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

Tom Ahn[*]
Naval Postgraduate School

Jacob L. Vigdor
University of Washington

August 2021

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 empirical studies, many based on the regression discontinuity design, contradict this basic prediction. We propose two rationalizations of this observed behavior with testable implications. One is based on information acquisition costs and the other focuses on learning and imperfect information. Using education data from North Carolina and exploiting a pay-for-performance system that sends a discrete, extraneous signal of school performance, we find support for the model of rational learning. These results imply that rational responses to policy interventions may take time to emerge, and evaluations based on short-term data may understate treatment effects.

[*] Ahn: sahn1@nps.edu; Vigdor: jvigdor@uw.edu. 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

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

A growing number of empirical analyses identify scenarios where this basic principle fails in practice. Numerous studies document economic agents reacting significantly to discrete categorizations based on an underlying continuous variable even when the variable is in their information set.[1] School performance is also commonly measured using continuous variables such as proficiency rates, but summarized with letter grades or binary designations.[2] In many cases, the failure of some agents to react rationally to irrelevant information creates profit opportunities for others. 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.

The phenomenon is particularly germane in studies employing regression discontinuity (RD) methodology, applied in scenarios where subjects are assigned to discrete categories from an underlying continuous measure. These studies often document significant reactions to discrete categorizations when the underlying difference in a continuous measure is trivial, such

[1] Examples include diners' response to sanitation grades, homeowners' electricity use based on "grades" from utility companies, and car buyers paying higher prices for small differences in odometer readings. See Jin and Leslie (2003), Anderson and Magruder (2011), Allcott (2011), Lacetera, Pope, and Sydnor (2012), Berger and Pope (2011), Card, Mas, and Rothstein (2008).

[2] Evidence from Figlio and Lucas (2004), Martinez (2010), Figlio and Kenny (2009), Ahn (2014), Papay, Murnane, and Willett (2011) shows teachers, parents, and students all respond to discrete grades in the education context.

as when a car's odometer rolls from 9,999 to 10,000 miles. The mounting empirical evidence motivates a question: is there a simple model of behavior that can explain significant responses to signals that should be irrelevant?

There are many important theoretical models of seemingly irrational behavior (Mullainathan 2002; Sarver 2008; see Gigerenzer and Goldstein 1996 for a review of the psychology literature). Informed by this literature, this paper presents two candidate explanations for significant responses to irrelevant information in a principal-agent style production framework with output incentives. It then uses data on school responses to an accountability program in North Carolina to evaluate their empirical implications.

The first model focuses on information costs, presuming that agents are "sparse maximizers" in the sense of Gabaix (2014), using the signal as a proxy for the underlying continuous variable. Intuitively, agents pay no attention to the underlying continuous variable until a discrete indicator informs them that something is wrong, much as a driver might only refuel their vehicle when the low fuel indicator illuminates.

The second posits that agents are Bayesian updaters who act rationally when fully informed. With imperfect information, however, incomplete understanding of the production process may lead them to react based on the signal (Sims 2003). Here, agents pay attention to the continuous measure but are not quite sure whether actions they take will have a beneficial or detrimental impact. The discrete signal, in certain circumstances, motivates them to take a risky action. These two models yield divergent empirical predictions. Most importantly, incomplete information problems should dissipate over time as agents learn the relationship between their actions and consequences. "Sparse maximizers," by contrast, should exhibit persistent behavior.

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 would expect rational, maximizing school personnel to behave similarly regardless of whether they barely missed or barely made the standard for receiving a bonus.

In contrast to this prediction, we show that schools below the threshold exhibited significant improvements in performance in following years relative to schools just above the threshold.[3] The discontinuity is greatest among schools with relatively inexperienced principals and schools with a 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 contributes to the large literature on educational accountability. Accountability systems are designed to improve student performance by introducing information, incentives, or market pressures that may be absent in public education systems.

Robust evidence exists from both U.S. and international implementations of merit pay.[4] Studies demonstrate small but statistically significant gains in math standardized test scores in elementary schools when merit pay systems induce teachers to higher effort. School-level incentives yield smaller effects compared to individual bonuses. Unsurprisingly, larger rewards

[3] One other study has confirmed these basic results. See Jinnai 2016.
[4] In the U.S., at least 40 states or localities have or had some type of incentive program. Internationally, well-known programs existed in China, England, India, Israel, Kenya, and Portugal.

are associated with larger effects. Studies find effects as large as a 4 percent of a standard deviation increase in math scores when the incentive is equal to 10 percent of per capita income.[5]

Numerous empirical studies assessing the impact of accountability policies take advantage of sharp thresholds for qualifications in many systems, where the treatment is unanticipated.[6] The implicit or explicit models of behavior underlying empirical analyses are usually one-shot games with fully informed agents, allowing little scope for effects of incentive programs to evolve as agents learn more about them.[7] While some research points to principals and teachers learning through trial-and-error and altering behavior, in many cases these are considered as threats to the validity of estimated treatment effects and controlled for or minimized using econometric techniques.[8]

This study adds to the evidence that accountability systems can be useful in inducing performance gains while emphasizing that agents may exhibit dynamic learning and respond rationally using the limited (and sometimes erroneous) information at hand.

The evidence presented carries implications for the design of incentive systems in education and other domains. Principal-agent models usually presume that agents fully understand the incentives and the process by which their own efforts translate into outcomes. Our results suggest that agents may require several iterations to attain this mastery. As such, a one-size-fits-all incentive will have different effects on agents with varying amounts of

[5] For the U.S., see, Koretz and Barron (1998) for KY, Fryer et al. (2012) for Chicago, Ahn (2013, 2017) for NC, Ladd (1999) and Imberman and Lovenheim (2015) for TX, Dee and Keys (2004) for TN, and Winters et al. (2008) for AR. Internationally, see Muralidharan and Sundararaman (2009), Lavy (2002, 2009), Glewwe, Ilias, and Kremer (2010), and Loyalka et al. (2019). See Neal (2011) and Pham et al. (2020) for a comprehensive review.
[6] See Hanushek and Raymond (2005) as well as Chakrbarti (2014), West and Peterson (2006), Chiang (2009), Rockoff and Turner (2010), and Craig, Imberman, and Perdue (2015) for RD implementations.
[7] Hemelt (2011), Mccartney (2012), Ahn (2016), and Ahn and Vigdor (2020) are some exceptions.
[8] Evidence on efficacy of principals is sparse. Some studies find experience and tenure at the school and managerial practices of principals matter. See Grissom and Bartanen (2019), Berteille, Kalogrides, and Loeb (2012), Miller (2013), Bloom et al. (2015), Clark Martorell, and Rockoff (2009), and Ahn and Vigdor (2020).

experience, implying that short-term evaluations of incentive programs may mis-state the long-run implications. Our results also suggest that efforts to "tweak" incentive programs to better align incentives need to be weighed against the costs of disrupting the learning process.

2. Conceptual Framework

As noted above, there have been numerous empirical exercises refuting traditional rational models in making decisions based on information when both a continuous measure and a more aggregated grade based on this continuous measure (and thus extraneous information) is available. Nonetheless, to frame our discussion of alternative models we begin by outlining a basic rational framework based on a principal-agent model. In this scenario, agents who find themselves on either side of a discontinuous incentive threshold face near-equivalent incentives to adjust their behavior.

2.1 Traditional 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. A set of parameters governs the translation of these input choices into output, but the final value incorporates an idiosyncratic shock ϵ_{it}. In the most basic formulation, production function parameters are fixed and known; the shock may exhibit serial correlation. 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.[9]

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 ϵ_{it}. Under a variety of assumptions regarding the evolution of ϵ_{it}, we should expect no discontinuous change in the optimal effort choice as a function of incentive receipt. Were ϵ_{it} entirely uncorrelated across years, rational employees would behave exactly the same in the subsequent period. If $E(\epsilon_{it})$ evolves as a random walk, agents on either side of an arbitrary threshold face nearly-equal incentives to adjust in response to a realization. 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

[9] 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.

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.[10] 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 would introduce costs to the agent of observing the realization of ϵ_{it} and thereby obtaining information on the distribution of future realizations. An alternate extension would be to introduce uncertainty and time-variation in the parameters translating input choices into output, with the agent required to incur some cost in order to update information on parameter values. 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 signal. Plausibly, if there is significant serial correlation in the value

[10] This basic scenario can be straightforwardly translated to simple models of consumer choice. When selecting a product, such as a used car, the quality of the product may vary monotonically as a function of a continuous measure such as mileage, 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 cars from the market given consumer discounting.

of ϵ_{it} or the parameters linking input choices to output 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 their vector of input choices.[11] 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 improvement in next-period performance 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.

Importantly, as this is a scenario where agents face real costs to obtain information, the model would not predict that more experienced agents exhibit any kind of differential response pattern. Conditional on paying the cost, all agents have full information. As such, while school-level resources or distribution of student or teacher ability may impact test scores, the school principal's knowledge gained through prior education or experience should not impact school's the ability to generate test score gains. One might expect that agents might gain better intuition for what is happening with state variables over time, which could be modeled as gradual reductions in information costs. Learning is the focus of our second candidate explanation.

2.3 Introducing incomplete information regarding the production process

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,

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

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 result 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 input choices translate into output (parameters) and uncertain of the role of exogenous factors in production (variance and other properties of ϵ_{it}). This arguably characterizes the status of new entrants to a profession such as teaching (or managing teachers).

After the first production period, agents receive information on a continuous measure of their output. This single data point is insufficient to identify the unknown parameters of the education production function and variance of the error term. They also observe a binary indicator of whether they met a production threshold and receive the cash bonus if they crossed the threshold. Agents receiving different signals, particularly inexperienced agents whose prior beliefs lean toward stationary and deterministic rather than stochastic production functions, face differing incentives to alter their effort allocation in the next period. Such agents would infer that there are minimal incentives to alter their input choices if they barely qualify for the bonus, given the probability that altering input would reduce rather than increase output. On the other hand, the impetus to alter effort allocations is strongest among those who receive the initial signal that their output was insufficient, as they are inclined to believe that they will continue to fail to cross the threshold in subsequent periods without adjusting their inputs. As agents gains a better appreciation for the role of stochastic factors in the production process, these discontinuous responses should dissipate.

In subsequent periods, agents who maintain the same input choices gain information on the variance and other properties (e.g., serial correlation) of the error term, but cannot update their prior beliefs on parameter values. Agents who alter their input choices have some capacity to update their prior beliefs on parameter values, limited by continued draws from the error distribution. For any equilibrium where the first period decision rule is to alter input choices only following the receipt of a negative discrete signal, decisions at intermediate experience levels may depend on the agent's history. Agents receiving their first negative signal after a string of positive signals have relatively accurate beliefs regarding the properties of the error term but limited information on how adjustments to their inputs might raise output. We might expect agents on either side of the signal threshold to be similarly motivated to make adjustments to their input choices, but similarly limited in their efficacy.

Agents who receive multiple negative signals will gain information on the relationship between input choices and output, while maintaining fairly uninformed beliefs on the role of error. If their prior beliefs ascribe a negligible role to stochastic determinants of output, negative signals in the present period may continue to motivate them to make further adjustments to input choices even when output falls arbitrarily close to the threshold. In certain scenarios, then, we expect agents with a history of past negative signals to respond more strongly and more effectively to a negative signal in the current period.

In the long run, agents converge to full information regarding the nature of the production process, at which point the model simplifies to the rational fully-informed agent model above. This presents a contrast with the information cost story as presented above. Discontinuous responses are expected to be concentrated among relatively inexperienced agents.

The learning model implies that the impact of an incentivization policy may take time to emerge. Policy evaluations based on short-term implementation may not reveal the long-run impact obtained when agents have more accurate beliefs regarding how their input choices translate into output.

Additionally, if the goal of an incentivization policy is to increase output, the learning model illustrates a clear trade-off between "tweaking" how the model operates and allowing agents to learn how to respond optimally to it. A "tweak" in this circumstance might involve an action as simple as changing the threshold for receiving a positive signal, or a more complicated reform such as altering how output is measured. In the context of education policy, reforms could include basing incentives on a different standardized test, a different school- or classroom-level summary statistic, or moving away from incentives based on standardized tests entirely.

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 50 and 90 percent 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 percent.

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 percent of schools qualify for some bonus payment even though they have failed to make AYP, and about 30 percent 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 differences in student performance on either side of the bonus discontinuity. 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, limited English proficiency, and free/reduced price lunch participation. Table 2 presents summary statistics for our sample, which consists of students enrolled in schools serving grades 3-5 in the 2005/06 and 2006/07 school years.[12] North Carolina is a racially and socio-economically 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 III indicating "sufficient

[12] 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 across classes and teachers, the teacher utility maximization problem is complicated.

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 principal. We use 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 in the 2005-06 academic year. Therefore, years of experience for principals suffers from right censoring. However, principals with 13 or more years of experience comprise less than 10 percent of the sample.

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, 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.[13]

4.2 Methodology

Our basic goal is to examine the impact of bonus receipt on student performance in the next academic year, using RD analysis. RD 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. 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

[13] 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 individual-level data. In addition, while the state 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.

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) (denoted in tables as IK), as well as the estimate for the robust bandwidth defined by Calonico, Cattaneo, and Titiunik (2014) (denoted in tables as CCT).

Our data contains student-level records, yet assignment to the treatment is at the school level. To estimate the impact of the treatment correctly, we collapse the 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 and pre-determined student socio-demographic variables.

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. Figure 2 shows the distribution of average growth performance measure across all school-year observations, with a normal density overlaid. If schools manipulated their growth performance to affect the bonus outcome, such manipulation would be expected to create a discontinuity in the density of school-year observations at the bonus threshold (McCrary 2008). Table 3 shows that using the entire sample, as well as all subsamples used to explore the deeper implications of the conceptual frameworks, the McCrary test statistic is uniformly insignificant.[14] Intuitively, as the composite growth index

[14] Standard McCrary test density figures for relevant subsamples are presented in the appendix Figures A1 to A4.

is a function of the performance of dozens to hundreds of students, calculated using a complex formula, and revealed *ex post*, it would seem difficult to engage in *ex ante* manipulation.

We next check for balance in covariates on both sides of the bonus threshold. Table 4 looks for discontinuities at the threshold for pre-determined student socio-demographic variables: percent of school population that is female, minority, free/reduced-price lunch, limited English proficient, and enrollment size of the school. Due to the large number of sub-sample analyses, we present only the optimal IK bandwidth result.[15] Alternative bandwidth results are presented in the appendix (Table A1-A3). Most of the RD estimates are indistinguishable from zero, which implies that the observed discontinuity in test scores is less likely to arise coincidentally from sharp changes in demographic characteristics. The one exception is that there is a discernable pattern in student enrollment size, especially for the sub-sample of schools that have experienced at least one bonus failure over the last five years. In these, conditional on the principal being inexperienced, schools that fail to get the bonus this year is substantively larger, compared to schools that just qualify. If these schools are led by a principal with 5 to 10 years of experience, the opposite holds, with schools that just fail being smaller than schools that just receive the bonus. Although odd, these results are not consistently replicated in other bandwidth estimates.

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 3 shows teachers' bonus receipt as a function of the composite growth index. There is a sharp discontinuity in probability of bonus receipt (from zero to one) at the assignment variable value of zero. Teachers

[15] A subset of RD figures showing the lack of discontinuity for these placebo runs is presented in the appendix (Figures A5-A8).

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 (LATE). 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.

5. Results

We next present our findings from the RD analysis. Throughout this section, we present results that are in line with, or contradict, each of our three conceptual frameworks, in order. Therefore, in Section 5.1, we explore regression results that evaluate whether principals operate with complete information. Section 5.2 examines whether the RD results corroborate information costs, and Section 5.3 investigates learning. Section 5.4 presents evidence of effect heterogeneity that suggest school leaders may harbor confusion about the summary statistic incentivized in the North Carolina bonus program.

5.1 Documenting the basic effect

Figure 4 presents a graphical representation of our most basic RD estimates, and Table 5 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 0.0564 with higher point estimates in models with narrower

[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. See the appendix for further discussion.

bandwidths (across the IK estimates).[17] The estimated discontinuity is inconsistent with the simple, traditional complete information model.

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 2020). In addition, Figure 4 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. This localization suggests that school personnel pay attention to more than just the discrete signal, as conditional on receiving a negative signal the response appears to vary as a function of distance to the threshold.[18] This could be consistent with either an information cost or learning framework.

For reading scores, the pattern of RD estimates across bandwidths is similar to math, with larger estimates at narrower bandwidths. However, most are statistically insignificant.[19] Point estimates are reported in Table 5. From this point, our discussion will center on math results as reading results are mostly statistically insignificant in all specifications.

The specification for Table 5 is the "standard" for almost every RD study of merit pay (or any accountability policy). Therefore, while finding a substantive effect for math in the context of the North Carolina program is interesting, it is certainly not novel. However, our point is that

[17] Figures with wider and narrower bandwidths are presented in the appendix (Figures A9-A11).
[18] A possible reason for the gains may be teachers moving in response to the bonus outcome. If *less* effective teachers in failing schools are more likely to transfer, this may generate the discontinuity. However, the modest bonus amount is most likely not enough to induce a large scale exodus. Similarly, while poorly performing students transferring might produce the discontinuity, student transfer rates in NC are low. See Ahn and Vigdor (2020).
[19] This is in line with most of the literature that finds teachers and schools less able to impact reading scores compared to math scores. See Jacob (2005), Reback (2008), and Rouse et al. (2013), among many others. A few studies have found the opposite. See Muralidharan and Sundararaman (2011).

the "standard" results are in tension with a story of principals and teachers operating with complete information.

We confirm the robustness of our initial findings by conducting a set of analyses to check for possible confounding factors. Detailed descriptions are in the appendix. To summarize, we 1) examine whether the estimated math treatment effect is robust to bandwidth choice (Figures A9 – A12), 2) run the analysis with artificial thresholds that do not relate to the policy in reality (Figure A13), and 3) estimate a set of parametric RD specifications that mirror our non-parametric model (Table A4, Specification (1)). We find that 1) the treatment effect is robust to bandwidth choice, 2) estimates away from the true threshold yield imprecise estimates, and 3) the parametric RD yield estimates for the discontinuity at about 0.048, which is bracketed by the IK estimate with the original bandwidth and the CCT estimate.

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.

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. This is, so far, in line with the information cost framework. As discussed further in Section 2.2, this explanation suggests that the signal value of bonus receipt (or non-receipt) is strongest when a school experiences a change in status after a period of

relative stability. For this reason, we now study how reactions to bonus receipt vary across schools with differing histories, and therefore differing expectations, regarding the bonus.

Table 6 shows RD estimates for subsets of schools divided by to their past performance in the bonus system. Schools are divided into those that have continuously qualified for the bonus, and those that have at least one failure in the last five years. If 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 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 6 contradict predictions of a stronger response by schools upon their first failure to qualify for the bonus. In fact, schools that have never failed do not exhibit any test score response 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 0.05 of a standard deviation after the next failure. See the appendix for confirmation of these results using parametric RD (Table A4, Specification (2)). As we will show in the next section, another prediction from the information cost framework fails to hold as well: principal experience impacts a school's ability to respond effectively to incentives. Clearly, the results 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

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

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 in Section 2.3. In this section, we present further tests based on principal experience levels.

Table 7 reproduces the basic results from Table 5 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.[21] Estimates indicate that school with principals of mid-level experience just below the bonus eligibility threshold exhibit higher test score gains relative to barely-eligible schools. The effect ranges from 0.0442 to 0.0999. These are large improvements, comparable to the impact of *replacing* an ineffective principal from NCLB sanctions (Ahn and Vigdor 2020). Similar effects are not observed for principals with low or high levels of experience. In fact, not only are the discontinuities statistically insignificant, 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 imperfect information and learning.

[21] 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.

Even with imperfect information and learning, we find discontinuities in Table 7 for a subset of principals because incentive effects are only observed after failing to get the bonus.

The statistically *insignificant* effects for principals with more than 10 years of experience is consistent with the prediction that highly experienced principals will have had periods of sustained success, allowing them to learn about the stochastic shocks (luck) in the bonus outcome. These principals converge to full information regarding the nature of education production, understanding that the impetus to adjust input choices in period t is not a discontinuous function of outcomes in period t-1 and possessing the expertise necessary to make these adjustments effectively.

On the other hand, the statistically significant discontinuity observed for principals with 5 to 10 year of experience aligns with the behavior of principals who have learned about the links between effort and output due to experimentation, but as yet are relatively uninformed about the role of exogenous factors.

The lack of an observed discontinuity for principals with less than 5 years of experience could be attributed to the lack of understanding about how to improve. Without the experience of learning about the production process, these principals may indeed be trying to improve, but with mixed success, leading to insignificant estimates.

Splitting schools by accountability history and principal experience further supports our learning framework. Table 8 presents these results. While schools with spotless records do not have statistically significant discontinuities in response to their first failure regardless of principal experience, 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.

At intermediate experience levels, the learning model predicts significantly different patterns of discontinuous response to the discrete signal. Those 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. Principals 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. 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 0.0438 to 0.1067 of a standard deviation.[22]

This result is also consistent with the principal's prior on the variance of shock (luck) evolving from a small value (where production is assumed to be mostly deterministic) to a larger value (where production is at least partially dependent on luck) with experience: principals initially learn on the job to systematically over-emphasize the role of effort relative to luck in scenarios where they have been repeatedly exhorted to exert more effort or find new ways to increase test scores. Upon finding sustained success, the experienced principal learns not to respond as he or she learns about the stochastic component of the education production function. Parametric RD results in the appendix replicate these findings (Table A4, Specifications (3) – (5)).

[22] 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, highly experienced principals may be tasked with resuscitating underachieving schools under the NCLB regime.

5.4 Effect heterogeneity and the impact of dueling accountability systems

As described above, North Carolina's accountability program paid bonuses on the basis of year-over-year test score gains. There is no reason to believe that gains are easier to produce among students in close proximity to the proficiency threshold. Indeed, depending on 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 proficiency threshold (Neal and Schanzenbach, 2010). Given the low degree of correlation between bonus receipt and NCLB sanction shown in Table 1, evidence that principals focus on students near the threshold when the state system gives them strong reasons to focus on generating gains would be consistent with a fundamental misunderstanding about the nature of the incentive system.

Table 9 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 level 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 proficiency. 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 bonus program.

These results suggest one specific dimension where there is some scope for learning among school leaders. In a system where there are dueling accountability systems, it may take time for principals to understand that the summary statistic used to assess a school in one system (mean year-over-year test score gain) is not necessarily maximized by a strategy that focuses on a different metric (percent proficient).

6. Conclusion

This paper identifies a scenario in the education context where, relative to a rational maximization model, agents exhibit too much sensitivity to discrete signals based on continuous variables available in their information sets. While we do not claim to have identified the only possible explanation for this phenomenon, we show evidence that this behavior is consistent with a Bayesian learning model, in which imperfectly informed agents focus on the discrete signal before they fully appreciate the continuous measure that underlies it. In our case, principals obtain information about their school's performance and learn to pay attention only to the underlying continuous information.

This paper also sheds light on issues in evaluating and operating incentive systems. Until agents learn, through repetition, how input choices map into output, incentive programs may appear to have no impact. This may explain why evaluations of one-shot educational incentive schemes sometimes find no effects, while schemes implemented over time show important effects. Our findings suggest that it may be inappropriate to evaluate incentive systems based on short-term implementation experiments. Administrators should also resist the temptation to calibrate the system to maximize effort, as this may hamper learning by principals.

Our findings of incentive effects dissipating as the RD bandwidths increase point to the limitation of one (or a few) threshold(s) for incentive qualification. In designing future policy, we may wish to consider a continuous payout that reward gains anywhere along the growth domain or place the threshold at the point with the highest density of schools.[23]

[23] Alternative systems may have their own issues. Continuous payout schemes may have resistant buy-in if it appears too opaque. Setting the threshold at the "fattest" part of the density may mean that it gets set too low (high), leading to the bonus being a foregone (impossible) outcome for many schools.

In addition, beyond models of competition among teachers to induce gains, information and learning may also have a role to play in education production. Ratcheting up pressure may lead to more effort by principals and teachers to produce gains, but only to the extent that they have the capacity to make improvements. Those without experience or insight into the education process may be unable to effectively respond. Related to this insight, incentive thresholds should be based on the ability of the principal. While experienced principals may be held to a higher standard, demands on new school leaders should be tempered and experimentation encouraged.

Finally, sharing of knowledge and best practices across schools may be helpful in speeding up gains. More collaborative incentives, such as rewards based on joint production of a group of experienced and new principals, may hit the sweet spot of inducing both effort exertion and information sharing. Ultimately, the research shows that, as is the case with most well-meaning incentive systems, a one-size-fits-all program that expects all principals to make similar gains, if only they were appropriately motivated to do so, may be both inefficient and unfair.

References

Ahn, T. (2013) "The Missing Link: Estimating the Impact of Incentives on Effort and Effort on Production Using Teacher Accountability Legislation" *Journal of Human Capital*, Vol. 7(3), pp. 230-273.

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. (2016) "A theory of dynamic investment in education in response to accountability pressure." *Economics Letters*, Volume 149, Pages 75-78.

Ahn T. (2017) "Strategic Matching of Teachers and Schools with (and without) Accountability Pressure." *Education Finance and Policy* 12:4, Pages 516-535.

Ahn, T. and J. L. Vigdor (2019) "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina, " *Working Paper*.

Ahn, T. and J. Vigdor (2020) “Opening the Black Box: Behavioral Responses of Teachers and Principals to Pay-for-Performance Incentive Programs.” *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.

Atkinson, A., S. Burgess, B. Croxson, P. Gregg, C. Propper, H. Slater, and D. Wilson (2009) “Evaluating the Impact of Performance-related Pay for Teachers in England,” *Labour Economics* 16:3 Pages 251-261.

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

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.*

Béteille, T., D. Kalogrides, and S. Loeb (2012) “Stepping stones: Principal career paths and school outcomes,” *Social Science Research,* Volume 41, Issue 4, Pages 904-919.

Bloom N., R. Lemos, R. Sadun, and J. Van Reenen (2015) “Does Management Matter in schools?” *The Economic Journal,* Volume 125, Issue 584, Pages 647–674.

Calonico S., M. Cattaneo, and R. Titiunik (2015) “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs” *Econometrica,* Volume 82, No. 6, Pages 2295-2326.

Clark, D., Martorell, P., Rockoff, J. (2009) “School principals and school performance.” Washington, DC. Retrieved from https://files.eric.ed.gov/fulltext/ED509693.pdf

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. “Incentives and Responses under No Child Left Behind: Credible Threats and the Role of Competition.” *Journal of Public Economics,* Volume 110, Pages 124-146

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

Clotfelter, C., H. Ladd, J. Vigdor, and R. Diaz (2004) “Do school accountability systems make it more difficult for low performing schools to attract and retain high-quality teachers?” *Journal*

of Policy Analysis and Management Volume 23, Pages 251-271.

Craig S. G., S. A. Imberman, and A. Perdue (2015) "Do administrators respond to their accountability ratings? The response of school budgets to accountability grades." *Economics of Education Review,* Volume 49, Pages 55-68.

Dee, T. S., and B. J. Keys (2004). "Does merit pay reward good teachers? Evidence from a randomized experiment." *Journal of Policy Analysis and Management*, 23(3), Pages 471–488.

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.

Figlio, D. N. and J. Winicki (2005) "Food for Thought: The Effect of School Accountability Plans on School Nutrition," *Journal of Public Economics* 89. Pages 381-394.

Fryer, R. G., S. D. Levitt, J. List, and S. Sadoff (2012) "Enhancing the efficacy of teacher incentives through loss aversion: A field experiment." NBER Working Paper w18237.

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.

Glewwe, P., N. Ilias, and M. Kremer (2010) "Teacher Incentives." *American Economic Journal: Applied Economics* 2(3): Pages 205-227.

Grissom, J.A., and B. Bartanen (2019) "Principal effectiveness and principal turnover", *Education Finance and Policy*

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.

Hemelt S. W. (2011) “Performance effects of failure to make Adequate Yearly Progress (AYP): Evidence from a regression discontinuity framework.” *Economics of Education Review*, Volume 30, Issue 4, Pages 702-723.

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.

Imberman, S. A., and M. F. Lovenheim (2015) “Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system.” *Review of Economics and Statistics*, 97(2), Pages 364–386.

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. (2016). “The effects of a teacher performance-pay program on student achievement: A regression discontinuity approach.” *Economics Bulletin,* 36(2), Pages 993–999.

Koretz, D. M., and S. I. Barron. (1998) “The Validity of Gains on the Kentucky Instructional Results Information System.” (KIRIS). Santa Monica: RAND.

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.

Ladd, H. F. (1999) “The Dallas School Accountability and Incentive Program: an Evaluation of its Impacts on Student Outcomes.” *Economics of Education Review* 18(1): Pages 1-16.

Lavy, V. (2002) “Evaluating the Effect of Teachers’ Group Performance Incentives on Pupil Achievement.” *Journal of Political Economy* 110(6): Pages 1286-1317.

Lavy, V. (2009) “Performance Pay and Teachers’ Effort, Productivity and Grading Ethics.” *American Economic Review* 99(5): Pages 1979-2021.

Loyalka, P., S. Sylvia, C. Liu, J. Chu, and Y. Shi (2019) “Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement.” *Journal of Labor Economics* 37:3, Pages 621-662.

Macartney, H. (2016) “The Dynamic Effects of Educational Accountability*.” Journal of Labor Economics*, Volume 34:1, Pages 1-28.

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

Martins, S. Pedro (2009) “Individual Teacher Incentives, Student Achievement and Grade Inflation,” IZA Discussion Paper No. 4051.

Miller, A. (2013) “Principal turnover and student achievement.” *Economics of Education Review,* Volume 36, Pages 60–72.

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.

Neal, D. (2011) “The Design of Performance Pay in Education” in E. A. Hanushek, S. Machin and L. (Eds.) Handbook of the Economics of Education, vol. 4. North-Holland: Amsterdam.

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*

Pham, L. D., Nguyen, T. D., and Springer, M. G. (2020) “Teacher Merit Pay: A Meta-Analysis.” *American Educational Research Journal.*

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.

Rockoff, J., and L. J. Turner (2010) "Short-Run Impacts of Accountability on School Quality." *American Economic Journal: Economic Policy*, 2 (4). Pages 119-47.

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.

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.

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.

Winters, M., J. P. Greene, G. Ritter, and R. Marsh (2008) “The Effect of Performance-Pay in Little Rock, Arkansas on Student Achievement,” National Center on Performance Incentives, Peabody College of Vanderbilt University, Working Paper.

Figures and Tables

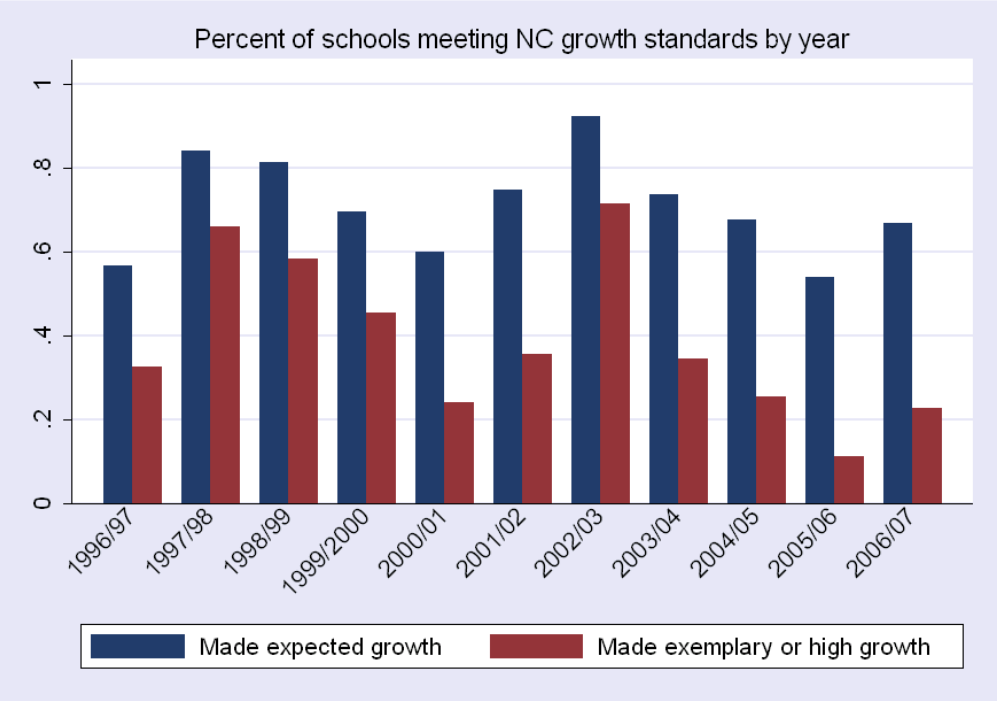

Figure 1: Proportion of schools qualifying for NC bonus. Expected growth rewards $750 per teacher, High growth is $1,500. (From Vigdor 2009)

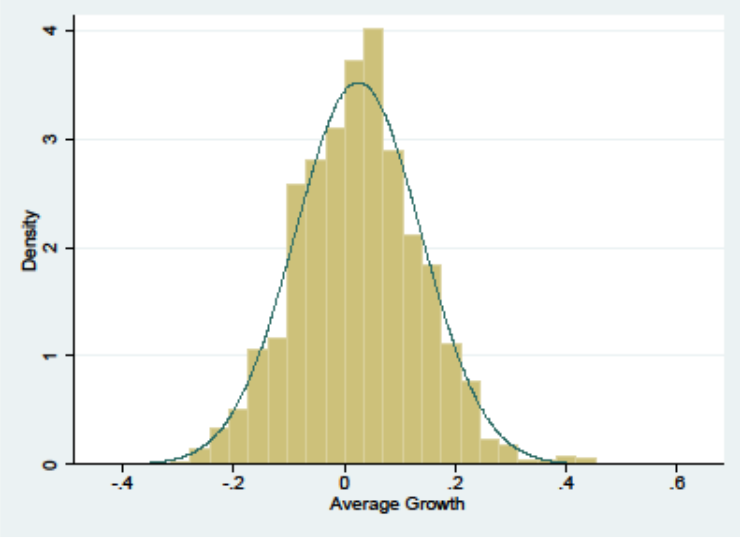

Figure 2: Density of observations across assignment variable of full sample of schools, which shows little evidence of manipulation at threshold cutoff. Normal density overlaid. Threshold cutoff is centered at zero.

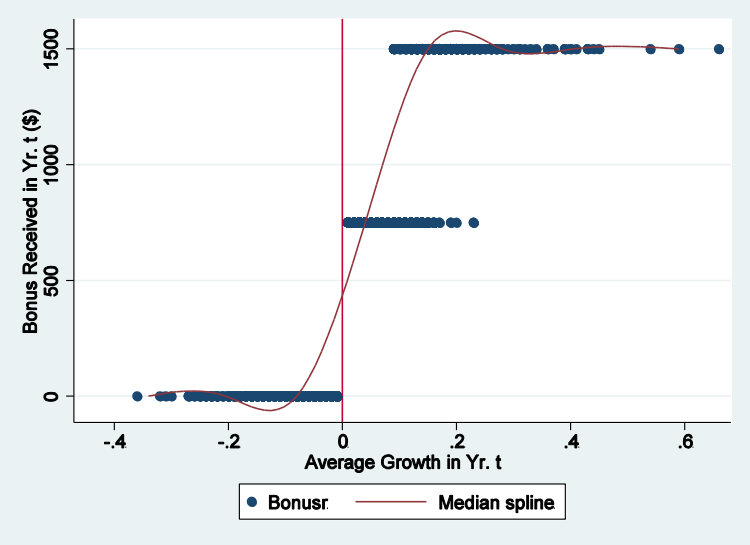

Figure 3: Existence of discontinuity in probability of bonus receipt at policy change. Sharp discontinuity at $750 for "expected" growth, fuzzy discontinuity at $1,500 for "greater than expected" growth.

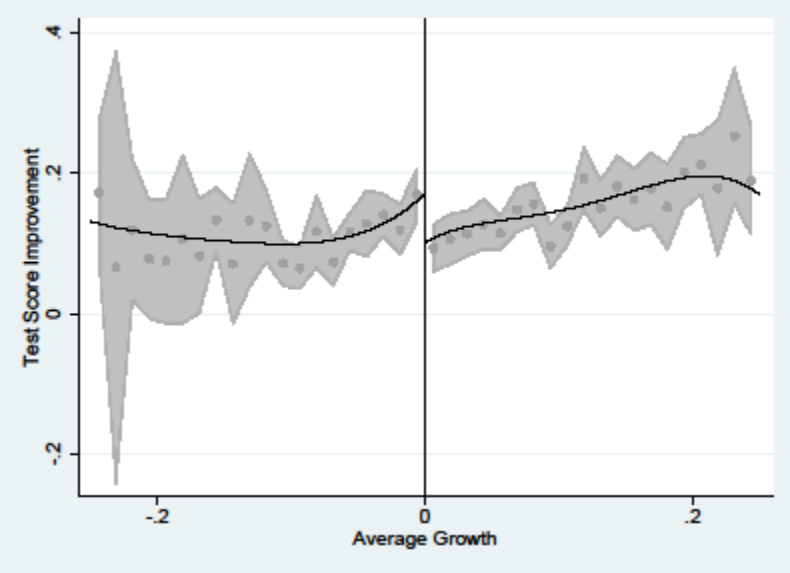

Figure 4: Simple RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t. Line is generated with polynomial of degree 4. Local averages presented with 40 bins. Observations for assignment variable less than -0.25 or greater than 0.25 are dropped for presentation purposes. Shading represents the 95% confidence interval.

Table 1: AYP and ABC Status

		ABC	
		Yes	No
AYP	Yes	956	284
	No	423	635

Note: Yes – qualified under accountability regime. Adequate Yearly Progress for NCLB, expected growth bonus receipt for ABC program. No – failed to qualify under accountability regime.

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: McCrary Test for Continuity of the Assignment Variable at the Threshold

Sample	Table.	Discontinuity	Std. Err.
All schools	5	0.0855	(0.0910)
No Fails	6	0.0851	(0.1717)
1 or more Fail		0.0627	(0.1346)
Experience < 5 yrs	7	0.0530	(0.1717)
5 yrs <= Experience <=10 yrs		-0.0576	(0.1403)
Experience>10 yrs		-0.0025	(0.1969)
No Fails / Exp < 5 yrs	8	0.0785	(0.2767)
No Fails / 5 yrs <= Exp <=10 yrs		0.1987	(0.3935)
No Fails / Exp>10 yrs		-0.2446	(0.4822)
1 or more Fails / Exp < 5 yrs		0.0681	(0.2147)
1 or more Fails / 5 yrs<= Exp<=10 yrs		0.0470	(0.2318)
1 or more Fails / Exp>10 yrs		0.0976	(0.3127)
Achievement Level 1	9	0.0593	(0.1151)
Achievement Level 2		0.0808	(0.1132)
Achievement Level 3		0.0828	(0.1137)
Achievement Level 4		0.0881	(0.1143)

Note: McCrary statistic calculated for every sub-sample. Relevant subsample analyzed in Table referenced above.

Table 4: Placebo RD with Predetermined Variables for all Sub-samples

	All principals	Exp. > 10	5<=Exp.<=10	Exp.<5
All Schools				
% Female	0.0008	-0.0098	-0.0005	0.0084
	(0.0041)	(0.0085)	(0.0064)	(0.0066)
% Minority	-0.0101	0.0454	-0.0330	-0.0102
	(0.0261)	(0.0500)	(0.0430)	(0.0467)
% Poverty	0.0099	0.0728	-0.0298	0.0288
	(0.0213)	(0.0512)	(0.0303)	(0.0344)
% LEP	-0.0059	0.0042	-0.0087	-0.0041
	(0.0073)	(0.0154)	(0.0115)	(0.0108)
School Size	-10.2669	-52.4741	41.7285*	-42.7783**
	(15.6733)	(42.4510)	(23.4730)	(19.5751)
Qualified for bonus every year				
% Female	-0.0003	-0.0184	-0.0018	0.0165**
	(0.0043)	(0.0127)	(0.0068)	(0.0081)
% Minority	-0.0040	0.0810	-0.0379	-0.0147
	(0.0379)	(0.0675)	(0.0593)	(0.0606)
% Poverty	0.0026	0.1184*	-0.0782*	0.0366
	(0.0314)	(0.0698)	(0.0466)	(0.0460)
% LEP	-0.0072	0.0016	-0.0214	0.0000
	(0.0109)	(0.0184)	(0.0182)	(0.0161)
School Size	2.0791	-52.4069	58.7390*	-28.9612
	(20.3082)	(41.4863)	(33.9442)	(22.5820)
Failed to qualify for bonus in at least 1 year				
% Female	-0.0023	0.0102	-0.0073	-0.0086
	(0.0092)	(0.0164)	(0.0125)	(0.0150)
% Minority	-0.0072	-0.0166	-0.0174	0.0268
	(0.0398)	(0.0674)	(0.0697)	(0.0737)
% Poverty	0.0218	-0.0142	0.0499	0.0521
	(0.0325)	(0.0746)	(0.0558)	(0.0535)
% LEP	-0.0007	-0.0053	0.0179	0.0011
	(0.0125)	(0.0210)	(0.0139)	(0.0232)
School Size	-16.7534	-10.0840	44.1183**	-81.1473**
	(21.6384)	(44.9934)	(19.4519)	(37.1202)

Note: Standard errors in parentheses. Dependent variables are school level demographic population measures. Results for optimal bandwidth from the Imbens-Kalyanaraman algorithm presented. Alternative bandwidth results, including estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014) are available in the appendix (Table A1-A3). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 5: RD estimates of the impact of failing to receive the bonus

	Reading (n=2,276)	Math (n=2,276)
IK optimal bandwidth	-0.0098	-0.0351**
	(0.0111)	(0.0140)
IK Half optimal bandwidth	-0.0226	-0.0458**
	(0.0155)	(0.0192)
IK Twice optimal bandwidth	-0.0057	-0.0260**
	(0.0096)	(0.0116)
CCT bandwidth (robust std. err.)	-0.0462	-0.0564**
	(0.0295)	(0.0287)

Note: Standard errors in parentheses. Dependent variable is reading and math standardized score growth.Bandwidth determination for the top three rows is by the Imbens-Kalyanaraman algorithm. The bottom row is the estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 6: RD estimates by ABC bonus history of success and failures.

Qualified for bonus every year	Reading (n=817)	Math (n=817)
IK optimal bandwidth	-0.0077	-0.0381
	(0.0194)	(0.0361)
IK Half optimal bandwidth	-0.0246	-0.0384
	(0.0252)	(0.0535)
IK Twice optimal bandwidth	-0.0045	-0.0440
	(0.0179)	(0.0304)
CCT bandwidth (robust std. err.)	-0.0452	-0.0243
	(0.0511)	(0.0768)
Failed to qualify for bonus in at least 1 year	Reading (n=1,431)	Math (n=1,431)
IK optimal bandwidth	-0.0139	-0.0415***
	(0.0137)	(0.0160)
IK Half optimal bandwidth	-0.0324	-0.0485**
	(0.0204)	(0.0221)
IK Twice optimal bandwidth	-0.0118	-0.0292**
	(0.0117)	(0.0132)
CCT bandwidth (robust std. err.)	-0.0523	-0.0699**
	(0.0339)	(0.0354)

Note: Standard errors in parentheses. Dependent variable is reading and math standardized score growth. Bandwidth determination for the top three rows is by the Imbens-Kalyanaraman algorithm. The bottom row is the estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 7: RD estimates by experience level of principal (math scores only)

Years of Experience	Exp. > 10 (n=456)	5<=Exp.<=10 (n=824)	Exp.<5 (n=968)
IK optimal bandwidth	-0.0302	-0.0456**	-0.0209
	(0.0310)	(0.0213)	(0.217)
IK Half optimal bandwidth	-0.0373	-0.0536**	-0.0507
	(0.0417)	(0.0274)	(0.0302)
IK Twice optimal bandwidth	-0.0161	-0.0442**	-0.0157
	(0.0253)	(0.0200)	(0.0175)
CCT bandwidth (robust std. err.)	0.0261	-0.0999**	-0.0682
	(0.0889)	(0.0424)	(0.0421)

Note: Standard errors in parentheses. Dependent variable is math standardized score growth. Bandwidth determination for the top three rows is by the Imbens-Kalyanaraman algorithm. The bottom row is the estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 8: RD estimates by ABC bonus history X experience level (math scores only)

Qualified for bonus every year	Exp. > 10 (n=180)	5<=Exp.<=10 (n=285)	Exp.<5 (n=352)
IK optimal bandwidth	0.0179	-0.0823	-0.0510
	(0.0665)	(0.0544)	(0.0557)
IK Half optimal bandwidth	-0.0095	-0.0746	-0.0618
	(0.0897)	(0.0685)	(0.0883)
IK Twice optimal bandwidth	0.0050	-0.0908*	-0.0549
	(0.0588)	(0.0522)	(0.0453)
CCT bandwidth (robust std. err.)	0.0853	0.0062	-0.1049
	(0.2151)	(0.1270)	(0.0852)
Failed to qualify for bonus at least once	Exp. > 10 (n=276)	5<=Exp.<=10 (n=539)	Exp.<5 (n=616)
IK optimal bandwidth	-0.0420	-0.0597**	-0.0109
	(0.0334)	(0.0273)	(0.0215)
IK Half optimal bandwidth	-0.0550	-0.0660*	-0.0317
	(0.0441)	(0.0376)	(0.0278)
IK Twice optimal bandwidth	-0.0314	-0.0438**	-0.0168
	(0.0295)	(0.0229)	(0.0184)
CCT bandwidth (robust std. err.)	0.0065	-0.1067**	-0.0624
	(0.0807)	(0.0538)	(0.0437)

Note: Standard errors in parentheses. Dependent variable is math standardized score growth. Bandwidth determination for the top three rows is by the Imbens-Kalyanaraman algorithm. The bottom row is the estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table 9: RD estimates by math proficiency level

Level I: insufficient mastery	Math (n=2,078)
IK optimal bandwidth	0.0121
	(0.0348)
IK Half optimal bandwidth	0.0092
	(0.0502)
IK Twice optimal bandwidth	0.0046
	(0.0283)
CCT bandwidth (robust std. err.)	-0.0015
	(0.0573)
Level II: inconsistent mastery	**Math (n=2,212)**
IK optimal bandwidth	-0.0342*
	(0.0182)
IK Half optimal bandwidth	-0.0324
	(0.0218)
IK Twice optimal bandwidth	-0.0336*
	(0.0175)
CCT bandwidth (robust std. err.)	-0.0310
	(0.0319)
Level III: sufficient mastery	**Math (n=2,195)**
IK optimal bandwidth	-0.0399**
	(0.0185)
IK Half optimal bandwidth	-0.0498*
	(0.0263)
IK Twice optimal bandwidth	-0.0317**
	(0.0152)
CCT bandwidth (robust std. err.)	-0.0633**
	(0.0306)
Level IV: superior mastery	**Math (n=2,161)**
IK optimal bandwidth	-0.0261
	(0.0168)
IK Half optimal bandwidth	-0.0255
	(0.0221)
IK Twice optimal bandwidth	-0.0236
	(0.0152)
CCT bandwidth (robust std. err.)	-0.0324
	(0.0370)

Note: Standard errors in parentheses. Dependent variables defined in the text. Bandwidth determination for the top three rows is by the Imbens-Kalyanaraman algorithm. The bottom row is the estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Appendix

Robustness Checks for RD

Figure A12 presents evidence of robustness of the math treatment effect estimate to bandwidth choice, by representing effect estimates and 95 percent 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 A13 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.

Parametric RD Specifications

We run a set of parametric RD estimations. The benefits of this analysis is two-fold: parametric RD allows us to use the entire sample of schools in one regression, increasing the power, and it allows us to replicate the non-linear effects of experience we show in later sections using non-parametric RD. Our base specification (1) is as follows:

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

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). 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. 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. Estimates of the parametric discontinuity is shown in the first set of results in Appendix Table A4 to be about 0.048. This value is bracketed by the IK estimate with the optimal bandwidth and the CCT estimate.

In addition to the base specification (1), we also run alternative models that are used to assess the other information models. Specification (2) includes accountability history of the school, with a dummy variable which equals one for schools that have consistently qualified for the bonus, Z_{it}. Each argument inside the function $f(\cdot)$ has polynomial controls in its own and all possible interaction terms.

$$(2) \qquad \Delta math_{it} = \beta_0 + X_{it}\beta_1 + \beta_2 I_{it} + \beta_3 Z_{it} + \beta_4 I_{it} \cdot Z_{it} + f(\Delta y_{i,t-1}, Z_{it}) + f(I_{it} \cdot \Delta y_{i,t-1}, Z_{it}) + \epsilon_{it}$$

Results from parametric RD Specification (2), which separately estimates treatment effects for schools by their accountability history show that the treatment effect on schools with consistent success in attaining the bonus is negligible from zero, while the treatment effect on schools with at least one failure is marginally statistically significant. The difference in estimated magnitude of the parameters is striking, as the latter estimate is over 8 times larger than the former estimate.

We also run models that include measures of the principal's experience. Specification (3) accounts for experience and learning in a basic way, as follows: [24]

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

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

LEV_{it} is a dummy variable for the experience level for the principal in school i. Specification (4) uses a continuous measure of principal experience and its squared term to confirm that results are not being driven by the cut-off values for experience:

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

The final specification (5) includes accountability history of the school interacted with the experience measure.

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

Parametric regression runs buttress our results from the non-parametric RD analysis. Continuing with Specification (3), which contains dummy variables for low, mid, and high experience level, the parameter of interest here is β_4. Following the results in Table 7, we would expect β_4 for mid-experience principals to be negative.

When the continuous measure of experience is used as in Specification (4), 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 replicate Table 8, we interact experience with school accountability history in Specification (5). Here, 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 three specifications above are presented in third, fourth, and fifith sets of results in Appendix Table A4.[25]

As demonstrated, the relevant parameter estimates in all three specifications are consistent with the non-parametric RD results. In particular, in specification (4), the shape of the experience response shows that the peak of response is between 6 and 7 years of experience.

The inclusion of experience as a *control* variable (in the parametric RD) is standard in many education production function estimations that include teacher or principal characteristics. Our unique set-up, in splitting the school sample by accountability history and principal experience in the non-parametric RD and including the experience and history measures interacted with the indicator for threshold in the parametric RD, allows us to find support for the gradual learning model, providing a consistent explanation for the seemingly irrational behavior of schools in response to a modest and straight-forward merit pay system.

The Impact of the $1,500 Threshold

The North Carolina incentive system is relatively straightforward in its implementation, which lends itself to a simple RD framework. One possible issue is the inclusion of a second, higher threshold which can introduce complications to the estimation of the LATE.

This second threshold requires a higher year-over-year test score gains than the "expected" growth criterion. In addition, other criteria, such as ensuring that gains are sufficiently broadly distributed across the student population must be met. This second threshold potentially creates threats to the validity of our estimate due to two factors:

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

1. The incentive impact of the second threshold may be *additive*, leading to an overestimate of the treatment effect. To the extent that the second threshold, which would double the monetary reward, would incentivize the principal and teachers to higher exertion to attain the bonus, this may lead to higher gains.
2. The incentive impact of the second threshold may be *subtractive*, leading to an underestimate of the treatment effect. Average test score growth at the school could decline if the additional conditions required for qualifying for the higher bonus pull effort and attention away from maximizing average gains.

The potential impact of these factors on the RD estimate depends on the following three factors:

1. The relative size difference between the additive and subtractive factors will determine how much "contamination" is possible.
2. The number of schools induced to alter behavior in response to the second threshold, especially schools close to our relevant threshold will impact the estimate. The number of schools induced to change behavior would depend on the perceived probability of success in attaining the second bonus.
3. Related to 2, bandwidth for RD estimation will change the number of potentially "contaminated" schools. Keeping in mind that the treated group is schools that just failed (schools to the left of the threshold), a wider bandwidth selection will contain more schools that are less likely to be impacted by incentives from the second threshold. It also increases the power of the estimator, since sample size increases. However, increasing the bandwidth also decreases the incentive effects from our relevant threshold. In fact, this tension underlies the selection of optimal bandwidth in all RD estimation.

As such, it is interesting to note that comparing the IK estimates in Table 5 at half, optimal, and twice optimal bandwidth, the size of the estimated treatment effect decreases in magnitude from about -0.046 to -0.026. This may imply that the overall incentive effect from the second threshold is additive. However, it is impossible to truly disentangle the relative sizes of these incentives.

It does appear that the second threshold is a difficult bar to cross, with approximately 15 percent of schools receiving the bonus in our sample. In contrast, over 60 percent manage to qualify for at least the lower bonus. This lends some cautious optimism that most schools would consider the second threshold as out of reach. Then, the impact on the first threshold may be modest.

There is room for some cautious optimism that whatever contamination exists, it would be relatively minor. For example, TN POINT system examined in Springer et al. (2010) offered much more generous benefits (up to $15,000) yet saw little incentive effects. This may have been because only 20 percent of teachers had a realistic chance to qualify for any bonus conditional on past performance, leading to an optimal non-response by most teachers. With the low fraction of schools that ended up qualifying for the second threshold, it is arguable that given the much smaller bonus amounts in NC, the contamination effects would also have been much more modest.

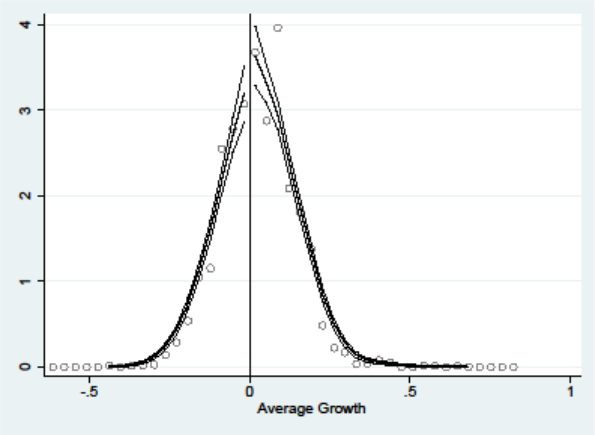

Figure A1: McCrary test figure with entire sample. 95 percent confidence interval brackets density plot.

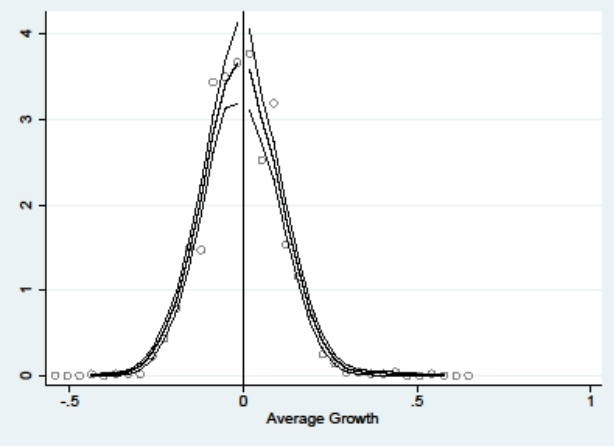

Figure A2: McCrary test figure with schools with at least one failure in the last five years. 95 percent confidence interval brackets density plot.

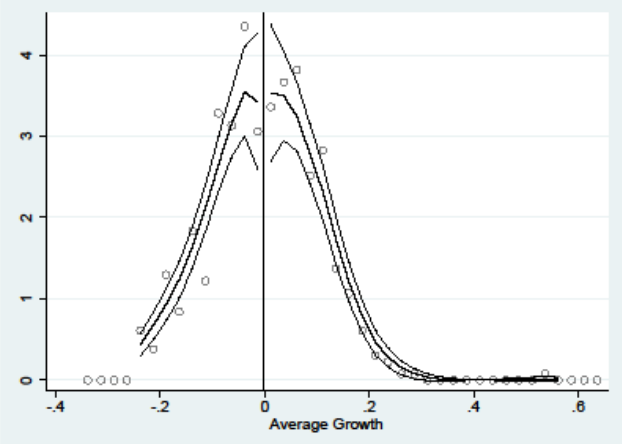

Figure A3: McCrary test figure with schools that have principals with 5 to 10 years of experience. 95 percent confidence interval brackets density plot.

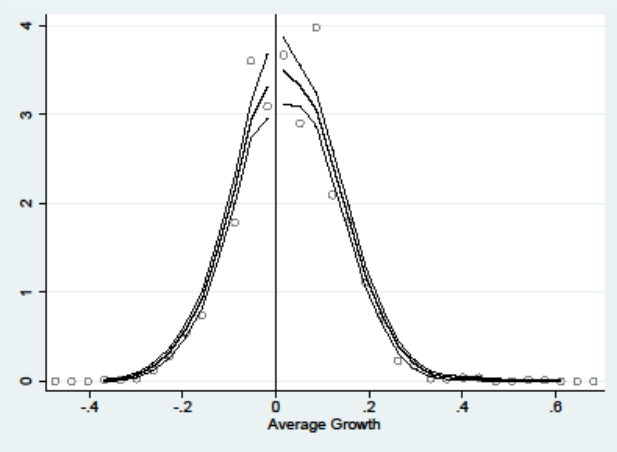

Figure A4: McCrary test figure with schools at math proficiency level III (grade-level competence). 95 percent confidence interval brackets density plot.

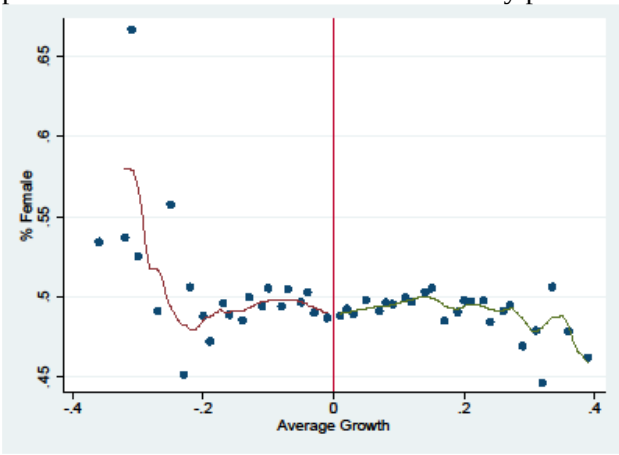

Figure A5: Placebo RD, dependent variable: school population female %. Figure generated with local polynomial of degree zero. Local averages presented with 50 bins. Observations for running variable less than -0.4 or greater than 0.4 (which comprise approximately 3.7 % of observations) dropped for presentation purposes.

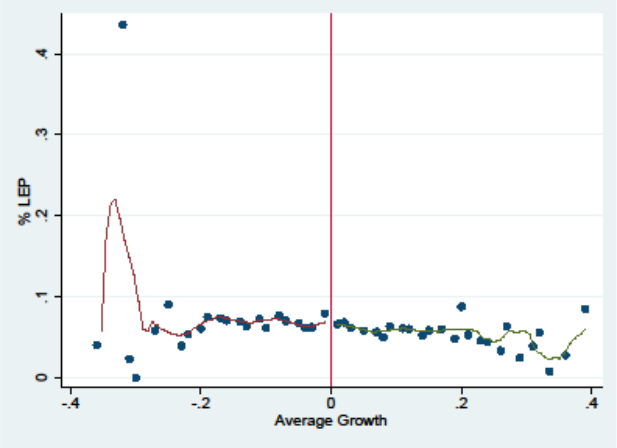

Figure A6: Placebo RD, dependent variable: school population LEP %. Figure generated with local polynomial of degree zero. Local averages presented with 50 bins. Observations for running variable less than -0.4 or greater than 0.4 (which comprise approximately 3.7 % of observations) dropped for presentation purposes.

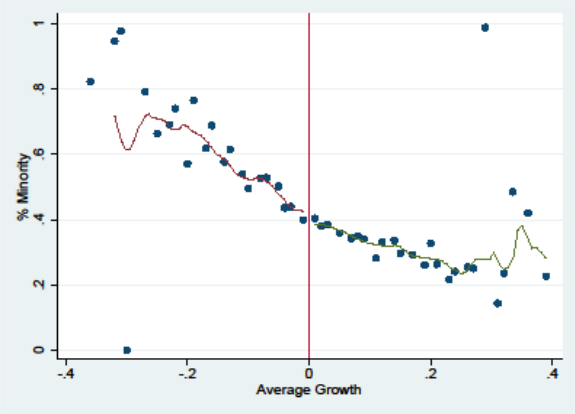

Figure A7: Placebo RD, dependent variable: school population minority %. Figure generated with local polynomial of degree zero. Local averages presented with 50 bins. Observations for running variable less than -0.4 or greater than 0.4 (which comprise approximately 3.7 % of observations) dropped for presentation purposes.

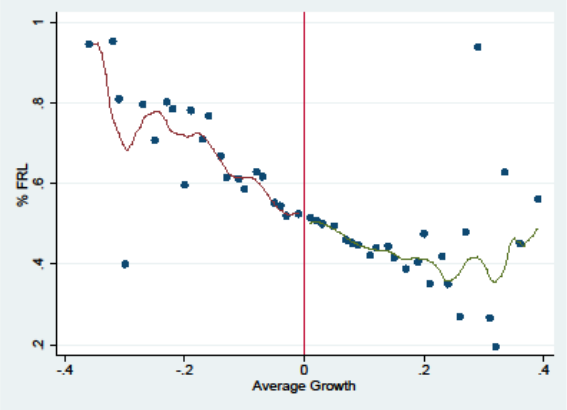

Figure A8: Placebo RD, dependent variable: school population % free, reduced price lunch students. Figure generated with local polynomial of degree zero. Local averages presented with 50 bins. Observations for running variable less than -0.4 or greater than 0.4 (which comprise approximately 3.7 % of observations) dropped for presentation purposes.

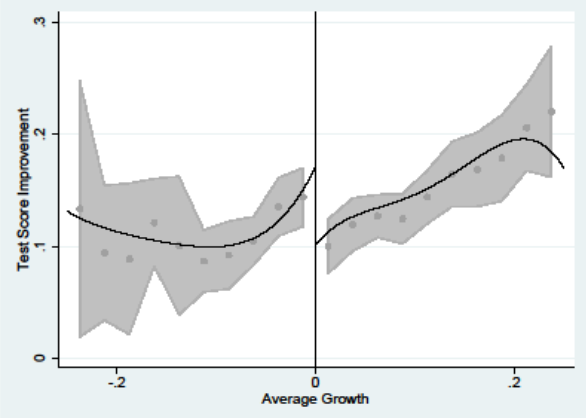

Figure A9: RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t. Figure is generated with polynomial of degree 4. Local averages presented with 20 bins. Observations for running variable less than -0.25 or greater than 0.25 are dropped for presentation purposes. Shading is the 95% confidence interval.

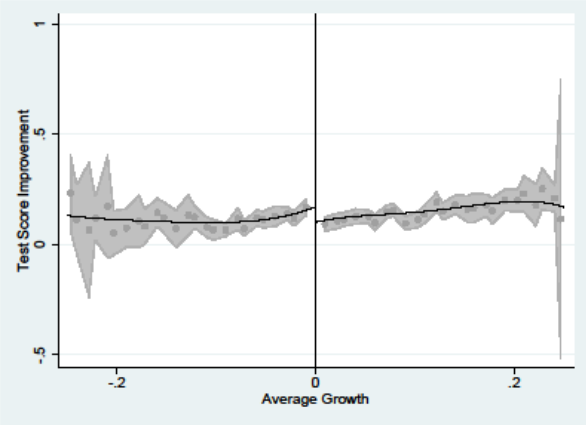

Figure A10: RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t. Figure is generated with polynomial of degree 4. Local averages presented with 80 bins. Observations for running variable less than -0.25 or greater than 0.25 are dropped for presentation purposes. Shading is the 95% confidence interval.

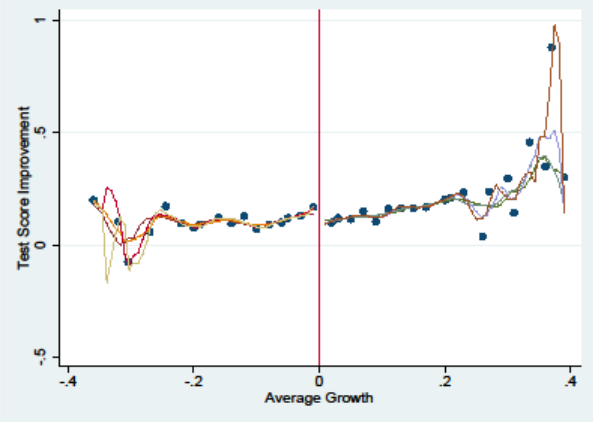

Figure A11: RD illustration of math score improvement in year t+1 conditional on just being below qualification for the bonus in year t. Figure generated with local polynomials of degree zero to degree three. Observations for running variable less than -0.3 or greater than 0.4 dropped for presentation purposes.

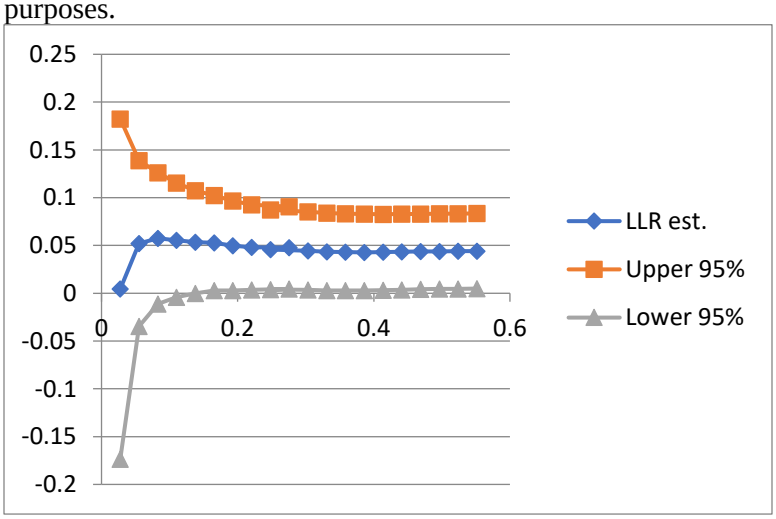

Figure A12: Local Linear Regression Estimates with Varying Bandwidths. Horizontal axis is bandwidth. Vertical axis is estimated treatment effect magnitude. Math score improvement in year t+1 conditional on just being below qualification for the bonus in year t in schools with mid-level experience principals.

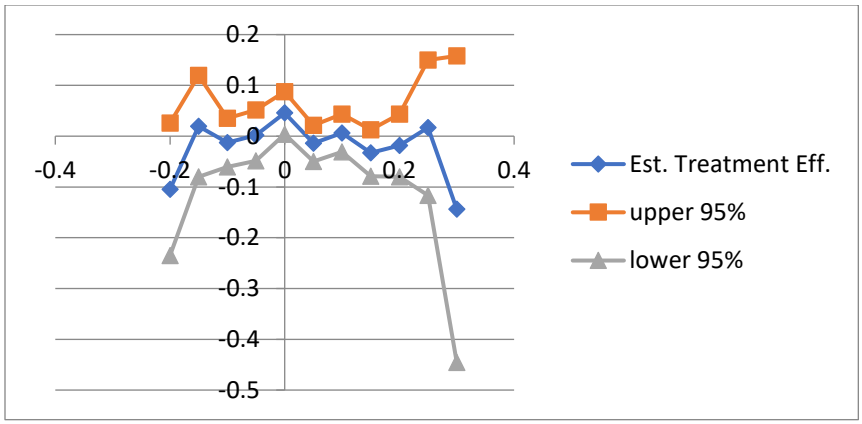

Figure A13: Local Linear Regression Estimates with Artificial Cut-Off Points. Horizontal axis is artificial threshold value. True threshold at 0. Vertical axis is estimated treatment effect magnitude. Math score improvement in year t+1 conditional on just being below qualification for the bonus in year t in schools with mid-level experience principals.

Table A1. Placebo RD with Pre-determined Variables for all schools X principal experience

All schools		All	Exp. > 10	5<=Exp.<=10	Exp. < 5
% Female	IK optimal	0.0008	-0.0098	-0.0005	0.0084
		(0.0041)	(0.0085)	(0.0064)	(0.0066)
	IK 0.5x optimal	0.0064	-0.0084	0.0049	0.0139
		(0.0056)	(0.0112)	(0.0084)	(0.0091)
	IK 2x optimal	0.0002	-0.0083	0.0004	0.0048
		(0.0034)	(0.0074)	(0.0055)	(0.0055)
	CCT	0.0130	-0.0088	0.0111	0.0249*
		(0.0091)	(0.0205)	(0.0119)	(0.0142)
% Minority	IK optimal	-0.0101	0.0454	-0.0330	-0.0102
		(0.0261)	(0.0500)	(0.0430)	(0.0467)
	IK 0.5x optimal	-0.0040	0.0397	-0.0457	-0.0363
		(0.0355)	(0.0645)	(0.0575)	(0.0669)
	IK 2x optimal	-0.0200	0.0177	-0.0442	-0.0216
		(0.0223)	(0.0420)	(0.0383)	(0.0369)
	CCT	-0.0130	0.0369	-0.0090	-0.0803
		(0.0518)	(0.1082)	(0.1011)	(0.0901)
% Poverty	IK optimal	0.0099	0.0728	-0.0298	0.0288
		(0.0213)	(0.0512)	(0.0303)	(0.0344)
	IK 0.5x optimal	0.0080	0.0658	-0.0394	0.0298
		(0.0298)	(0.0705)	(0.0367)	(0.0518)
	IK 2x optimal	0.0010	0.0475	-0.0327	0.0103
		(0.0182)	(0.0418)	(0.0291)	(0.0276)
	CCT	-0.0272	0.0687	-0.0923	-0.0564
		(0.0534)	(0.1091)	(0.0955)	(0.0864)
% LEP	IK optimal	-0.0059	0.0042	-0.0087	-0.0041
		(0.0073)	(0.0154)	(0.0115)	(0.0108)
	IK 0.5x optimal	-0.0067	0.0092	-0.0163	-0.0006
		(0.0099)	(0.0198)	(0.0157)	(0.0139)
	IK 2x optimal	-0.0062	-0.0051	-0.0059	-0.0044
		(0.0064)	(0.0126)	(0.0107)	(0.0103)
	CCT	-0.0296	0.0019	-0.0645	-0.0219
		(0.0216)	(0.0276)	(0.0401)	(0.0349)
School Size	IK optimal	-10.2669	-52.4741	41.7285*	-42.7783**
		(15.6733)	(42.4510)	(23.4730)	(19.5751)
	IK 0.5x optimal	-16.8675	-60.0398	48.4007	-60.5118***
		(21.4177)	(62.1447)	(33.6117)	(22.6186)
	IK 2x optimal	-6.4109	-44.3210	37.0807*	-29.6402*
		(13.0207)	(32.5272)	(21.2442)	(16.6786)
	CCT	-20.3536	-67.6484	44.6793	-43.8053
		(39.3275)	(103.3714)	(65.4574)	(51.9305)

Note: Standard errors in parentheses. Dependent variables are school level population measures. Bandwidth determination for the top 3 rows is by the Imbens-Kalyanaraman algorithm. The bottom row is the estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table A2. Placebo RD with Pre-determined Variables for schools with at least 1 failure X principal experience

Failed to qualify at least 1 yr		All	Exp. > 10	5<=Exp.<=10	Exp. < 5
% Female	IK optimal	-0.0003	-0.0184	-0.0018	0.0165**
		(0.0043)	(0.0127)	(0.0068)	(0.0081)
	IK 0.5x optimal	0.0038	-0.0297*	0.0023	0.0300***
		(0.0055)	(0.0180)	(0.0083)	(0.0114)
	IK 2x optimal	0.0002	-0.0157	-0.0019	0.0097
		(0.0040)	(0.0100)	(0.0064)	(0.0065)
	CCT	0.0144	-0.0540*	0.0184	0.0397**
		(0.0106)	(0.0301)	(0.0188)	(0.0155)
% Minority	IK optimal	-0.0040	0.0810	-0.0379	-0.0147
		(0.0379)	(0.0675)	(0.0593)	(0.0606)
	IK 0.5x optimal	-0.0177	0.0505	-0.0291	-0.0342
		(0.0560)	(0.0837)	(0.0860)	(0.0899)
	IK 2x optimal	-0.0126	0.0562	-0.0491	-0.0189
		(0.0298)	(0.0562)	(0.0485)	(0.0475)
	CCT	-0.0115	0.1003	-0.0272	-0.0243
		(0.0653)	(0.1686)	(0.1072)	(0.1081)
% Poverty	IK optimal	0.0026	0.1184*	-0.0782*	0.0366
		(0.0314)	(0.0698)	(0.0466)	(0.0460)
	IK 0.5x optimal	-0.0178	0.0925	-0.1079	0.0307
		(0.0468)	(0.0959)	(0.0668)	(0.0713)
	IK 2x optimal	0.0060	0.0928*	-0.0600	0.0321
		(0.0246)	(0.0551)	(0.0385)	(0.0359)
	CCT	-0.0410	0.1270	-0.1411	-0.0029
		(0.0697)	(0.1689)	(0.1183)	(0.1093)
% LEP	IK optimal	-0.0072	0.0016	-0.0214	0.0000
		(0.0109)	(0.0184)	(0.0182)	(0.0161)
	IK 0.5x optimal	-0.0163	0.0123	-0.0414	0.0048
		(0.0158)	(0.0225)	(0.0261)	(0.0228)
	IK 2x optimal	-0.0069	-0.0059	-0.0116	0.0007
		(0.0084)	(0.0158)	(0.0139)	(0.0134)
	CCT	-0.0358	-0.0112	-0.1054**	-0.0061
		(0.0264)	(0.0312)	(0.0517)	(0.0457)
School Size	IK optimal	2.0791	-52.4069	58.7390*	-28.9612
		(20.3082)	(41.4863)	(33.9442)	(22.5820)
	IK 0.5x optimal	-8.1219	-61.1657	46.0301	-43.0077*
		(28.1384)	(55.6294)	(53.0092)	(25.6467)
	IK 2x optimal	1.5589	-44.5182	48.8386*	-23.5783
		(16.7618)	(36.1177)	(28.5425)	(19.7112)
	CCT	-17.7314	-97.1022	36.6106	-35.0157
		(50.6042)	(132.7873)	(93.3691)	(65.0315)

Note: Standard errors in parentheses. Dependent variables are school level population measures. Bandwidth determination for the top 3 rows is by Imbens-Kalyanaraman algorithm. The bottom row is the estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table A3. Placebo RD with Pre-determined Variables for schools with no failures X principal experience					
Qualified for bonus every year		All	Exp. > 10	5<=Exp.<=10	Exp. < 5
% Female	IK optimal	-0.0023	0.0102	-0.0073	-0.0086
		(0.0092)	(0.0164)	(0.0125)	(0.0150)
	IK 0.5x optimal	0.0050	0.0400*	-0.0090	-0.0169
		(0.0130)	(0.0228)	(0.0145)	(0.0204)
	IK 2x optimal	-0.0045	0.0091	-0.0066	-0.0110
		(0.0079)	(0.0141)	(0.0122)	(0.0128)
	CCT	0.0268	0.0804	0.0401	-0.0091
		(0.0218)	(0.0550)	(0.0832)	(0.0331)
% Minority	IK optimal	-0.0072	-0.0166	-0.0174	0.0268
		(0.0398)	(0.0674)	(0.0697)	(0.0737)
	IK 0.5x optimal	0.0071	-0.0175	-0.0248	-0.0085
		(0.0461)	(0.0895)	(0.0914)	(0.0925)
	IK 2x optimal	-0.0117	-0.0222	-0.0242	0.0143
		(0.0378)	(0.0611)	(0.0667)	(0.0643)
	CCT	-0.0203	-0.0183	0.2707	-0.1688
		(0.0957)	(0.3521)	(0.3583)	(0.1755)
% Poverty	IK optimal	0.0218	-0.0142	0.0499	0.0521
		(0.0325)	(0.0746)	(0.0558)	(0.0535)
	IK 0.5x optimal	0.0433	-0.0546	0.0287	0.0612
		(0.0425)	(0.1069)	(0.0767)	(0.0864)
	IK 2x optimal	0.0076	0.0068	0.0278	0.0136
		(0.0304)	(0.0675)	(0.0531)	(0.0453)
	CCT	0.0484	0.0689	0.1017	-0.0113
		(0.0744)	(0.3182)	(0.1326)	(0.1338)
% LEP	IK optimal	-0.0007	-0.0053	0.0179	0.0011
		(0.0125)	(0.0210)	(0.0139)	(0.0232)
	IK 0.5x optimal	-0.0026	-0.0031	0.0177	-0.0104
		(0.0152)	(0.0295)	(0.0158)	(0.0311)
	IK 2x optimal	-0.0006	-0.0122	0.0241*	-0.0024
		(0.0117)	(0.0186)	(0.0134)	(0.0210)
	CCT	-0.0064	-0.0070	0.0593	-0.0347
		(0.0369)	(0.0998)	(0.0497)	(0.0549)
School Size	IK optimal	-16.7534	-10.0840	44.1183**	-81.1473**
		(21.6384)	(44.9934)	(19.4519)	(37.1202)
	IK 0.5x optimal	-32.1159	-32.0789	35.4749	-73.1363*
		(24.3204)	(55.8020)	(23.0889)	(40.1655)
	IK 2x optimal	-8.4285	2.9969	50.6010***	-55.7719
		(20.8356)	(41.0666)	(18.8038)	(34.0292)
	CCT	77.9242	-15.0868	92.6551	129.7609
		(58.0762)	(97.8231)	(82.2141)	(98.6491)

Note: Standard errors in parentheses. Dependent variables are school level population measures. Bandwidth determination for the top three rows is by Imbens-Kalyanaraman algorithm. The bottom row is the estimator with robust bias-corrected CI from Calonico, Cattaneo, and Titiunik (2014). *** denotes an estimate significant at the 1% level; ** the 5% level; * the 10% level.

Table A4. Parametric Robustness Checks

Specification (1): base case equiv. to Table 5	Coefficient (Std. Error)
β: Treatment at Threshold	-0.0480** (0.0211)
Specification (2): accountability history equiv. to Table 6	
β_2: No fail schools	-0.0133 (0.0586)
β_4: Schools with at least 1 failure	-0.1105* (0.0670)
Specification (3): by experience dummies equiv. to Table 7	
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 (4): by years of experience equiv. to Table 7	
β_5: Linear Experience	-0.0155** (0.0066)
β_6: Quadratic Experience	0.0010** (0.0005)
Specification (5): interaction of experience and accountability history equiv. to Table 8	
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 variable is math standardized score growth. Specifications control for minority, limited English proficient, free/reduced price lunch eligible, female percents, 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