Opening the Black Box: Behavioral Responses of Teachers and Principals to Pay-for-Performance Incentive Programs

Tom Ahn
Naval Postgraduate School

Jacob Vigdor
University of Washington

Educator incentive programs offer the promise of better aligning teacher compensation with effort, but potentially at the cost of introducing conflict or stress. Using North Carolina data, we examine how teachers' receiving a performance bonus influences perceptions of the efficacy of school leadership and work environment. Surprisingly, results show perceived improved workplace quality in schools that barely missed the bonus, suggesting that the failure prompts positive change in the organization. Further, favorable survey responses and test score improvements are isolated to schools with some history of failures, suggesting a role for experimentation by school leadership to find reforms that work.

I. Introduction

The economics-of-education literature has established that accountability rewards and punishments can lead to improvements in student academic outcomes (Lavy 2002; Jacob 2005; Figlio and Rouse 2006; West and Peterson 2006; Chiang 2009; Chakrabarti 2013, among others). However, with few exceptions, the rewards or punishments themselves are not expected to be the proximate cause of test score gains; their purpose is usually to serve as a prod to get the school to do something different.[1] Spurring action without prescribing what to do is not necessarily effective, as any positive mechanisms invoked by accountability sanctions may be offset

Any opinions expressed herein are those of the authors and not any affiliated organization.

[1] One prominent exception is the No Child Left Behind Act, which prescribed specific corrective measures for schools with each failure to meet its standards. See Ahn and Vigdor (2014).

Electronically Published July 26, 2023.

Journal of Human Capital, volume 17, number 3, fall 2023.

https://doi.org/10.1086/724365

by negative mechanisms. Prior studies have documented undesirable impacts of accountability pressure on teacher turnover rates (especially among new teachers), school disciplinary practices, and other outcomes (Clotfelter et al. 2004). Arguments in favor of accountability systems rely on the presumption that the positive impacts of incentivization are strong enough to outweigh the possible negative impacts of unintended responses or losses to morale. This paper seeks to quantify the extent to which failure to receive a positive reward is linked to losses in self-reported morale or changes perceptions about the quality of a teaching workplace.

While the body of evidence on the effectiveness of accountability systems is growing, most studies are concerned with identifying a clean treatment effect of the reward or punishment itself and not with the actual policy or practice that may mediate or moderate the observed impact. Therefore, while we may now have more confidence to declare whether (and what kind of) accountability is effective, we are still very much in the dark about why certain rewards or sanctions work. As we accumulate more evidence, it is becoming clear that the academic improvements from accountability pressure are not uniform across different schools and students (Ahn and Vigdor 2014). A previous study of the merit-pay system in North Carolina revealed that the effectiveness of accountability pressures varies greatly with the school's history of successes and failures in attaining the bonus (Ahn and Vigdor 2021). Schools that had a small number of failed attempts to meet the standards showed the largest response in test score increases to a failure to receive a bonus in the previous year. Schools that had never previously failed or had chronically failed to qualify for the bonus were less able to respond effectively. In the conclusion of that study, we speculated that schools that had prior experiences with failure also had prior experiences with policy experimentation, allowing those schools to implement adjustments more effectively in the intermediate run. We remained agnostic about what these adjustments were.

In this study, we shed more light on how schools may be responding to accountability pressure. Augmenting North Carolina's administrative data with a Working Conditions Survey (WCS) administered to teachers allows us to examine whether schools facing heightened accountability pressure exhibit differential survey response patterns on items related to school safety, teacher empowerment, professional development (PD), mentoring, and availability of resources and facilities. In the WCS, teachers and principals are asked a battery of questions about teacher time use, the state of school leadership, the relationship between school leadership and faculty, opportunities for PD and mentoring (receiving and providing), and the school's ability to provide the resources and facilities for teachers to educate students in an effective manner.

Using a regression discontinuity (RD) framework, we run two sets of analyses. First, we replicate the main results from Ahn and Vigdor (2021) for our sample to confirm that schools' academic performance in response to a failure to qualify for a pay-for-performance bonus is dependent on their

accountability history. Second, we use the school averages of the WCS responses as outcome variables to examine the treatment effect on teachers' perception of the school environment. Schools with different histories of failures exhibit divergent patterns of changes in their perceptions of the school environment upon treatment. We can identify the set of school environment changes that are observed in schools that generate test score gains. We also characterize school environment changes that are observed in treated schools that fail to generate academic gains.

It is worth emphasizing up front that we cannot infer the change in school environment as producing (or failing to produce) test score growth. While extensive, the WCS may still be incomplete, and the working-conditions change may be correlated with (or caused by) only some unmeasured policy change that directly leads to increased test scores. In addition, the survey primarily measures perceptions of the school environment, not the actual allocation of resources and time or measurable changes in school policy. While we utilize the quasi-experimental RD setup, take advantage of the timing of the administration of the survey and end-of-year tests, and buttress the main findings with several robustness checks, these limitations temper our ability to ascribe causality to the survey responses. Instead, we use the simultaneous changes in teachers' perceptions about the school and test score responses to the treatment to identify potentially important mediating factors associated with observed increases in student academic achievement and describe the schools that implement them.

Conditional on its history of successes or failures in the past, a school's reaction to success or failure this year differs. These differing reactions, exhibited by changes in teachers' perceptions of the school environment, suggest that school leadership may be changing school policies or implementing new ones. An extensive literature examines the efficacy of school policies such as PD, investment in facilities, teacher empowerment, mentoring, and allocation of time for instructional planning.[2] These policies may be some of the reforms that the school attempts to implement when faced with accountability pressure. Intuitively, we may expect academic outcomes to improve when teachers are provided good PD opportunities and high-quality mentoring. Teacher morale may derive from feelings of empowerment and participation in school-level decision-making. A collaborative relationship between teachers and the principal could also lead to higher productivity. These policies could conceivably backfire. If PD and investment in facilities are a wasteful (or at least inefficient) use of funds, if too much teacher autonomy leads to a dilution of the school's mission, or if time and effort spent on mentoring and instructional planning could be put to better use, students may actually derive higher benefit from reforms that are not specifically focused on positively motivating or training teachers.

[2] Below, we review the literature that examines the school environment policies in this study.

The principal responses to accountability failure intimated by this study suggest limitations of our current understanding of how accountability systems lead to academic improvements. In contrast to the popular perception that accountability failure elicits improvements despite adverse morale impacts, we find suggestive evidence of morale improvements alongside a specific set of policy or practice changes in schools with experienced principals. Teachers report positive feelings about implementation of reforms that empower teachers, provide additional resources, and take steps to ensure that teachers feel that they have a bigger stake in the school's success. These changes in the school environment appear simultaneously with improvements in academic performance in the next year, at least in schools that have had a chance to experiment with reforms to find the set of effective policies.

Less experienced principals also appear to attempt to change the school environment. Although the reforms ultimately prove to be less successful, they are not policies that are obviously damaging or wasteful. Indeed, well-informed and well-intentioned policy practitioners and researchers recommend some of the very reforms that teachers report these principals implemented.

As the No Child Left Behind (NCLB) era has given way to an era of greater experimentation in school accountability regimes, it would be most helpful to learn which (if any) of these potential education reforms are associated with effective and ineffective accountability pressures, to provide guidance to policy makers and practitioners on how to better design and implement accountability systems to invoke positive causal mechanisms without activating negative ones.

Literature on school environments/working conditions.—The WCS asks about six broad categories of school/professional environment for teachers that may affect education production and teacher morale: safety, fair and effective use of performance evaluation, teacher empowerment (mutual respect between administration and faculty), adequate resource provision, mentoring, and availability of noninstructional time. Broadly, there is ample qualitative or indirect evidence that working conditions can affect test scores, but credible quantitative studies illuminating the causal pathways by which these improvements arise for some policies are harder to find. Below, we briefly describe what the literature has to say about each of these school environments.

There is a sizable literature examining the role of school safety and test score outcomes, which consistently finds that measures of safety are positively associated with academic performance. However, most of the focus is on how students' perception of danger or misbehavior of peers may directly affect test scores (Fryer and Levitt 2004; Carrell and Hoekstra 2010; Ahn and Trogdon 2017). To date, there has been little examination of how perceptions of school safety may affect teachers. As we show later in the paper, while we find no evidence of an impact of school safety on teachers' likelihood of leaving the school, we do find that teachers'

perception of the safety of the school improves simultaneously with test score increases.

Performance evaluations in economics of education have centered around value-added models. This literature quantifies a teacher's contribution to students' test scores by netting out observable student, classroom, school, and district characteristics, leaving the residual as a measure of the teacher's "value-added" (plus a shock). Studies have also advocated for the use of value-added measures in personnel decisions, most often by firing and replacing teachers with low value-added (Hanushek 2009; Chetty, Friedman, and Rockoff 2014). Recent studies have found that systematic and detailed evaluations can identify good teaching practices that lead to sizable test score gains. In addition, teachers who have received high-quality, classroom-observation-centered evaluations improve teacher performance and retain this increased productivity for years (Kane et al. 2011; Taylor and Tyler 2012; Briole and Maurin 2022). The causal pathway is hypothesized to be from teachers learning to invest strategically in human capital after the evaluation.

There is robust exploration in the psychology and education literatures of teacher empowerment/mutual respect between school leadership and teachers, with most finding a positive relationship among self-efficacy beliefs, job satisfaction, and student achievement (where test score data are available). With some exceptions, many of the papers' methodological focuses are in-depth case studies/interviews or theory-based conceptual models or are reliant on modestly sized surveys (Marks and Louis 1999; Marks and Printy 2003; Caprara et al. 2006). Overall, there is a wealth of corroborating evidence that teacher empowerment and mutual respect positively affect the school environment and teacher job satisfaction, which may lead to increased student test scores. However, we have not been able to find papers that demonstrate direct causal evidence using large datasets and econometric techniques that account for the usual endogeneity and selection issues.

A robust literature in the education-finance field examines the relationship between student achievement and availability of school resources. Evidence is mixed, with many studies showing little impact of increased resources while others show that increased resources can be effective, especially in low-income school districts. Often, the analysis is forced to treat the use of funding as a black box. Resources are funneled to districts or schools, but researchers in many cases cannot observe in detail how those funds are used (Card and Krueger 1992; Burtless 1996; Lemasters 1997; Ladd and Zelli 2002; Hanushek 2003; Hannaway and Stanislawski 2005; Hamilton et al. 2007; Uline and Tschannen-Moran 2008; Lafortune, Rothstein, and Schanzenbach 2018).[3]

As for specific programs that resources could be directed toward, the use of multimedia and technology (such as online tools) has received

[3] One exception is Cellini, Ferreira, and Rothstein (2010).

some attention, but the bulk of the literature has focused on PD activities (Champoux 1999; Sugar, Crawley, and Fine 2004; Yoon et al. 2007). The education literature identifies focus on content, active learning, collective participation by faculty and principals, alignment with school policies and practice, and sufficient duration of training as traits defining a good PD program (Desimone et al. 2002; Penuel et al. 2007). Although there is a sizable qualitative literature affirming the effectiveness of PD in improving teacher pedagogy, measurable impact of PD with the desirable characteristics, as defined by the education literature on student achievement, has been difficult to find (Garet et al. 2008, 2011; Wilson 2013).

As we show in the data, there is near-universal acclaim of mentors among North Carolina teachers. More than 90% of surveyed teachers state that mentors have been important in their careers. The literature has explored the role of mentoring, especially for new teachers, on student academic as well as long-term career outcomes. Mentoring has been suggested as a method to decrease the high rates of early attrition of young teachers (Grissmer and Kirby 1997; Ingersoll and Smith 2004). Mentoring has been found to be effective in increasing retention rates of teachers, but only in low-poverty schools. Most importantly for our study, there is credible evidence that good mentoring can lead to increased student performance. Evidence points to intensity of mentoring, as measured by time spent receiving mentoring, as important for test score increases (Rockoff 2008; Glazerman et al. 2010).

Research on the use of time in education has mainly focused on the efficacy of increased instruction time on test score outcomes (Pischke 2007; Marcotte and Hemelt 2008; Bellei 2009; Taylor 2014). Papers evaluating the efficacy of policy interventions often compare gains to additional school instructional days (Ahn, Aucejo, and James 2022; Aucejo et al. 2022). The WCS focuses on the amount of noninstructional time teachers have, which is assumed to be used for pedagogical planning and collaboration and consultation with other teachers (Reeves, Emerick, and Hirsch 2006). Noninstructional time is almost always unobserved in large administrative datasets. As a result, most research tend to be case studies or small surveys, with almost no evaluation of whether allocation of noninstructional time actually leads to improved academic achievement (Roth et al. 2003; Dever and Lash 2013; Philipp and Kunter 2013).

II. North Carolina ABC Accountability System

In the 1996–97 school year, North Carolina introduced a system of cash bonuses awarded to all teachers in schools meeting test-score-based performance goals as part of its "ABCs of Public Education accountability plan."[4] Schools were assessed by calculating year-over-year growth in math

[4] The acronym stands for strong accountability, teaching the basics, and emphasis on local control.

TABLE 1
AYP and ABC Status from 2006–7

	ABC	
AYP	Yes	No
Yes	472	74
No	254	266

and reading end-of-grade (EOG) test scores. Details of the performance measure computation can be found in Vigdor (2009). Initially, each teacher was paid a $1,000 bonus, but after one year, the state switched to a system that paid out $750 for "expected" performance and $1,500 for "exceptional" performance. A school had "expected" performance if the test score growth in the school exceeded a predetermined threshold based on gains measured from 1993 to 1994, the first two years of test administration, and correction factors for macroshocks in students' performances in the last year. "Exceptional" performance goals were met if the school passed a higher threshold. Approximately 65% of public schools qualified for some bonus payment in 2006–7, the year we examine most closely in this analysis. Of these, about one-third received the $1,500 bonus. Our previous study that measured the treatment effect on scores used an RD design to take advantage of the structure of the merit-pay program.[5]

In the sample academic year for this study, 2006–7, the ABC program and the NCLB program were simultaneously administered in public schools in North Carolina. Because NCLB placed sanctions on schools based on proficiency rates, and not test score growth, there is only a modest correlation between failing to make adequate yearly progress (AYP) for NCLB and failing to qualify for either bonus in the ABC system. Table 1 shows a cross-tabulation of AYP status and bonus receipt for 2006–7. More than 30% of schools are on the "off-diagonal," succeeding under one accountability system but failing under the other. In addition, the escalating nature of NCLB sanctions meant that no schools in 2007 were subject to the "restructuring" sanction, the only sanction out of the set of NCLB corrective measures showing evidence of significant test score effects in RD analysis (Ahn and Vigdor 2014).

III. Conceptual Framework: Learning about the Production Process through Failures

RD analysis often focuses on comparing two sets of observations, of which one is exposed to a "treatment," when the rule for administering the treatment hinges on an observed assignment variable exceeding a prespecified threshold. In this application, the "treatment" can be considered either

[5] The bonus program was discontinued in 2010 because of budgetary pressures from the Great Recession. At its peak, payouts per year reached upward of $100 million.

receiving a monetary bonus in the system described in section II or failing to receive that bonus. In Ahn and Vigdor (2021), we find evidence of significant differences in next-year student performance among schools on either side of the bonus threshold. From a traditional full-information rational perspective, these differences pose a puzzle. The risk of failing to receive a bonus in year t should vary approximately continuously across the threshold for bonus receipt in year $t - 1$; thus, it is hard to explain why a significant difference exists.

In Ahn and Vigdor (2021), we propose that school leaders do not fully understand the education production process and utilize the information contained in a discrete performance signal much the way a motorist reacts to the "check engine" light in a vehicle. If the light is not illuminated, drivers assume that the underlying process is working well and does not require intervention. Similarly, when the school qualifies for the performance bonus, the faculty continues to assume that they are doing well and that substantive reforms are not required.

When the discrete signal suggests that the underlying process is not working well, school leaders are incentivized to change it. In Ahn and Vigdor (2021), we posit and find evidence to support the notion that the inexperienced school leader faces a problem similar to the inexperienced motorist broken down on the side of a highway; they may think of ways to intervene but have little basis for understanding whether their intervention will improve matters or make them worse. More experienced leaders, and specifically those who have prior experience in attempting to adjust the education production process in their school, may be more likely to react effectively in the presence of a signal that adjustments are warranted.[6]

Applied to analysis of the WCS, a model of fully informed rational school leadership would continue to predict no significant difference between schools on either side of the bonus threshold. The incomplete-information model is consistent with differential responses and with those differences varying according to past experience, with a hypothesis that schools that have had some past failures will exhibit responses more closely associated with improved student performance. It is not clear a priori how these responses might affect measures such as teacher morale. While intuitively one might expect education production to be increasing in teacher morale, other things equal, implementing curricular or pedagogical shifts that entail short-run switching costs or increasing effort expectations could negatively affect morale. To the extent that teacher morale is a function of

[6] Schools that fail to receive a bonus may replace school leadership or other personnel. Then, any benefits of experience would be lost. Of course, it may be the case that a school exhibiting repeated failures suffers from leadership incapable of mastering the production process at that school, in which case a new principal could improve performance without prior experience at the school; this is a key finding of Ahn and Vigdor (2014). We show evidence that failing to receive a bonus is not significantly associated with teacher or principal turnover.

perceived efficacy, and perceived efficacy is in turn a function of the discrete bonus signal, failing to receive the bonus could also have a direct negative impact on morale.

IV. Data

We use the public school administrative dataset and the WCS provided by the North Carolina Education Research Data Center, as well as official records from the North Carolina Department of Public Instruction (NCDPI), which contain the official score used to calculate whether the school qualified for the ABC bonus, the composite growth index score. The normalized index score ranges from −0.45 to 0.66; schools with values above zero qualify for at least the $750 bonus.[7] The NCDPI composite index score data for our sample is publicly available online.[8] The WCS is an anonymous survey, and it is impossible to link these responses back to administrative data on teachers. Therefore, most of our analysis is conducted with data collapsed to the school-level observations.

From the administrative dataset, we gather individual-level math test scores.[9] Importantly, we also extract test scores from consecutive years, allowing us to measure test score growth. We construct our test score outcome variable by aggregating these individual test score gains up to the level of the school. In addition to test score data, we have access to a battery of demographic and socioeconomic variables, including race, gender, poverty (via participation in free- or reduced-price-lunch programs), and disability status. We also collect demographic and experience information on teachers.

Summary statistics from the administrative dataset are presented in table 2. These consist of students and teachers in elementary schools serving grades 3–5 in the 2006–7 school year. The dataset contains roughly 350,000 student and 87,000 teacher data points across over 1,000 schools. It should be noted that while student data are restricted to grades 3–5 (since only tested students are entered into the administrative data), the teachers may be assigned to any grade in an elementary school, from kindergarten to grade 5. This dataset also includes substitute teachers.[10] Because we need two years of test score data to calculate growth measures for the student, out-of-state transfer students and others (moving from

[7] Estimating the impact of the $1,500 bonus was not possible, because of small sample sizes.

[8] While we have access to additional years of the administrative and WCS data, we could gain access to only one year of the composite growth index score data. We were unsuccessful in replicating the composite index score with the administrative data, which restricted our sample to only one wave of the WCS.

[9] Reading scores were almost uniformly unresponsive to incentives. Reading and language arts skills may be more dependent on home production, as numerous education policy interventions show lesser impacts in this domain.

[10] Test score growth for students in grade 3 is measured by the difference between the EOG exam and the pretest exam that is administered within the first three weeks of the academic year.

TABLE 2
Summary Statistics from Student-Level and Teacher-Level Administrative Data

Variable	Mean	SD	Minimum	Maximum
	Student Statistics (n = 343,817, Grades 3–5 Only)			
Standardized reading score	1.33E−09	1.000	−3.40	2.67
Standardized math score	−1.78E−09	1.000	−3.84	2.58
Female	.494	.499	0	1
Minority	.430	.495	0	1
Free/reduced-price lunch	.436	.496	0	1
Limited English proficiency	.062	.241	0	1
Disabled	.212	.409	0	1
School made AYP	.488	.500	0	1
	Teacher Statistics (n = 87,522, K–Grade 5)			
Female	.919	.272	0	1
Minority	.138	.345	0	1
New teacher	.067	.249	0	1
Transfer	.056	.231	0	1

private or home schools, for example) who lack previous-year test scores are excluded from the analysis. Figure 1*B* shows how student socioeconomic characteristics are distributed in schools by their index score. Percent eligible for free- and reduced-price-lunch (FRL) and percent minority students are

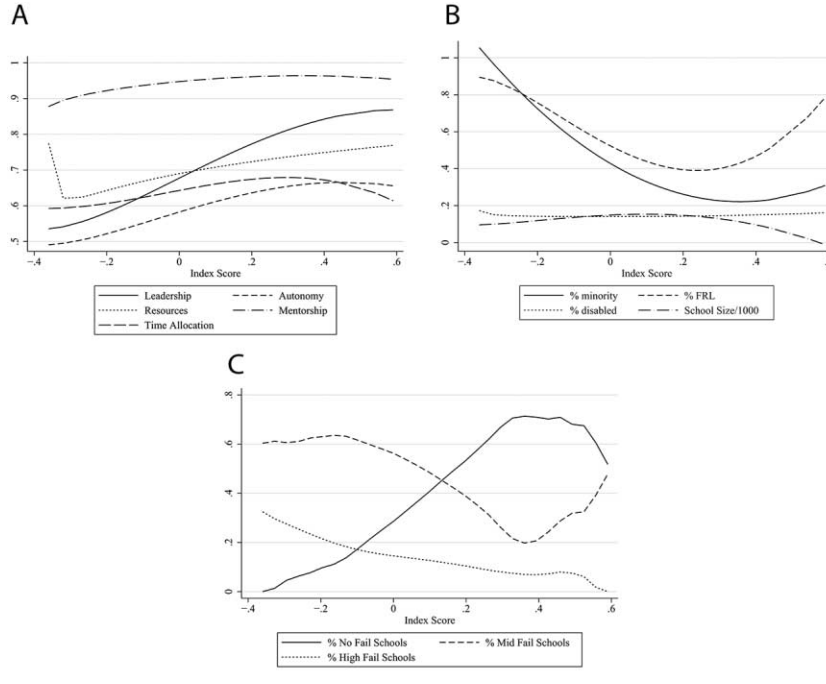

Figure 1.—Distribution of WCS opinion constructs (*A*), student sociodemographic characteristics (*B*), and school accountability history (*C*) across the index score (assignment variable). FRL = eligible for free/reduced-price lunch.

negatively correlated with school performance, while percent disabled and school population are uncorrelated with the index score.[11]

To examine whether schools respond differentially, depending on their history of receiving the bonus, we then divide our sample of schools into three subgroups. The first group of schools (No-Fail) qualified for bonus payments consistently in the five years before 2007. The second group (Mid-Fail) had one or two failures, and the third group (High-Fail) failed to qualify for the bonus three or more times in those five years. Summary statistics for each subgroup from the administrative data are presented in table A.1 (tables A.1–A.18 are available online). Figure 1*C* shows the distribution of schools by their accountability history across the index score. Unsurprisingly, No-Fail schools tend to have high index scores, while Mid-Fail and High-Fail schools are more prevalent in the low-index-score group of schools.

Table 3 contains school-level summary statistics of teacher opinions from the WCS for the entire sample. Overall, teachers feel that they are provided adequate support and feedback from leadership. While they may not have a large voice in some school-level decision-making, such as how the school budget is spent, they are often consulted on education issues and are given autonomy in selecting their methods of instruction. On average, teachers seem to be satisfied with their work environment and their relationship with school leadership. Figure 1*A* presents the distribution of the WCS responses after survey responses are collapsed into five main category of responses (leadership quality, teacher autonomy, resource availability, time allocation, and mentoring) using factor analysis. All five constructs are positively correlated with the index score, showing that high-performing schools are populated with teachers who feel good about the school environment.

Tables A.2.1 and A.2.2 present the same summary statistics from the WCS for our school subgroups. Teachers tend to hold more negative views about their work environment in schools that have failed to qualify for bonus payments. That said, even in the High-Fail group of schools, teachers generally have an optimistic and positive view of their school. Indeed, plans to leave the school within the next few years correlate only weakly with past bonus receipt.

The WCS has a high response rate, at 77.1%. In addition, two-sample *t*-tests of available teacher characteristics (gender, race, and newly hired) from the administrative data and the WCS reveal no statistically significant differences, giving us cautious optimism that response bias should be relatively small.[12] WCSs in subsequent years show even higher response rates

[11] The upward swing noticeable at very high values of the index scores is due to a small number of outliers.

[12] Opinions from the school leadership were also solicited. Perhaps unsurprisingly, principals have a much more positive opinion of their abilities to connect with and lead teachers. They also feel that teachers have a large degree of autonomy, with minimal interference from the leadership. In general, principals and teachers largely agree on the state of the school and the nature of the relationship between leadership and faculty.

TABLE 3
School-Level Survey Summary Statistics from WCS Data (N = 1,063)

Variable	Mean	SD	Minimum	Maximum
	A. Demographic Statistics			
Proportion female respondents[a]	.923	.053	.682	1
Proportion nonwhite respondents[a]	.143	.164	0	1
Proportion of respondents with no prior experience[a]	.067	.067	0	.667
Average years of experience (capped at 25)	12.95	2.728	1.75	25.000
Proportion of respondents national board certified	.106	.084	0	.667
Proportion with MA degree or higher	.310	.128	0	1
Noninstructional hours per week (capped at 12)	2.912	.659	1.125	6.436
Lesson-planning hours per week (capped at 12)	2.043	.472	.729	4.463
Collaborative planning hours per week (capped at 12)	1.491	.390	.375	4.265
Planning outside regular school hours per week (capped at 12)	6.030	1.036	3.071	9.188
Assigned a mentor in the first 3 years of teaching	.903	.184	0	1
Served as mentor within the past five years	.727	.117	0	1
Response rate to WCS	.770	.147	.217	1
	B. Opinion Questions for Teachers: Proportion Responding Somewhat/Strongly Agree			
Overall opinions about the school:				
My school is a good place to teach and learn	.760	.158	0	1
My school is a safe environment	.867	.134	.250	1
I plan to leave this school within the next two years	.150	.079	0	.539
Opinions about school leadership:				
The school leadership in my school is effective	.628	.174	0	1
There is an atmosphere of trust and mutual respect	.670	.204	0	1
The school leadership consistently supports teachers	.707	.191	0	1
Performance evaluations are handled in an appropriate manner	.796	.155	0	1
Procedures for teacher performance evaluations are consistent	.766	.162	0	1
Teachers receive feedback that can help improve teaching	.768	.157	0	1

Opinions about teacher autonomy and involvement in school-level decision-making:				
Teachers are involved in decision-making about educational issues	.608	.190	0	1
Teachers trusted to make decisions about instruction	.743	.168	0	1
Teachers have a role in selecting instructional materials	.500	.185	.031	1
Teachers have a role in devising teaching techniques	.613	.175	.056	1
Teachers have a role in setting assessment practices	.470	.146	.061	1
Teachers have a role in hiring new teachers	.104	.119	0	.815
Teachers have a role in implementing discipline policies	.360	.149	.030	1
Teachers have a role in deciding how budget will be spent	.114	.123	0	1
Opinions about provision of resources:				
Teachers have appropriate instruct. materials and resources	.755	.155	0	1
Funds/resources available for PD activities	.577	.192	0	1
Opinions about mentoring:				
Mentor provided support on instructional strategies	.963	.041	.750	1
Mentor provided support on curriculum	.952	.049	.571	1
Mentor provided support on classroom management	.962	.042	.625	1
Mentor provided support on school/district policies	.961	.044	.688	1
Mentor provided support with encouragement	.975	.033	.750	1
Met often with mentor to plan during the school day	.905	.072	.571	1
Met often with mentor to be observed teaching	.864	.086	.500	1
Met often with mentor to observe teaching	.851	.088	.500	1
Mentor has been important in my career	.907	.070	.595	1
Opinions about allocation of time:				
Hours per week available for noninstructional time	3.333	.964	1.125	6.563
Hours per week available for individual planning	2.314	.675	.729	5.471
Hours per week available for collaborative planning	1.630	.498	.300	4.708
Hours per week spent on work outside of regular hours	5.978	1.032	2.591	9.188

[a] We fail to reject the null hypothesis that the sample means from the teacher-level administrative and WCS data are equal at the 0.01 significance level.

(2010 WCS: 80.4%, 2012 WCS: 86.0%), suggesting that teachers perceive the survey as truly anonymous and perhaps even a useful device for delivering feedback to school and district leadership. Internal reliability testing of WCS from a more recent wave (2014) shows that the survey will produce similar results with similar populations, with Cronbach's alpha around 0.9 (New Teacher Center 2014). Further, as table A.2.1. shows, baseline response rates are very similar across all schools with different accountability histories.[13] All this supporting evidence makes us cautiously optimistic that responses from the WCS can be considered representative opinions of the teachers at the schools.

However, because the survey is voluntary, we cannot completely exclude the possibility that our analysis is affected by response bias. In particular, there are schools with teacher response rates below 25%. Figure 2 shows a histogram of response rates by school. Even in schools with relatively high response rates, teachers who fill out the questionnaire may be more likely to provide a positive report or may be motivated to respond because they are particularly disgruntled. Response motivations could vary systematically on either side of the bonus threshold.

Combining the three datasets to analyze teacher and principal sentiments about the school environment when affected by the accountability system is facilitated by the timing of the standardized exam, results notification to schools, and the WCS administration. Schools are "treated" on the basis of EOG exam outcomes from the 2006–7 academic year. In our sample, EOG exams were administered in May 2007. Teachers and principals were informed of the bonus outcome several months later, at the latest by mid-to-late December 2007. The survey was live for 1 month, during March 2008, with teachers free to access the questionnaire at any time online. EOG exams for the 2007–8 academic year were in May 2008, with bonus results published in December 2008. Therefore, teachers knew before the survey whether they had qualified for the bonus in the previous year. However, the survey was completed before testing in the current year, and certainly before the results were known. Figure 3 summarizes this timeline.[14]

V. Econometric Analysis

In order to estimate the true impact of the failure to qualify for the bonus, we take advantage of the structure of the ABC accountability system, which defines a sharp threshold using the composite index score. Teachers at schools just below this threshold receive no bonus, while those at

[13] Difference in response rates across the threshold is statistically indistinguishable from zero, implying that bonus outcome did not appreciably affect teachers' inclination to fill out the survey. See table 5.

[14] The bonus outcome may be revealed 1–2 months earlier. Anecdotal evidence from teachers' bulletin board web sites show that teachers were receiving the bonus in their paychecks as early as October (http://www.proteacher.net/discussions/showthread.php?t=116168).

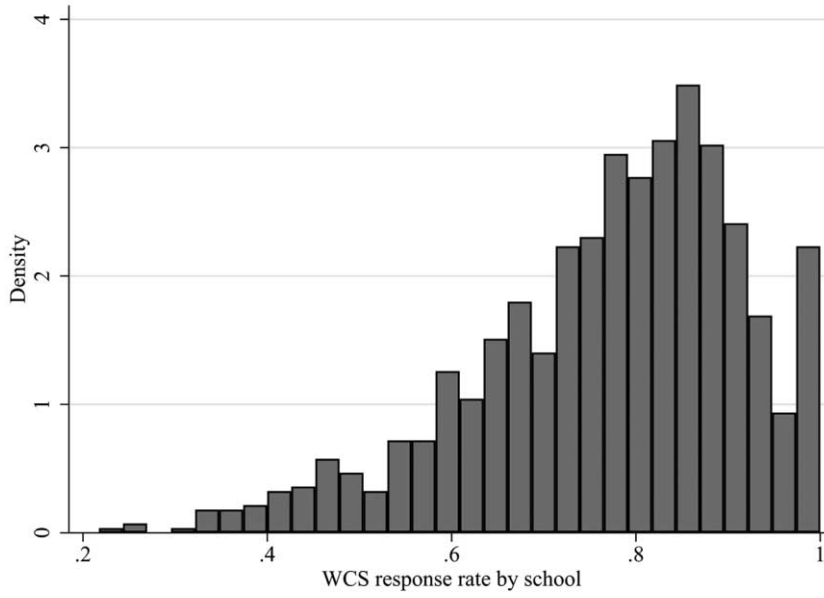

Figure 2.—WCS response rate by school. Minimum: 21.7%, maximum: 100%, mean: 77.1%.

schools just above receive a $750 bonus. With this structure, we use RD analysis to estimate the local average treatment effect of the failure to qualify for the bonus at the threshold. The index score is used as the assignment variable, and school-level averages of the WCS responses are used as the dependent variables. If there are noticeable differences in practices and policies across schools that just failed or just qualified for the bonus, some or all of these could be important in driving the differences in test score growth in the next period.

It is important to emphasize that our local average treatment effects pertain to survey response patterns and not necessarily to the underlying constructs the survey is designed to measure. A positive impact of failure to receive a bonus on measures of teacher autonomy, for example, could reflect an actual impact on teacher autonomy—that principals react to failure to receive a bonus by granting teachers more autonomy. But it

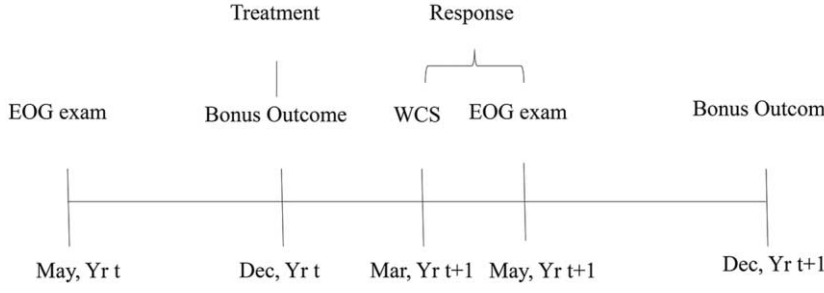

Figure 3.—Timeline of EOG exam, bonus award, and WCS administration.

could also merely change teachers' perceptions of (unchanged) autonomy levels or shift the mix of teachers who choose to respond toward teachers who consistently have perceptions of greater autonomy. Without direct measures of teacher autonomy, we must remain agnostic on this matter of interpretation.

A. How to Think about the Treatment

In this study, the "treated" group contains schools that did not qualify for the bonus in the prior year. It is fair to ask whether the carrot (receiving the bonus), rather than the stick, is the correct treatment indicator. Acknowledging that this is in some sense a semantic decision, since the convention can be easily reversed by flipping the sign of our point estimates, three interconnected pieces of evidence motivate our decision to define the treatment as failure to receive a bonus.

First, between 65% and 80% of schools in any given year received the bonus. That is, many schools in North Carolina were able to cross the threshold. In this sense, achieving score growth to qualify for the bonus was the status quo. Indeed, the $750 bonus was paid out to teachers in schools that had "expected" growth in test scores, using the state's own language. See table 1.

Second, this statistical fact is buttressed by opinions held by teachers and principals at the time. The bonus was looked upon not as a rare reward for extraordinary performance but as something teachers were due as just compensation. Not receiving the bonus, for example, even for exogenous reasons such as budgetary shortfalls, was considered a pay cut.[15] Costly investments in reforms that could lead to a loss of bonus in the current year would then have to be made judiciously.

Third, whether one thinks that the mechanism at work is about the carrot or the stick, we should expect abrupt changes in behaviors and outcomes from the treated schools. Ahn (2016) and Ahn and Vigdor (2021) confirm that it is the group of schools that fail to qualify for the bonus in the previous year that experience a sharp increase in test scores, while schools that did receive the bonus exhibit normalized test score growth closer to zero, and consistent with a smooth trend, in the next year.[16] The clearest departures from a smooth trend are observed in schools below the bonus threshold, especially for the Mid-Fail schools. Changes in dependent variables are flatter for schools above the threshold.[17]

B. RD Model

RD analysis can be performed parametrically or nonparametrically. We use the Hahn, Todd, and Van der Klaauw (2001) nonparametric specification.

[15] See, e.g., http://obsyourschools.blogspot.com/2010/08/good-work-no-bonus.html.

[16] Similar patterns in the WCS response constructs are visible as well. See fig. A.2 (figs. A.1–A.3 are available online).

[17] We discuss this point further in sec. V.B.

We estimate a local linear regression to fit a smooth function to both sides of the discontinuity. Although we have teacher-level records to analyze mediator variables and student-level data for academic outcomes, the qualification for the bonus, and thus the treatment, is determined for the entire school. In order to address this issue, we collapse the WCS and the administrative data to the school level, weighting each school-level observation by the number of teacher-level observations used in computing school-specific means of the dependent variable and covariates.[18] Collapsing the WCS variables to the school level also helps to account for teachers in charge of untested grades.

We conduct several standard tests to confirm that RD is the appropriate econometric framework for analysis (Imbens and Lemieux 2008; Lee and Lemieux 2010). First, in figure 4, we check for evidence of manipulation at the threshold, using a density plot of the index score. An abnormally thick mass immediately to the right of the threshold could indicate that some schools can manipulate test scores in order to ensure that they qualify for the bonus. We also conduct the McCrary test (McCrary 2008) to confirm continuity of the assignment variable at the threshold for the entire sample as well as each subsample in figure 5 and table 4.

Next, we check for balance in covariates on both sides of the threshold. We look for sharp differences in student demographic characteristics, WCS response rates, predetermined test scores, or pressure from other accountability systems. We do not see evidence of data manipulation or abrupt sociodemographic, academic, or teacher salary differences across the threshold for the entire sample as well as for the school subgroups. See table 5.[19] Finally, we also run tests where the treatment is estimated at the threshold values where sanctions are not applied. Results from table 6 imply that treatment occurs only at the threshold point.

The columns labeled "All Schools" in table 7 show our base-case analysis for nonparametric RD methods. Aggregating across all schools, we attempt to capture differences in teacher responses as a function of whether the school qualified for the bonus under the ABC system. The point estimates in the first two rows show the change in standardized test score crossing the discontinuity threshold to the side with no bonus payment. Estimates in the remaining rows show the difference in the fraction of teachers who agree (strongly or somewhat) with the statement in the first column, crossing the discontinuity threshold to the side with no bonus payment. Optimal-bandwidth results are presented (see Imbens and Kalyanaraman 2012). Alternate results at half and double bandwidths, as well as for the robust bandwidth defined by Calonico, Cattaneo, and Titiunik

[18] We estimate an alternative model keeping observations at the teacher level and clustering standard errors at school level. We also estimate RD unweighted and weighted by school enrollment. Results were qualitatively similar. This may be because variation in school enrollment size (and teacher counts) is relatively modest near the threshold.

[19] See fig. A.2 for the standard RD effect figures showing the lack of impact. See table A.3 for alternative-bandwidth estimation results.

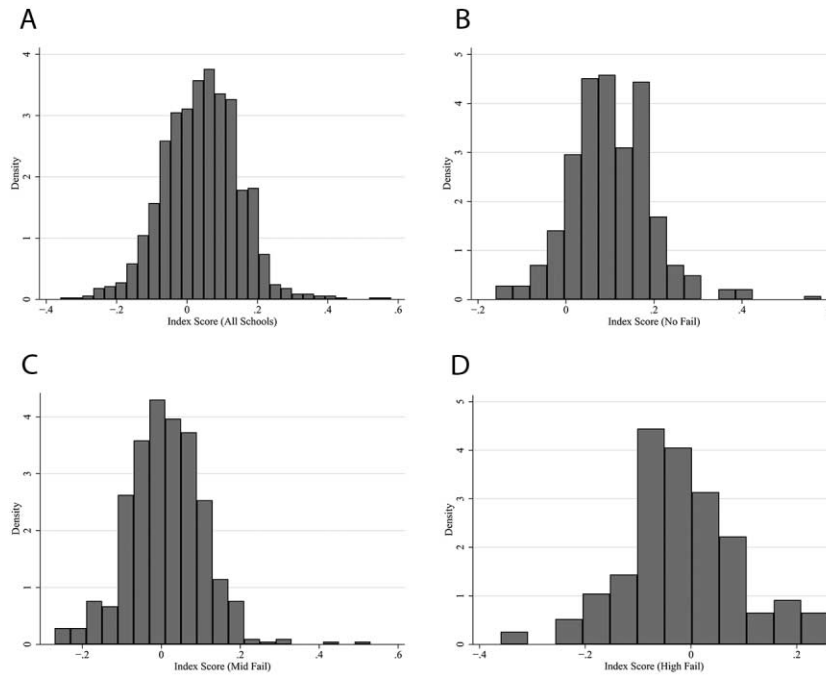

Figure 4.—Density of observations across assignment variable for all (*A*), No-Fail (*B*), Mid-Fail (*C*), and High-Fail (*D*) schools to check for possible presence of manipulation.

(2014; denoted CCT in table 6 and online tables), are presented in table A.4.[20]

Beginning at the top of table 7, the overall impact of just failing to qualify for the bonus in the academic performance of schools next year, compared to schools that just qualified, is indistinguishable from zero. The direction and sign of the parameter estimate point to a modest increase in performance, but the estimate is imprecise. Consistent with Ahn and Vigdor (2021), schools with a history of moderate performance (Mid-Fail) exhibit stronger gains in math, the only statistically significant coefficient of the eight test score gain coefficients presented. We discuss this pattern more fully below.

While we find no significant relationship between failure to receive a bonus and test score gains across all schools, subsequent rows in table 7 document that WCS responses change with the failure to qualify for the bonus. In particular, three of the five measures of teacher autonomy show statistically significant differences across the threshold, implying

[20] RD plots (of constructs of the five domains of teachers' opinions about the school described in table 3, panel B, using factor analysis) splitting the sample of schools by their accountability history are shown in fig. 6 and summarized in fig. A.3. In fig. A.3, time management treatment effects are excluded from the figure, because of parameter size differences. Parameter estimates are available in table A.5.

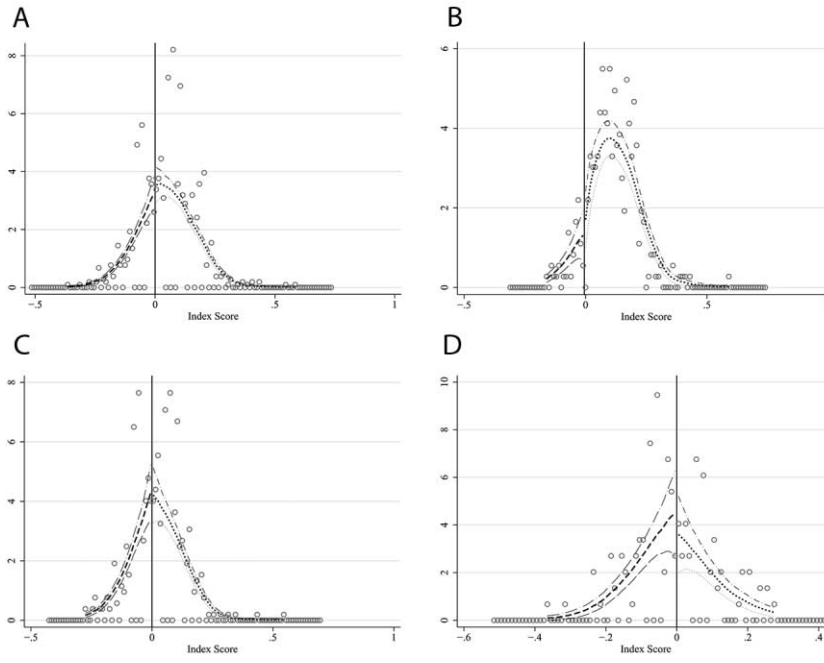

Figure 5.—McCrary test for all (*A*), No-Fail (*B*), Mid-Fail (*C*), and High-Fail (*D*) schools.

that teachers report being more involved in education decision-making, are trusted more to make decisions about class instruction, and have a bigger role in setting grading and assessment practices at the school. Teachers report significantly better resources for PD in schools that barely miss qualifying for a bonus, with favorable responses averaging 5 percentage points higher on a statewide mean of 58%. Four of the five measures of mentoring support also show statistically significantly higher values in schools that fail to receive a bonus.[21] While the magnitudes are small, baseline rates of favorable mentoring support are very high, well above 90%, implying that there is little additional room for improvement.

The absence of significant test score gain results might suggest that the shifting survey responses are rooted in perception or sample selection rather than reality, that these particular working conditions do not influence test scores, or that the well-documented low statistical power of RD designs masks significant test score impacts.[22] However, as described in

[21] An earlier draft considered the opinions of the principal. Results for almost all variables were insignificant. Because principals uniformly report that the working environment is good and that they have good rapport with faculty, negative bonus outcome cannot increase the fraction of principals who change their behaviors or beliefs.

[22] We calculate the statistic $(\tau_\nu - \tau - B)/\sqrt{V} \rightarrow^d N(0,1)$ where $B = h_+^{1+p-\nu}B_+ - h_-^{1+p-\nu}B_-$ and $V = (V_+/nh_+^{1+2\nu}) + (V_-/nh_-^{1+2\nu})$, to derive the value of the power function at the null hypothesis that the treatment effect is zero. Here, τ_ν is the RD estimate with bandwidths from the left and right, respectively, as h_- and h_+; τ is the true impact of the treatment; B_- and B_+

TABLE 4
McCrary Test for Continuity of the Assignment Variable at the Threshold

	Discontinuity	Standard Error	*p*-Value
All schools	.055	.135	.684
No-Fail	.094	.352	.789
Mid-Fail	−.037	.168	.826
High-Fail	−.216	.327	.509

the literature review in section I, domains such as mentoring support, autonomy, and PD have prior research attesting to their effectiveness.[23] Confidence intervals on the test score gain estimates suggest that point estimates are consistent with math test score gains as high as 7% of a standard deviation; reading score gains could reach 4% of a standard deviation.

The absence of significant test score results might also indicate that positive changes captured in the WCS are offset by other, counterproductive leadership reactions. As noted above, we might expect the frequency of counterproductive reactions to depend on the accountability history of the school. Isolating schools by their accountability history reveals different survey response patterns and academic results. A more nuanced story emerges.

As noted above, the treatment results in a statistically significant gain in math test score growth in the Mid-Fail sample. The increase is about 5% of a standard deviation, which is modest in absolute terms but is quite efficient in terms of the overall amount of bonus paid out, if results could be replicated across the entire state.[24] The results may suggest that principals in these schools pursue different policies to increase education production.

The treated No-Fail schools have experienced failure for the first time in at least five years and thus have little recent incentive to alter their educational production process. In contrast with the results for the entire sample, there is no evidence of shifts in teacher-reported autonomy, with coefficients uniformly small and varying in sign. There is more consistent evidence of shifts in teacher-reported mentoring support, suggesting that this avenue is among the first that school leaders pursue when faced with a negative performance signal. Researchers and practitioners have long highlighted the ability of quality mentoring, especially for new teachers,

---

are misspecification biases; and V_- and V_+ are asymptotic variance of the RD estimator, from the left (−) and right (+) sides of the thresholds. See Calonico, Cattaneo, and Titiunik (2014) for more details. Overall, we have sufficient power for the Mid-Fail group. However, some No-Fail and High-Fail estimates suffer from marginal or insufficient power because of the small sample size and estimated effect. See table A.6 for results.

[23] Some of this evidence is qualitative in nature.

[24] Out of 1,063 schools in the sample, 541 are Mid-Fail schools. Fourth- and fifth-grade teachers and the principal comprise roughly one-third of staff subject to the bonus, and they teach approximately 150 students per school. Of the $90 million annually in total bonus payouts, $15.3 million is spent on bonuses that directly affect these students. Then, it costs about $38 per student to raise math scores by 1% of a standard deviation.

TABLE 5
Falsification Test Using Predetermined Demographic Characteristics as Dependent Variables (Optimal Bandwidth)

	All Schools		No-Fail		Mid-Fail		High-Fail	
Placebo Dependent Variable	Coefficient	SE	Coefficient	SE	Coefficient	SE	Coefficient	SE
% Minority	−.033	(.040)	.088	(.128)	−.069	(.056)	.034	(.117)
% Poverty	−.038	(.038)	.071	(.080)	−.076	(.053)	−.003	(.103)
% LEP	−.014	(.014)	−.004	(.042)	−.013	(.018)	.016	(.037)
% Female	−.002	(.021)	.005	(.021)	−.013	(.025)	.046	(.035)
% Disabled	−.010	(.012)	−.018	(.022)	−.003	(.014)	−.024	(.020)
Response rate	.018	(.029)	.025	(.042)	.006	(.031)	−.062	(.106)
No. of AYP failures	−.144	(.167)	.126	(.416)	−.483	(.375)	−.576	(.561)
Reading score prior year	.060	(.045)	.034	(.077)	.078	(.060)	.052	(.096)
Math score prior year	.051	(.043)	−.097	(.104)	.081	(.064)	.057	(.123)
Average salary at school	7.11	(36.99)	−89.00	(70.89)	14.10	(45.82)	36.35	(97.90)

Note.—Bandwidth determination is by the Imbens-Kalyanaraman (2012) algorithm. LEP = low English proficiency.

TABLE 6
Falsification Test using Noncritical Assignment Cutoff Points

	No-Fail		Mid-Fail		High-Fail	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
			At Critical Point +.1			
At optimal bandwidth	.019	(.025)	−.021	(.021)	.026	(.063)
× .5 optimal bandwidth	.031	(.033)	.007	(.028)	.085	(.085)
× 2 optimal bandwidth	.023	(.021)	−.024	(.018)	−.009	(.056)
CCT bandwidth	.004	(.032)	.007	(.033)	.072	(.072)
			At Critical Point −.1			
At optimal bandwidth	−.174	(.162)	−.042	(.032)	.025	(.043)
× .5 optimal bandwidth	−.735***	(.062)	−.011	(.040)	.045	(.048)
× 2 optimal bandwidth	−.122	(.116)	−.035	(.031)	.034	(.040)
CCT bandwidth	−.050	(.152)	−.064	(.151)	.020	(.095)

Note.—Bandwidth determination is by the Imbens-Kalyanarama (2012) algorithm. The bottom row in each panel is for the estimator with robust bias-corrected confidence interval from Calonico, Cattaneo, and Titiunik (2014; CCT).
*** Significant at the 1% level.

to increase teaching effectiveness and reduce turnover (see Glazerman et al. 2010). There is a positive coefficient in the specification tracking PD resources, but it is one-third the magnitude of the result for all schools and not statistically significant. One significant result appears among No-Fail schools where no pattern was detected in the full sample: teachers report spending significantly less time on collaborative preparation.

Despite this evidence of improved mentoring, math scores are either unchanged or actually down, on average, after the treatment in No-Fail schools. Even if there is a direct impact of mentoring on test score gains, perhaps it is not surprising that we see little impact. Effective mentoring might take more than a year to exhibit positive effects. Mentoring requires protected time for the mentor to attend training programs, prepare materials, and observe and provide feedback to mentees. Similar time for mentees to absorb and incorporate feedback must be carved out as well (see Goldrick 2016). However, our estimates show that noninstruction and preparation time is either unchanged or reduced after the first accountability failure. It is telling that while mentoring activity has increased for this sample, the fraction of teachers who report that mentoring is important to their careers does not increase.

Mid-Fail schools show different patterns in school environment changes. Teachers at these schools report greater autonomy and a larger role in school-level decision-making. This autonomy does not necessarily rise to the level of an equal partnership between faculty and leadership. We do not see statistically significant changes in teacher reports of having a say in student assessment practices or hiring decisions. However, teachers report that the leadership trusts them more in the classroom. Principals also work to carve out additional individual instructional planning time.

TABLE 7
WCS RD Results—Schools by Accountability History (Optimal Bandwidth)

	All Schools ($N = 1{,}063$)		No-Fail ($N = 372$)		Mid-Fail ($N = 541$)		High-Fail ($N = 150$)	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Math score improvement	.029	(.019)	−.045	(.047)	.050**	(.020)	.068	(.052)
Reading score improvement	.001	(.019)	.043	(.034)	.002	(.028)	−.055	(.036)
Overall opinions about the school:								
My school is a safe environment	.020	(.016)	−.010	(.062)	.059***	(.019)	−.019	(.040)
I plan to leave this school within the next 2 years	.008	(.012)	−.008	(.024)	.002	(.014)	.031	(.025)
Opinions about school leadership:								
There is an atmosphere of trust and mutual respect	.042	(.027)	−.072	(.069)	.097**	(.039)	−.038	(.077)
School leadership consistently supports teachers	.040*	(.024)	−.068	(.070)	.089***	(.033)	.001	(.054)
Opinions about autonomy/involvement—teachers:								
Are involved in decision-making about educational issues	.031	(.023)	−.095	(.068)	.068**	(.033)	.128**	(.060)
Are trusted to make decisions about instruction	.069***	(.024)	.012	(.060)	.088**	(.034)	.130**	(.057)
Have a role in devising teaching techniques	.060**	(.024)	−.018	(.061)	.079**	(.033)	.066	(.048)
Have a role in setting student assessment practices	.051**	(.022)	−.011	(.048)	.048*	(.028)	.096**	(.046)
Have a role in hiring new teachers	.014	(.014)	.019	(.054)	−.002	(.016)	.094***	(.036)
Opinions about provision of resources:								
Teachers have appropriate instructional materials and resources	.002	(.019)	−.060	(.041)	.023	(.022)	−.031	(.047)
Funds/resources available for PD activities	.053**	(.026)	.016	(.053)	.052	(.035)	.048	(.064)
Opinions about mentoring—mentors:								
Provide support on instructional strategies	.015**	(.006)	.020**	(.010)	.007	(.007)	.037**	(.017)
Provide support on curriculum	.021***	(.007)	.038***	(.012)	.013	(.009)	.024	(.017)
Provide support on classroom management	.013**	(.006)	.011	(.010)	.009	(.008)	.011	(.013)
Provide support on school/district policies	.013**	(.006)	.031***	(.011)	.004	(.008)	.031*	(.016)
Has been important in my career	.009	(.010)	.017	(.018)	−.001	(.015)	.028	(.024)
Opinions about allocation of time:								
Noninstructional hours have increased by × hours	.127	(.087)	−.114	(.208)	.209*	(.108)	.344	(.240)
Individual planning hours have increased by × hours	.074	(.054)	−.039	(.138)	.191**	(.075)	.343*	(.195)
Collaborative preparation hours have increased by × hours	−.022	(.051)	−.204**	(.102)	.054	(.061)	−.106	(.145)

Note.—Bandwidth determination is by the Imbens-Kalyanaraman (2012) algorithm.
* Significant at the 10% level.
** Significant at the 5% level.
*** Significant at the 1% level.

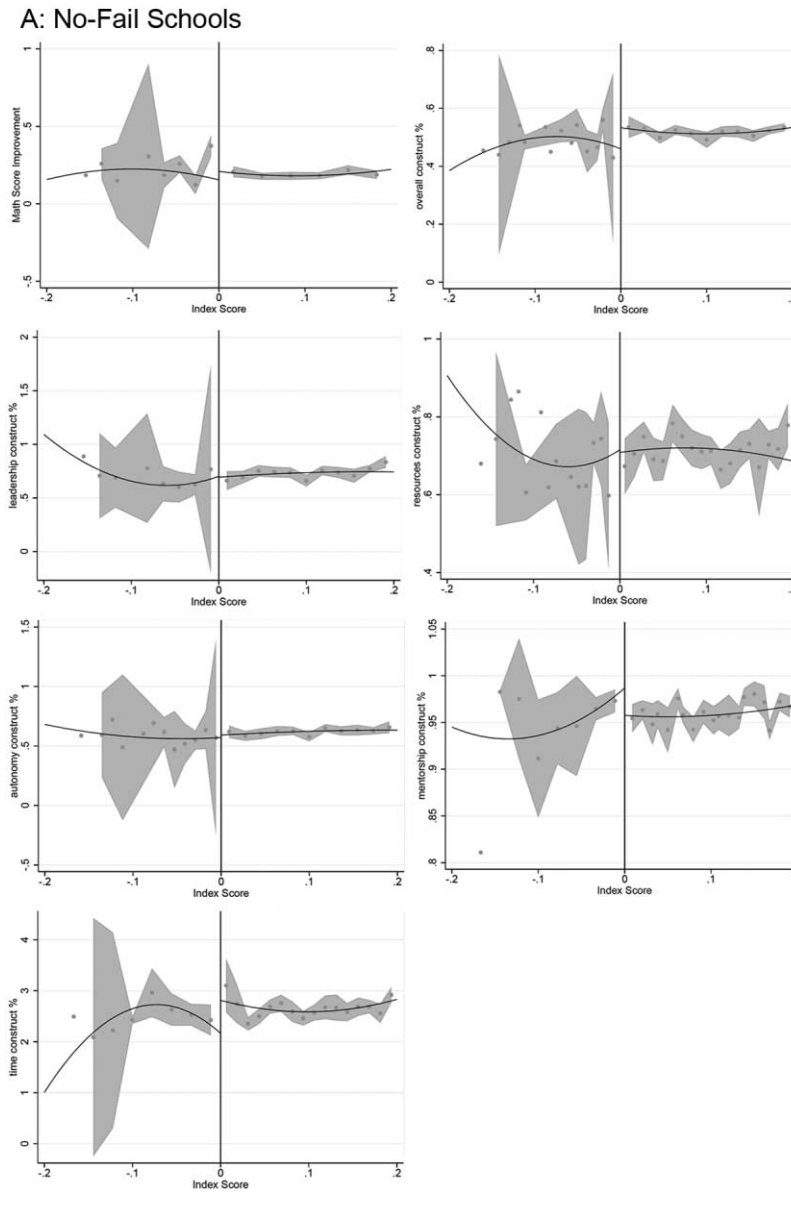

Figure 6.—RD illustration of math score improvement, overall, leadership, resources, autonomy, mentorship, and time management constructs for No-Fail (*A*), Mid-Fail (*B*), and High-Fail (*C*) schools.

RD estimates on mentoring in this subsample are uniformly insignificant. We interpret the small point estimates and standard errors to mean that mentoring levels are maintained at the No-Fail sample level. Schools to the left of the bonus threshold do not appear to be deemphasizing

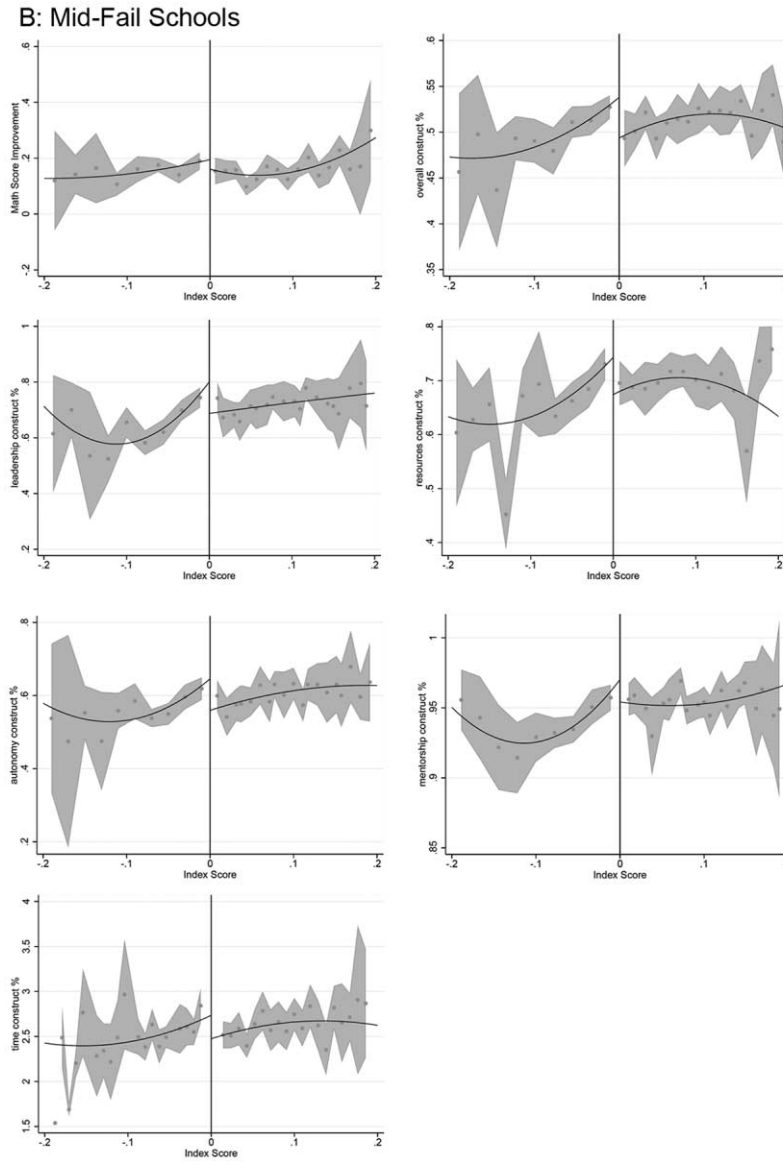

Figure 6. (*Continued*)

mentoring, which would have shown up as negative and statistically significant estimates, nor do they seem to ramp up mentoring further.

The new reforms Mid-Fail schools appear to pursue when ineligible for a bonus are concurrently seen with an overall improvement in faculty perceptions about the school leadership. Teachers report that they feel more supported and that their relationship with leaders is more collaborative. This general feeling of well-being even spills over into the perception that the school has become safer.

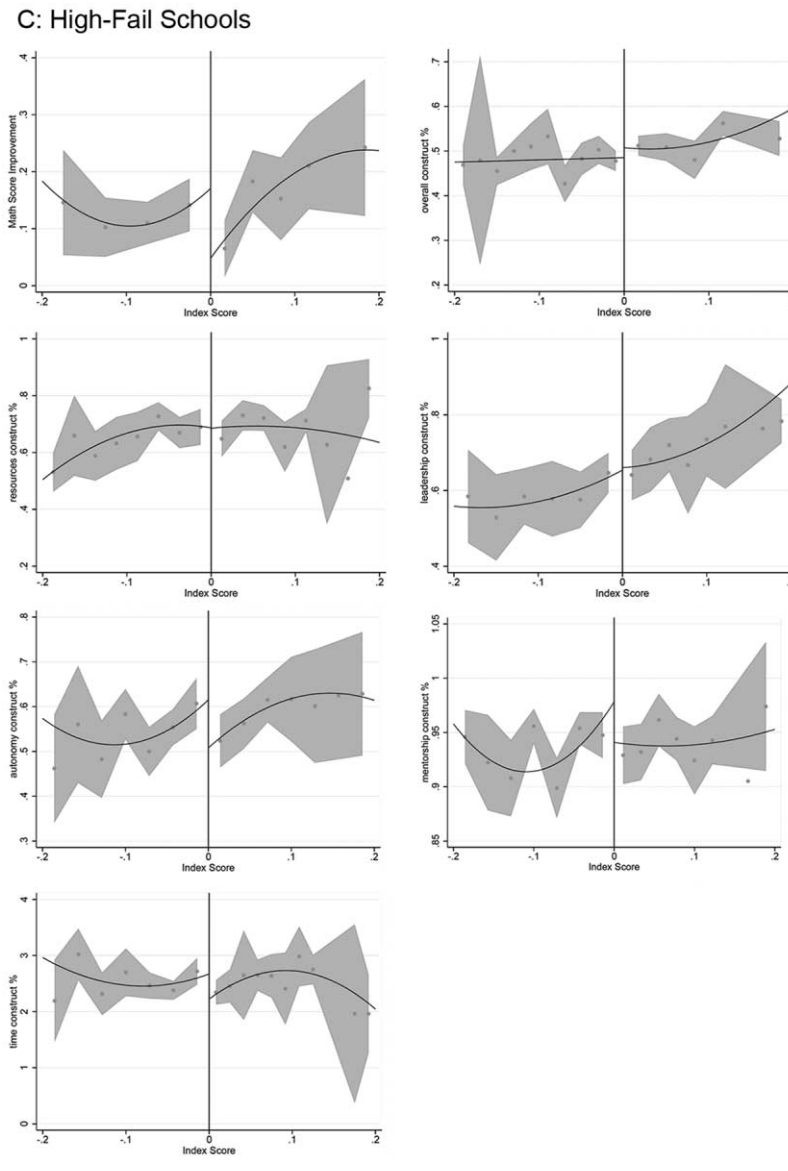

Figure 6. (*Continued*)

The High-Fail sample yields imprecise estimates of test score effects, although point estimates are substantial in magnitude. RD estimates imply that among these schools perceived teacher involvement in school-level decision-making increases after a school misses the bonus threshold. Teachers report having more input into setting student assessment practices and hiring decisions. In fact, many of the point estimates are substantially larger than in the Mid-Fail sample. There is also some weaker evidence of carving

out more time for instructional preparation. Again, it is worth emphasizing that zero or the statistically insignificant point estimates do not necessarily indicate that the school is doing away with reforms of the Mid-Fail sample. In the RD framework, evidence of such a move would most clearly appear as a negative and significant estimate.

These reforms, unlike those in the Mid-Fail sample, are not associated strongly with an additional improvement in teacher perception of leadership quality. The null hypothesis that teachers at Mid-Fail schools have a more positive opinion of the leadership, compared to teachers at High-Fail schools, cannot be rejected, yielding a p-value of less than .0002.[25]

The failure of these additional measures to improve the teacher-principal relationship suggests limits to the efficacy of reforms, or at least a large degree of heterogeneity in their impacts, in schools that have had persistent trouble in meeting state-established standards. Schools in this category suffer from larger deficits in overall morale (see table A.2.2), and even sustained efforts by an incumbent principal to reform the school may not be enough to improve the relationship between leadership and faculty.[26]

While the narrative above may be provocative, analysis with RD suffers from low power, especially for a nonparametric local-linear regression framework with relatively small sample sizes. In order to provide some evidence on the robustness of these results and our interpretation of these findings, we replicate our main nonparametric RD estimates using a variety nonparametric and parametric specifications.

Nonparametric RD results using factor-loaded survey constructs and teacher-level observations with clustered errors largely correspond to findings from the main model (see tables A.5 and A.8). Parametric specifications in both OLS and logistic frameworks also broadly agree with the main model (see tables A.9 and A.10.).[27] The one exception is parametric RD estimates with linear functions of assignment variable, which return very few statistically significant results. More flexible functional forms seem to better align with nonparametric RD results, with responses of schools below the threshold displaying especially large movement.[28]

[25] A systemic test for differential effects between Mid- and High-Fail schools across all outcome measures in a parametric RD framework yielded few statistically significant parameter estimates. See table A.7.

[26] Ahn and Vigdor (2014) showed that for schools that continually fail to make NCLB's AYP (5–6 consecutive years), restructuring that replaces the leadership is the only effective means of increasing test scores.

[27] Parametric specifications control for linear, quadratic, or cubic functions of the assignment variable that vary on both sides of the threshold. Control variables are gender, limited English proficiency, race, disability, and school size. Of 152 estimates ([2 specifications] × [4 subsamples] × [19 outcomes]), quadratic specification returns 30 and cubic 43 coefficients significant at the 5% level. For No-Fails, of the 11 significant responses, 9 concerned mentoring. For Mid-Fails, of the 30 significant responses, 8 were about leadership, 13 on autonomy, and 5 were about time management. For High-Fails, of the 11 significant responses, 6 concerned autonomy, and 4 regarded mentoring.

[28] There were only four estimates significant at the 5% level. The poor results may be due to sharply localized responses, such that a discontinuity exists only near the threshold. Such an outcome is difficult to fit with a line. Some suggestive evidence is in the local-linear figures

The closer alignment of specifications with flexible functional forms to nonparametric RD results is related to the discussion above about whether the treatment is the failure or the success in qualifying for the bonus. Parameter estimates of the flexible functional forms buttress our argument that it is the "stick" that matters more. See table A.11, which shows parameter estimates for the linear and quadratic functions of the assignment variable below and above the threshold with math score growth as the outcome variable. For the linear specification, the Mid-Fail schools show steeper slopes when below the threshold. For the quadratic specification, both the linear and the squared terms from below are statistically significant, while from above, only the linear term is significant, and it is smaller in absolute magnitude than the linear term from below.

Importantly, because our empirical strategy estimates treatment effects on many school environment outcome variables, we must account for multiple-hypothesis testing. Even if the treatment is ineffective for all school environment variables, we may falsely reject the null hypothesis at a critical value of 0.05 (or 0.1) if we assess enough outcome variables. To account for this "false discovery" of statistically significant results, we calculate false discovery rate–sharpened $q$values (Benjamini and Hochberg 1995) in tables A.12.1–A.12.5. Unsurprisingly, some salient treatment effects become statistically insignificant. However, treatment effects for Mid-Fail schools on math score improvement, trust in leadership, teacher autonomy, and individual time management remain statistically significant across many specifications.

Finally, we estimate several models to buttress our conceptual framework. Using data from the 2006 WCS, we use the changes in mean opinions of teachers at the school as the new dependent variables (see table A.13). While point estimates are mostly statistically insignificant, the estimated impacts are qualitatively similar to our main results.[29] To test whether nonresponse of No-Fail schools may be from principals' belief that the first failure was an anomaly, we estimate RD with bonus receipt in the next year as the outcome. Table A.14 shows that results are insignificant regardless of accountability history.[30] To test whether observed treatment effects are due to the arrival of new leadership or teachers, we replace the outcome variable with composition of the faculty. Table A.15 shows that there is no significant change in turnover for either principal or teachers.[31]

---

in fig. 6. Especially for Mid-Fails, a sharp upturn to the left of the threshold is visible. A best-fit line would have large standard errors, which may explain why flexible functions have better agreement with the nonparametric estimates. We caution that this is speculative.

[29] Differences in mean opinion do not necessarily reflect true changes. Teacher turnover and differing response rates make comparison across time less robust. Schools could also have received a treatment in 2007 or new leadership. Given these issues, and because we lose one-third of the sample, it is not surprising that the results are noisy.

[30] If this belief is true, then nonresponse is the optimal response.

[31] There is one interesting pattern to principal turnover. While results imply higher probability of new leadership at No- and High-Fail schools after treatment, the opposite

VI. Discussion

We use a unique survey of teachers and principals and a student administrative dataset of public schools in North Carolina to analyze changes in responses to opinion questions about the school's environment after it has been "treated" with a failure to qualify for the state's merit-pay bonus. Taking advantage of the sharp state-defined criterion that determines whether teachers at a school receive merit bonuses, we use nonparametric and parametric RD analysis to find that treated schools with some history of accountability failures demonstrate both significant differences in teachers' perceptions of their working conditions and sizable growth in test scores. Schools without a history of failure and experimentation and schools with a chronic history of failures, on the other hand, show weaker evidence on both fronts.

This paper pushes the literature forward in two important directions. First, we take advantage of the quasi-experimental, near-random assignment nature of the sample near the threshold to treat the different accountability history sample runs as snapshots of a learning process. Our empirical evidence is consistent with a model of a decision maker who experiments with a set of reforms from a menu of possible choices, gradually learning which are effective but eventually exhausting effective reforms to increase test scores.[32]

Second, by focusing on the subset of schools that actually deliver higher test scores, we provide exploratory evidence on possible mechanisms linking the stimulus of accountability incentives to the response of improved achievement. Improvements in teacher morale, largely ignored in the literature—usually because of measurement issues—appear hand-in-hand with academic gains in schools with a history of learning and experimentation in education reforms. In particular, teachers in these schools exposed to negative sanctions report improvements in the relationship

is true for Mid-Fails. Mid-Fails may have principals who have experimented with reforms. Table A.16 examines whether the decision to transfer or leave the profession depends on the school's academic achievement. For teachers, better performance is associated with a lower likelihood of transfer or exit. Estimation with splines shows the likelihood of transfer or exit increasing faster as performance declines, especially when the score is below the threshold. We do not find significant results for principals. Table A.17 isolates those who transfer to a new school. Teachers move to schools with higher performance. For principals, the only salient result is that years of missing the bonus is positively correlated with higher academic performance at the new school. Overall, changes in perception are likely not due to turnover in leadership, but we cannot fully rule out the possibility of churn in faculty affecting WCS results.

[32] An additional piece of evidence that points to the role of accountability history is shown in table A.18. We split the sample using average salary instead of accountability history and run the RD model with factor-loaded outcome variables. One potential explanation for the efficacy of the treatment for some schools may be due to the composition of the faculty. Experienced and higher-paid faculty may be more receptive to cooperation with school leadership as equal partners. Overall, we see very few statistically significant results. The earnings of teachers do not seem to play a role in school environment/working-conditions change in response to accountability pressure.

between faculty and school leadership and increased teacher autonomy. This pattern contradicts concerns that negative sanctions would cause reductions in morale. Even with the specter of previous failure hanging over them, principals willing to work with teachers appear to substantively improve the school environment. At the same time, empowerment may have its limits, and chronically underperforming schools do not exhibit significant test score gains even as they continue to implement reforms.

If we better understand the mechanisms linking accountability reforms to higher test scores, we may be able to design more efficient and equitable accountability policies. Ideally, we may be able to disburse information about effective reforms to be implemented across all schools without the need to incentivize by identifying winners and losers.[33]

It is worth mentioning again that caution is warranted in attributing test score gains to improved teacher perceptions. Improved opinions about the school environment in survey responses do not necessarily imply improved educational experiences for students. It is unclear which changes (if any) are directly causing the test score increase. Some of the observed changes may merely correlate with the proximate cause of the test score gains. For example, in schools that successfully raise test scores, teachers report that principals foster a more mutually respectful work environment and grant teachers more autonomy in setting classroom education practices. If increased autonomy is one of the ways in which mutual respect is shown to teachers, the improvement in morale, and not the increased authority in class, could be the cause of test score increases. On the other hand, perhaps morale has nothing to do with test score gains. Allowing teachers the freedom to teach the way they want and getting out of their way may be the key to raising student achievement. It is also possible that the survey is incomplete and that it fails to ask about relevant changes implemented by the school leadership. Ultimately, more research is required to truly understand not only the ifs of accountability systems' effectiveness but also the whys and the hows.

How generalizable are the effects?—The RD estimates in this paper reflect local average treatment effects. That is, we can attribute the differences in test scores and survey responses recorded only to schools immediately surrounding the bonus threshold. To maximize the usefulness of the insights from this analysis, it is worth exploring how generalizable the conclusions are.

Whether schools farther away from the threshold would have exhibited similar patterns had they been in the counterfactual scenario depends on the expected growth in test scores, how school leadership responded to sanctions if at all, and how teacher perceptions evolved as a result. Schools that failed to meet the bonus threshold by more than a trivial amount may

[33] We caution that any actual changes in the school environment were implemented voluntarily by the principal. Mandating such reforms may not result in the same observed impact in morale or test scores.

have implemented very different policy or practice interventions, or none at all. The net effect on teacher perceptions is similarly difficult to infer. The demoralizing effect of not receiving a bonus might outweigh the positive impact of any intervention in schools well below the threshold.

At the other end of the spectrum, schools far above the threshold may experience a demotivating effect. Schools that are assured of the bonus would not feel additional incentive from the accountability policy. Some corroboration is provided by Ahn (2013), finding that teachers in schools far away from the threshold (thus close to 0% or 100% possibility of receiving the bonus) are absent from work more often, compared to comparable teachers in schools nearer the threshold.[34]

Even if schools farther below the bonus threshold pursued interventions identical to the ones in schools seeing the test score gains and improved teacher perceptions shown here, the effects could be heterogeneous. While it is impossible to definitively claim that the reforms we highlight would be effective in all schools, there is some suggestive evidence that schools below the threshold may be well positioned to implement effective reforms. Figure 1*A* plots the average WCS survey responses reduced to single constructs for each of the five broad categories. Higher values on the vertical axis represent a higher fraction of respondents at the school having a positive view of the school's environment. The positive trends for all five categories indicate that there may be more room for improvement in schools with lower index scores. These schools may have "low-hanging fruit" reforms that are easier to implement, while schools with higher index scores (and already better work environments) may not have a ready list of new changes that will further improve morale.

References

Ahn, Tom. 2013. "The Missing Link: Estimating the Impact of Incentives on Teacher Effort and Instructional Effectiveness Using Teacher Accountability Legislation Data." *J. Human Capital* 7 (3): 230–73.

———. 2016. "A Theory of Dynamic Investment in Education in Response to Accountability Pressure." *Econ. Letters* 149:75–78.

Ahn, Tom, Esteban Aucejo, and Jonathan James. 2022. "The Importance of Matching Effects for Labor Productivity: Evidence from Teacher-Student Interactions." Working Paper no. 2106, Dept. Econ., California Polytechnic State Univ., San Luis Obispo.

Ahn, Tom, and Justin G. Trogdon. 2017. "Peer Delinquency and Student Achievement in Middle School." *Labour Econ.* 44:192–217.

Ahn, Tom, and Jacob L. Vigdor. 2014. "The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina." Working Paper No. 20511 (September), NBER, Cambridge, MA.

[34] We are not claiming that teachers at schools far from the threshold would cease to care about students' education. We are claiming that the accountability system would not provide more motivation if the threshold is too far away.

———. 2021. "When Incentives Matter Too Much: Explaining Significant Responses to Irrelevant Information." *J. Human Capital* 15 (4): 629–64.
Aucejo, Esteban, Patrick Coate, Jane Cooley Fruehwirth, Sean Kelly, and Zachary Mozenter. 2022. "Teacher Effectiveness and Classroom Composition: Understanding Match Effects in the Classroom." *Econ. J.* 132 (648): 3047–64.
Bellei, Cristián. 2009. "Does Lengthening the School Day Increase Students' Academic Achievement? Results from a Natural Experiment in Chile." *Econ. Educ. Rev.* 28 (5): 629–40.
Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *J. Royal Statis. Soc. B* 57 (1): 289–300.
Briole, Simon, and Éric Maurin. 2022. "There's Always Room for Improvement: The Persistent Benefits of a Large-Scale Teacher Evaluation System." *J. Human Resources*. https://doi.org/10.3368/jhr.1220-11370R1.
Burtless, Gary, editor. 1996. *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success*. Washington, DC: Brookings Inst. Press.
Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6): 2295–326.
Caprara, Gian Vittorio, Claudio Barbaranelli, Patrizia Steca, and Patrick S. Malone. 2006. "Teachers' Self-Efficacy Beliefs as Determinants of Job Satisfaction and Students' Academic Achievement: A Study at the School Level." *J. School Psychology* 44 (6): 473–90.
Card, David, and Alan B. Krueger. 1992. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *J.P.E.* 100 (1): 1–40.
Carrell, Scott E., and Mark L. Hoekstra. 2010. "Externalities in the Classroom: How Children Exposed to Domestic Violence Affect Everyone's Kids." *American Econ. J. Appl. Econ.* 2 (1): 211–28.
Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010. "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design." *Q.J.E.* 125 (1): 215–61.
Chakrabarti, Rajashri. 2013. "Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida." *Econ. Inquiry* 51 (1): 500–526.
Champoux, Joseph E. 1999. "Film as a Teaching Resource." *J. Management Inquiry* 8 (2): 206–17.
Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *A.E.R.* 104 (9): 2593–632.
Chiang, Hanley. 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement." *J. Public Econ.* 93 (9–10): 1045–57.
Clotfelter, Charles T., Helen F. Ladd, Jacob L. Vigdor, and Roger Aliaga Diaz. 2004. "Do School Accountability Systems Make It More Difficult for Low-Performing Schools to Attract and Retain High-Quality Teachers?" *J. Policy Analysis and Management* 23 (2): 251–71.
Desimone, Laura M., Andrew C. Porter, Michael S. Garet, Kwang Suk Yoon, and Beatrice F. Birman. 2002. "Effects of Professional Development on Teachers' Instruction: Results from a Three-Year Longitudinal Study." *Educ. Evaluation and Policy Analysis* 24 (2): 81–112.
Dever, Robin, and Martha J. Lash. 2013. "Using Common Planning Time to Foster Professional Learning: Researchers Examine How a Team of Middle School Teachers Use Common Planning Time to Cultivate Professional Learning Opportunities." *Middle School J.* 45 (1): 12–17.
Figlio, David N., and Cecilia Elena Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *J. Public Econ.* 90 (1–2): 239–55.

Fryer, Roland G., Jr., and Steven D. Levitt. 2004. "Understanding the Black-White Test Score Gap in the First Two Years of School." *Rev. Econ. and Statis.* 86 (2): 447–64.

Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, et al. 2008. "The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement." Report no. NCEE 2008-4030, Nat. Center Educ. Evaluation and Regional Assistance, Inst. Educ. Sci., US Dept. Educ., Washington, DC.

Garet, Michael S., Andrew J. Wayne, Fran Stancavage, James Taylor, Marian Eaton, Kirk Walters, Mengli Song, et al. 2011. "Middle School Mathematics Professional Development Impact Study: Findings after the Second Year of Implementation." Report no. NCEE 2011-4024, Nat. Center Educ. Evaluation and Regional Assistance, Inst. Educ. Sci., US Dept. Educ., Washington, DC.

Glazerman, Steven, Eric Isenberg, Sarah Dolfin, Martha Bleeker, Amy Johnson, Mary Grider, Matthew Jacobus, and Melanie Ali. 2010. "Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study." Report no. NCEE 2010-4027, Nat. Center Educ. Evaluation and Regional Assistance, Inst. Educ. Sci., US Dept. Educ., Washington, DC.

Goldrick, Liam. 2016. "Support from the Start: A 50-State Review of Policies on New Educator Induction and Mentoring." Policy report, New Teacher Center, Santa Cruz, CA.

Grissmer, David, and Sheila Nataraj Kirby. 1997. "Teacher Turnover and Teacher Quality." *Teachers Coll. Rec.* 99 (1): 45–56.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1): 201–9.

Hamilton, Laura S., Brian M. Stecher, Julie A. Marsh, Jennifer Sloan McCombs, Abby Robyn, Jennifer Lin Russell, Scott Naftel, and Heather Barney. 2007. *Implementing Standards-Based Accountability under No Child Left Behind: Responses of Superintendents, Principals, and Teachers in Three States.* Santa Monica, CA: RAND.

Hannaway, Jane, and Maggie Stanislawski. 2005. "Responding to Reform: Florida School Expenditures in the 1990s." Urban Inst., Washington, DC.

Hanushek, Eric A. 2003. "The Failure of Input-Based Schooling Policies." *Econ. J.* 113 (485): F64–F98.

———. 2009. "Teacher Deselection." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway, 165–80. Washington, DC: Urban Inst. Press.

Imbens, Guido, and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Rev. Econ. Studies* 79 (3): 933–59.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *J. Econometrics* 142 (2): 615–35.

Ingersoll, Richard M., and Thomas M. Smith. 2004. "Do Teacher Induction and Mentoring Matter?" *NASSP Bull.* 88 (638): 28–40.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *J. Public Econ.* 89 (5–6): 761–96.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying Effective Classroom Practices Using Student Achievement Data." *J. Human Resources* 46 (3): 587–613.

Ladd, Helen F., and Arnaldo Zelli. 2002. "School-Based Accountability in North Carolina: The Responses of School Principals." *Educ. Admin. Q.* 38 (4): 494–529.

Lafortune, Julien, Jesse Rothstein, and Diane Whitmore Schanzenbach. 2018. "School Finance Reform and the Distribution of Student Achievement." *American Econ. J. Appl. Econ.* 10 (2): 1–26.

Lavy, Victor. 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." *J.P.E.* 110 (6): 1286–317.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *J. Econ. Literature* 48 (2): 281–355.
Lemasters, Linda Kay. 1997. "A Synthesis of Studies Pertaining to Facilities, Student Achievement, and Student Behavior." PhD diss., Virginia Polytechnic Inst.
Marcotte, Dave E., and Steven W. Hemelt. 2008. "Unscheduled School Closings and Student Performance." *Educ. Finance and Policy* 3 (3): 316–38.
Marks, Helen M., and Karen Seashore Louis. 1999. "Teacher Empowerment and the Capacity for Organizational Learning." *Educ. Admin. Q.* 35 (5): 707–50.
Marks, Helen M., and Susan M. Printy. 2003. "Principal Leadership and School Performance: An Integration of Transformational and Instructional Leadership." *Educ. Admin. Q.* 39 (3): 370–97.
McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *J. Econometrics* 142 (2): 698–714.
New Teacher Center. 2014. "2014 North Carolina Teacher Working Conditions Survey: Design, Validity, and Reliability." Research brief, New Teacher Center, Santa Cruz, CA.
Penuel, William R., Barry J. Fishman, Ryoko Yamaguchi, and Lawrence P. Gallagher. 2007. "What Makes Professional Development Effective? Strategies That Foster Curriculum Implementation." *American Educ. Res. J.* 44 (4): 921–58.
Philipp, Anja, and Mareike Kunter. 2013. "How Do Teachers Spend Their Time? A Study on Teachers' Strategies of Selection, Optimisation, and Compensation over Their Career Cycle." *Teaching and Teacher Educ.* 35:1–12.
Pischke, Jörn-Steffen. 2007. "The Impact of Length of the School Year on Student Performance and Earnings: Evidence from the German Short School Years." *Econ. J.* 117 (523): 1216–42.
Reeves, Cynthia, Scott Emerick, and Eric Hirsch. 2006. "Creating Non-Instructional Time for Elementary School Teachers: Strategies from Schools in North Carolina." Center for Teaching Quality, Chapel Hill, NC.
Rockoff, Jonah E. 2008. "Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City." Working Paper no. 13868 (March), NBER, Cambridge, MA.
Roth, Jodie L., Jeanne Brooks-Gunn, Miriam R. Linver, and Sandra L. Hofferth. 2003. "What Happens during the School Day? Time Diaries from a National Sample of Elementary School Teachers." *Teachers Coll. Rec.* 105 (3): 317–43.
Sugar, William, Frank Crawley, and Bethann Fine. 2004. "Examining Teachers' Decisions to Adopt New Technology." *J. Educ. Tech. and Soc.* 7 (4): 201–13.
Taylor, Eric. 2014. "Spending More of the School Day in Math Class: Evidence from a Regression Discontinuity in Middle School." *J. Public Econ.* 117:162–81.
Taylor, Eric S., and John H. Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *A.E.R.* 102 (7): 3628–51.
Uline, Cynthia, and Megan Tschannen-Moran. 2008. "The Walls Speak: The Interplay of Quality Facilities, School Climate, and Student Achievement." *J. Educ. Admin.* 46 (1): 55–73.
Vigdor, Jacob L. 2009. "Teacher Salary Bonuses in North Carolina." In *Performance Incentives: Their Growing Impact on American K–12 Education*, edited by Matthew G. Springer, 227–50. Washington, DC: Brookings Inst. Press.
West, Martin R., and Paul P. Peterson. 2006. "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments." *Econ. J.* 116 (510): C46–C62.
Wilson, Suzanne M. 2013. "Professional Development for Science Teachers." *Science* 340 (6130): 310–13.
Yoon, Kwang Suk, Teresa Duncan, Silvia Wen-Yu Lee, Beth Scarloss, and Kathy L. Shapley. 2007. "Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement." Issues and Answers report, REL 2007–No. 033, Regional Educ. Laboratory Southwest, Nat. Center Educ. Evaluation and Regional Assistance, Inst. Educ. Sci., US Dept. Educ., Washington, DC.