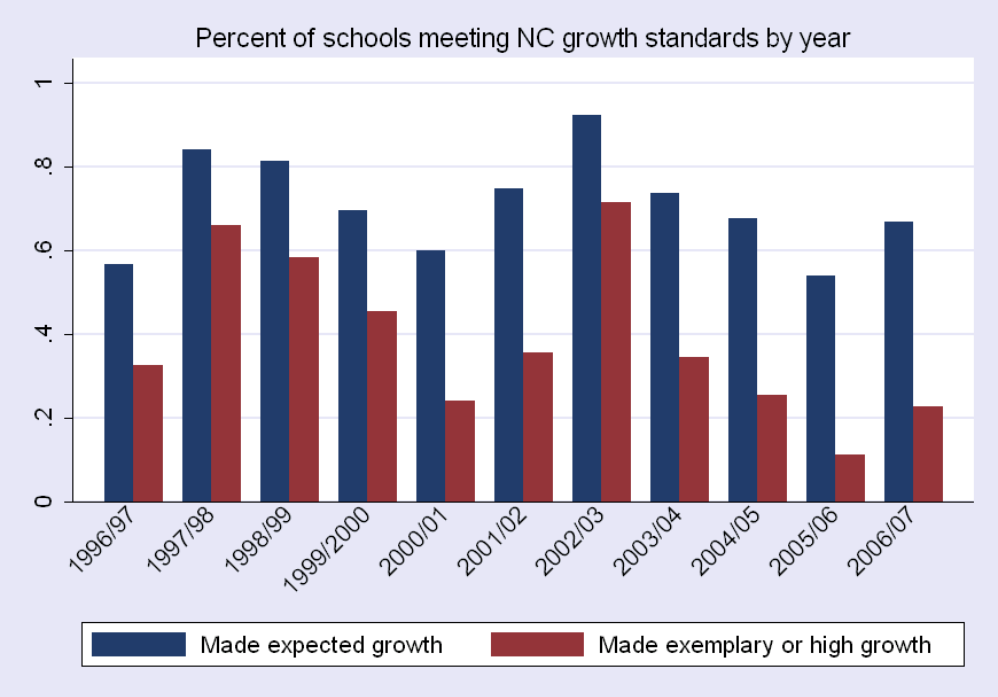

Figure 1: Proportion of schools qualifying for NC bonus. (From Vigdor 2009)

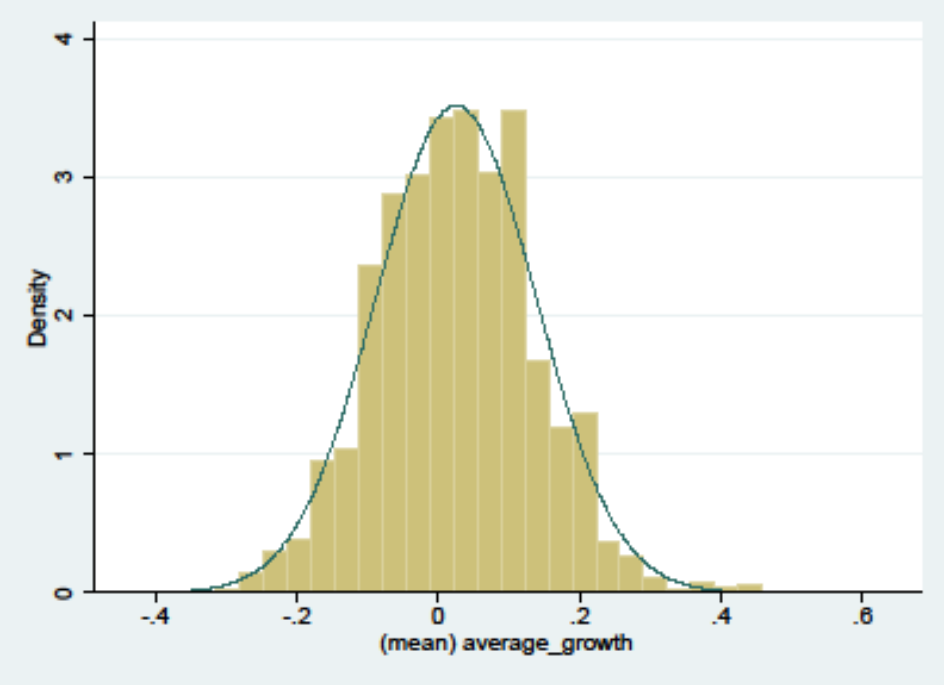

Figure 2: Density of observations across assignment variable.

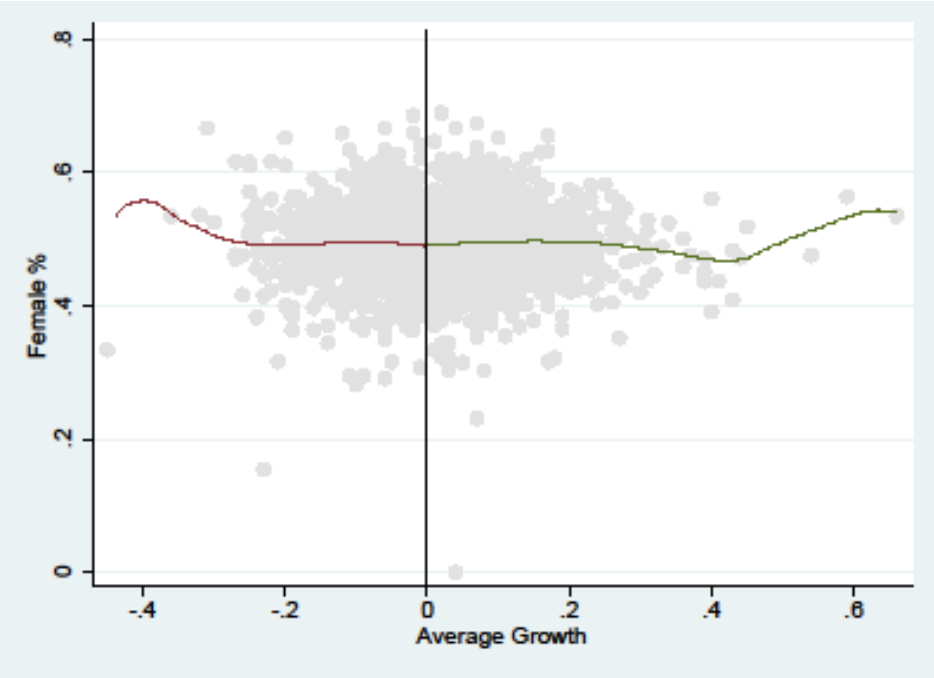

Figure 3: RD of female percent.

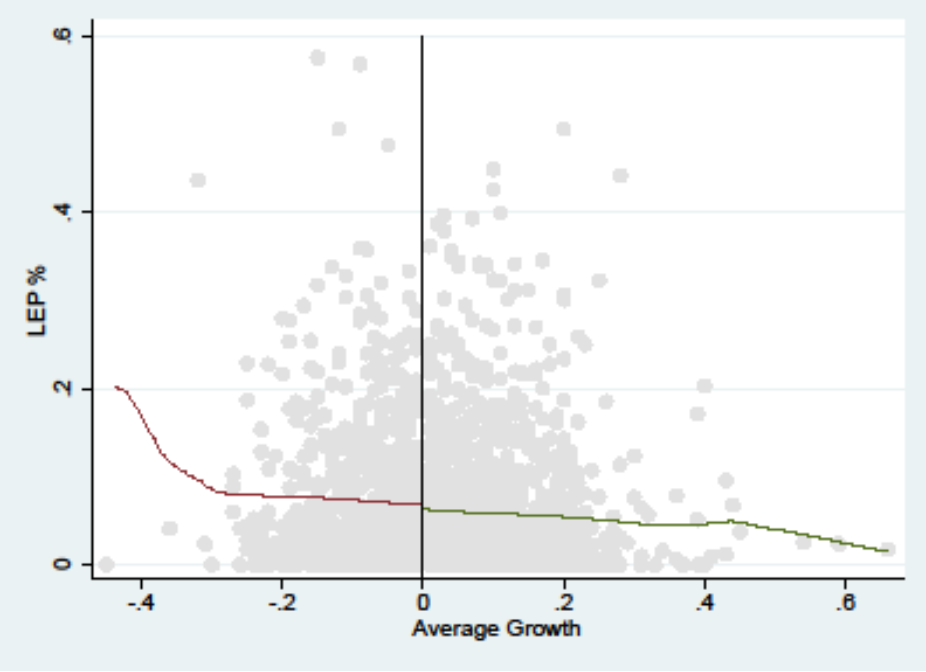

Figure 4: RD of LEP percent.

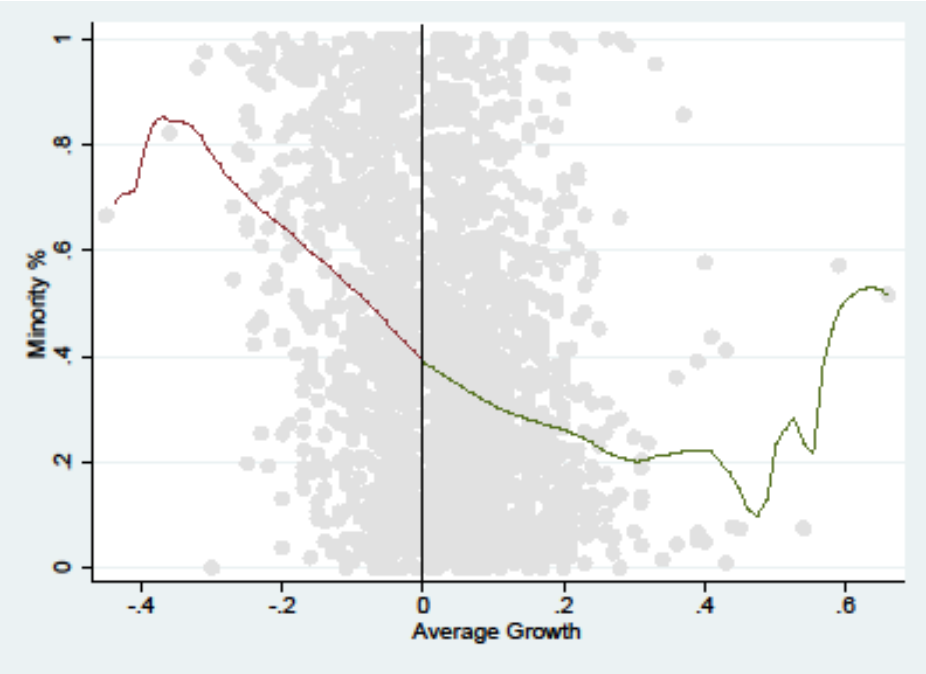

Figure 5: RD of minority percent.

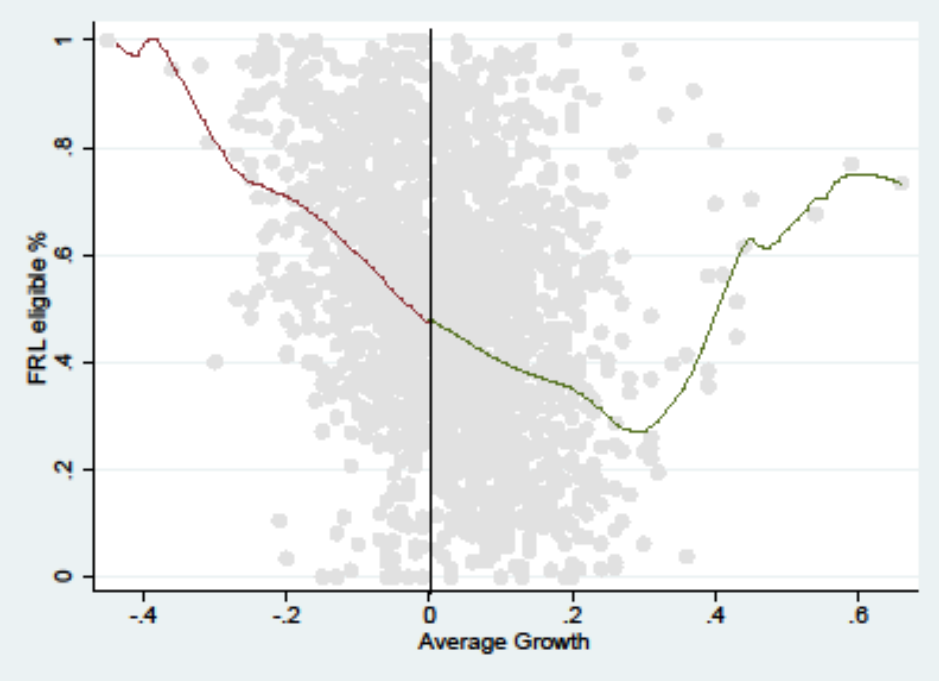

Figure 6: RD of LEP percent.

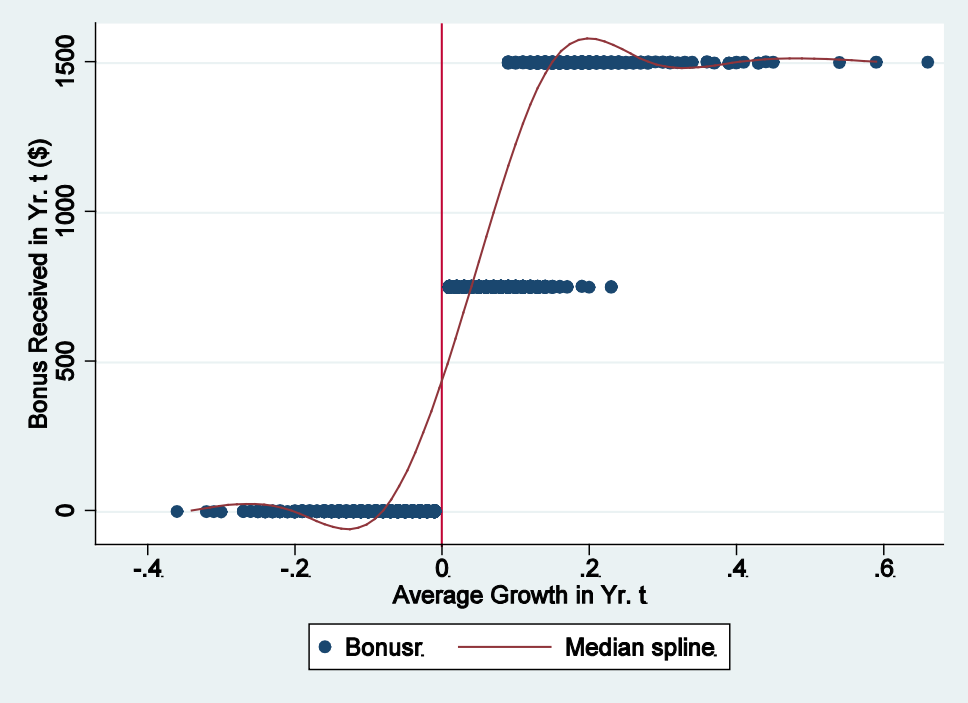

Figure 7: Existence of discontinuity in probability of bonus receipt at policy change.

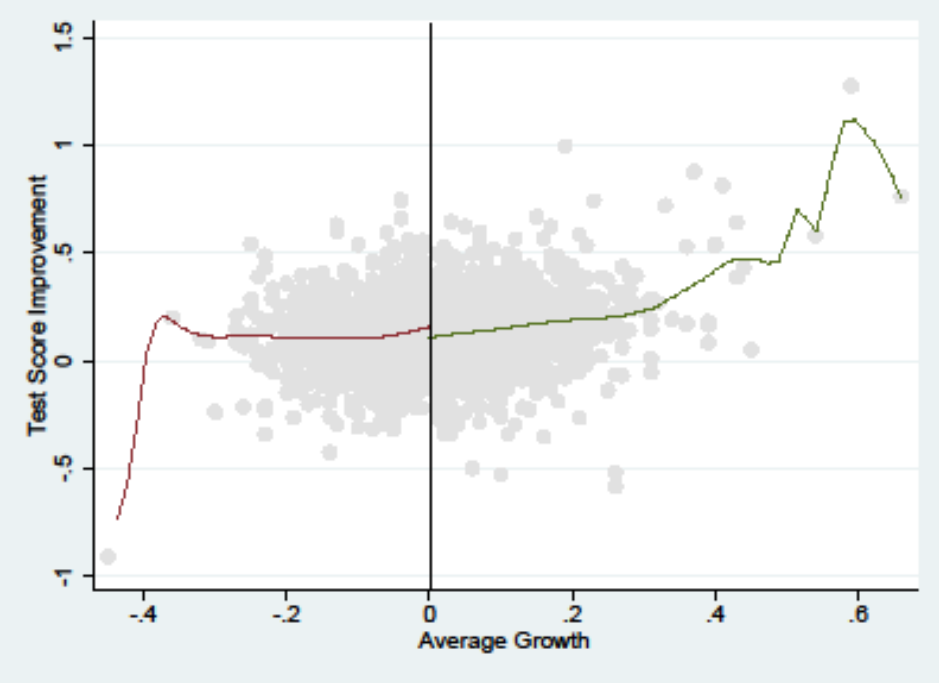

Figure 8: Simple RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t in schools with mid-level experience principals.

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: 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.