DOES THIS MEAN WE GET AN A? CAUSAL IMPLICATIONS OF CHANGES IN SCHOOL ACCOUNTABILITY

SAMUEL KAMIN, PH.D.
POSTDOCTORAL RESEARCH ASSOCIATE
NEAG SCHOOL OF EDUCATION, UNIVERSITY OF CONNECTICUT
ORCID 0009-0001-6049-3420

ABSTRACT

Many states and districts in the United States use school report cards to share accountability data in which K–12 schools are rated on a variety of metrics, including test scores, which create a categorical grade or rating. These report cards are shared with the public as a mechanism of school accountability and in the process of school choice. This paper explores the causal impact of a school report card used by the New York City Department of Education which was not attached to specific rewards and/or sanctions. I use a regression discontinuity approach to analyze the impact of receiving a lower rating. I find that just receiving a low rating leads to an increase in Math score growth in comparison to similar schools just beyond the cut point, although no such effect is found in English score growth. I also explore implications in the context of school/district policy and leadership.

Keywords: School accountability, testing, regression discontinuity, bounded rationality

The introduction of the No Child Left Behind (NCLB) Act ushered in a host of changes to U.S. public schooling, including new content standards, the introduction of Annual Yearly Progress requirements for all students, and a substantive increase in testing requirements (Linn et al., 2002). While schools in the past were accountable to municipalities and states to varying degrees, NCLB formally required all schools to collect annual testing data to verify their progress towards the goal of all students achieving proficiency in math and reading by 2013–2014 (Dee & Jacob, 2011).

Another key element of NCLB was a new focus on the public sharing of these aforementioned school-level data (Dee & Jacob, 2011), with most states and many districts beginning to publish reports on individual school quality. States and districts often formatted these reports as report cards, sometimes even mimicking classic A to F grades. For these report cards, schools are rated on a collection of measures of varying scales, but a final grade is determined through some scaling mechanism. School report cards have been widely examined and researchers have found far-reaching consequences of their implementation, including changing parents' choice of schools away from low scoring schools (Friesen et al., 2012), and impacting housing markets as high scoring schools drive up property prices (Figlio & Lucas, 2004). There is also evidence that school report cards shift behavior within schools; Chakrabarti (2007) found that schools receiving low scores on school report cards focus on students at or near minimum criteria cutoffs for proficiency.

A salient question, then, is whether school report cards are working efficiently and as intended: to communicate school quality to parents, as well as share data with district and school employees to effect change. A second related question is whether schools substantively change practices based on the information provided to increase student achievement, rather than limited and particular effects. For example, if test scores are increasing, they may only be increasing for specific subgroups within a given school, suggesting only certain students are receiving increased attention because of a new focus on the rating. Last, it is possible that schools and their leaders may respond to the report card rating itself as an inherent signal, as opposed to some particular reward or sanction that may come attached to a particular rating.

New York City provides a particularly interesting opportunity for investigating the impact of school report card systems. In 2015, the New York City Department of Education (NYCDOE) transitioned from an A to F report card system with attached consequences and rewards to a goal-based system with less specific grade metrics. In this paper, I contribute to the causal literature on mechanisms of school accountability by examining the impact of this post-2015 report card system in New York City. Specifically, this paper addresses the following research questions:

1. What is the causal impact of just receiving a lower school report card rating on exam scores?

2. What is the causal impact of just receiving a lower school report card rating on relevant achievement-oriented subgroups of students?

In this paper, I leverage the fact that NYCDOE-defined categorical Student Achievement ratings are sharply determined from a continuous score and use a regression discontinuity approach to examine the causal impact of just receiving particular low Student Achievement ratings in comparison to schools just receiving the higher score. This quasiexperimental approach yields causal estimates of the impact of just receiving the lower rating.

While prior papers have examined the NYCDOE's A to F report card system and found positive impacts on learning outcomes (Rockoff & Turner, 2010; Winters & Cowen, 2012), this paper examines the impacts of a newer, more holistic report card system which is substantively different in design and intention (discussed more in the following pages). Because the new, post-2015 system was entirely separate from sanctions and rewards, as opposed to the A–F system of the past, any measurable impacts on learning outcomes from this new system can be directly attributed to the rating itself and not any potential consequences. To measure the potential impact on learning outcomes, I developed multiple test score growth metrics from the years since the shift in policy to examine the impact of just receiving a given rating, namely score growth across grades and movement of specific student subgroups.

In summary, I find that being just assigned a particular low rating ("Approaching Target") has three notable impacts: first, there is a positive impact on math score growth; second, there is a negative impact on the proportion of students in the lowest, Level 1 math achievement category (i.e., there are proportionally fewer students in the lowest performance category the following year); third, there is a positive impact on the proportion of students in the proficient categories. There is not statistically significant evidence of similar trends in English test scores, however. Additionally, I do not find any significant impact of just receiving a rating of "Meeting Target" in comparison to similar schools receiving a rating of "Exceeding Target."

In the following sections, I discuss prior research on accountability systems and provide context regarding the specific report card policy in New York City. I then discuss my analytic approach and report my findings. Finally, I discuss these findings in the context of policy and note potential future areas of research on the subject.

BACKGROUND

PRIOR RESEARCH ON ACCOUNTABILITY SYSTEMS

Researchers have examined the impact of the strict accountability imposed and inspired by NCLB. Indeed, extant literature found evidence that the use of strict accountability scores had some notable positive impacts on student achievement and school practice. Chiang (2009) examined the threat of low accountability scores using a regression discontinuity design and found evidence that the pressure of a low score and the sanctions that are threatened therein increase math scores, shift pedagogy, and lead to the introduction of new curricula. The evidence that accountability scores lead to real, substantive shifts is largely replicated by Rouse et al. (2013). Relatedly, Carnoy and Loeb (2002) examined the relative strength of accountability systems across states, defining strength as the amount of pressure placed on schools to improve test scores based on state mandates; they found low strength accountability systems have little to no state-level accountability to increase student test scores, while high strength accountability systems place specific demands (including rewards or sanctions) on schools that meet or fail to meet testing thresholds. They also found that the strength is positively related to NAEP math scores.

Prior research has also examined temporal differences within states as they shift from simply providing public reports of achievement to threatening sanctions and offering rewards for low or increasing performance (respectively). Hanushek and Raymond (2005) defined two distinct categories of accountability systems. First, they described a "system" as a mechanism in which states "[publish] outcome information on standardized tests for each school along with providing a way to aggregate and interpret the school performance" (p. 12). They differentiated, however, between "report card" states and "consequential" states; the former simply report out the data, while the latter attach specific consequences. Throughout the 1990s and early 2000s, many states transitioned from no data-based accountability system to a low (or no) consequence "report card" system to a "consequential" system with rewards and/or punishments; Hanushek and Raymond (2005) leveraged that shift to find that the introduction of consequential systems increased state-level NAEP scores, although not for all subgroups of students.

There is no consensus that strict accountability mechanisms are a panacea for issues of low student achievement, however. Jacob (2005) demonstrated that while there were increases in test scores in Chicago Public Schools after the introduction of a strict accountability system; those gains were driven by positive shifts in test-specific skills and student effort on exams. These gains may not be entirely productive or efficient if the long-term goal is raising student achievement, given the mechanisms identified are limited and test-specific. Deming et al. (2016) found that while the risk of receiving a low school rating may have positive impacts on schools receiving a high score had little impact. Further, low-scoring students in schools pressured to receive a higher rating may have actually experienced negative impacts on exam scores, as well as an increased likelihood of being classified into special education. Last, Deming and Figlio (2016) demonstrated that high-stakes testing (and its related accountability measures) led to increased and disproportionate attention being paid to "bubble" students (students on the threshold of achieving proficiency on a particular exam).

SCHOOL PROGRESS REPORTS AS AN ACCOUNTABILITY SYSTEM IN NEW YORK CITY

In 2007, the NYCDOE created a new "School Progress Report" protocol to assess its schools. Using a combination of student achievement data including test scores and credit accumulation, parent survey data, and other observational data gathered during superintendent review, numeric scores were calculated on a 1–100 scale, which were then collapsed by predefined bands into A–F grades. The grades were also intended to be linked to rewards and consequences, including bonus pay for teachers for successful schools and potential school closure for those with lower grades (Gootman & Medina, 2007). Further, schools with a D or F rating were required to implement formal plans of school improvement, students in F schools were eligible for a special transfer process, and schools that met high grade thresholds were eligible for school-based budget bonuses, as well as principals earning personal bonuses (Rockoff & Turner, 2010). Low scores are also used as justification by district administration for staffing and administrative changes in schools in which they were received (Winters & Cowen, 2012). The fact that the scores were directly linked placed this policy squarely in the "consequential" bucket, as defined by Hanushek and Raymond (2005, p. 306).

These reports were also widely available for public consumption, and designed to be interpretable by parents, educators, and others. The NYCDOE created carefully presented digital and print versions of these reports which prominently featured schools' assigned letter grade, as well as selected other information. These reports were circulated at schools, in school selection publications developed by the NYCDOE, and made available both at each school's official website and that of the NYCDOE, including past years' reports (Corcoran & Pai, 2013).

There were novel elements to these report cards, beyond simply their accessibility to the public, that attempted to correct prior issues in school accountability policies. Specifically, they used school peer groups, used to compare schools within more similar groupings, as opposed to comparing against the entire city school population. There were numerous ways the use of peer groups is important. First, strict and universal school accountability policies are often influenced by out-of-school factors (Hamilton & Koretz, 2002). Knowing this, the report cards were supposed to allow schools to be compared against peer groups with similar out-of-school circumstances (e.g., number of students in poverty, entering student academic preparation, etc.), and provide a report card grade that was contextualized in the reality that different schools serve different students. Second, these report cards included some growth measures instead of city-wide normed achievement metrics, again presenting the opportunity for equity in the consideration of schools with differing circumstances out of their control, in this case prior student achievement. Simply, schools would not be punished with a low accountability grade for serving students who entered with lower prior test scores than other schools in the city. Yet it is still unclear whether the use of these peer groups had the intended balancing impact on schools' grades; according to Corcoran and Pai (2013), the Peer Index (the collapsed measure developed by NYCDOE which was used to group schools) did not have a notable impact on schools' overall grades due to the diversity within the peer groups. This suggests the use of peer groups may not have actually adjusted the scores towards the end of providing balance across differing out-of-school circumstances.

The causal impact of the NYCDOE version of school report cards on student achievement has been investigated in two papers: Rockoff and Turner (2010) and Winters and Cowen (2012). Moreover, both studies used regression discontinuity approach to examine the impact of receiving a particular grade on exam scores. Rockoff and Turner (2010) examine grade 3–8 test scores and find significant positive impacts of receiving an F relative to a D, or a D grade relative to a C in both math and reading scores. Winters and Cowen (2012) add specificity to a similar analysis by adding student-level characteristics and identifiers, providing the ability to follow students from school to school across years. They find positive impacts on student test scores of receiving an F relative to a D, particularly in English scores, and those gains were persistent across multiple years. Together, the pieces suggest that NYCDOE's school report cards do have a positive impact on test scores for schools at or near the cut points.

A NEW PROGRESS REPORT: THE "SCHOOL QUALITY REPORT"

Despite the positive impacts of the prior report card system, NYCDOE made substantive changes to the School Progress Report in 2015. There were a number of elements to this shift in policy. First, as noted by Corcoran and Pai (2013), the peer groups that were designed to balance the prior report grades by comparing schools against similarly situated “peer” schools actually had little impact on the overall scores in the initial iteration of the report cards. Not only did the peer groups not work as designed, but the NYCDOE also believed these peer groupings created an unfair competitive attitude between the schools being compared (NYCDOE, 2018). Thus, the new progress reports removed the use of the comparison group in score calculation, although interestingly the NYCDOE did choose to include some reference to an unpublished comparison group on the reports themselves, merely suggesting the relevance of the comparison group and not actually using the group to calculate scores and/or ratings.

The most substantive change, though, was a shift from the aforementioned A to F categorical grade scale to a new four-level categorical scale, which labeled schools as Excellent, Good, Fair, or Poor in reports designed for parents. In reports designed for teachers and administrators, the same four-level categorical scale was labeled as Exceeding Target, Meeting Target, Approaching Target, and Not Meeting Target (the latter of these labeling schemes will be referred to for the rest of the paper). These new metrics were described as “gentler” (Wall, 2014, title), and described by then-chancellor of the NYCDOE Carmen Fariña as “looking beyond test scores and focusing on making sure… each school has what it needs for sustained and continuous growth” and a “transformed… approach [to] school accountability” (Darville, 2014, quoted speech).

Still, the most prominently placed measure was for “Student Achievement” which combined student test scores and credit accumulation. As before, these four-tier categorical ratings are assigned by collapsing a continuously calculated numeric score. While the numeric score was and remains publicly available, it is published in a format perhaps too complex for the general population and not formatted, designed, or documented for those without some knowledge of statistics. This suggest any decisions by parents, students, or teachers may be made, not from the continuous numeric score, but rather from the EGFP label. In addition, the scores were no longer criterion referenced — rather, they were built on pre-set targets determined by the NYCDOE, although these targets were not consistent across years.

Last, the new School Quality Reports were also no longer tied to accountability measures or bonuses; instead, these reports were designed for schools and leaders to inform their planning and allow families to learn more about their school (NYCDOE, 2015). This marks a distinct shift away from the “consequential” school accountability mechanism as described by Hanushek and Raymond (2005) towards one in which school performance is still aggregated and publicized, but without the same predetermined rewards or sanctions.

CONCEPTUAL FRAMEWORK

This paper examines the causal impact of receiving a particular categorical Student Achievement rating on a school report card beyond the impact of the numeric Student Achievement score. Policymakers and reformers in NYC adopted report cards to “set expectations for schools and promote school improvement” (NYCDOE, 2018, p.1). However, if there are measurable impacts of the categorical rating beyond that of the numeric rating, it is not obvious why the categorical Student Achievement rating specifically would have any impact on the activities of a school or leader. Indeed, all the information (e.g., test scores and survey responses) used to build the numerical achievement score, which then determines the rating, is known to school leaders ahead of the release of the grade. Further, the information contained in the newer, more holistic report card was specifically designed to help schools "identify and address specific strengths and weaknesses" (NYCDOE, 2015, Overview section). Last, this policy shift represents a move away from a consequential accountability system (as defined by Hanushek & Raymond, 2005) to one without specifically pre-known consequences. Why, then, might the categorical rating itself have any impact above the variety of known information that informs the rating?

From a purely rational approach, schools and their leaders should work to maximize student achievement outcomes and thus improve all progress report numerical scores, regardless of the cut points and letter grades with which those numeric scores are associated. Simply, if you raise test scores, you raise your achievement score. Unless there are stated punishments or rewards for entering/exiting certain categorical ratings, there is no obvious reason why a rating category would cause any change above and beyond the impact of the numeric score. Further, given the information for specific schools within each school's report card, the most efficient or rational behavior may be to specifically target areas of weakness in the report card.

However, rational choice theory (Simon, 1956) explains that not all behavior is as rational as expected. Actors may not search for the best option; rather, a good move might be chosen as it is safer. When actors respond in these ways, they are "satisficing" (Simon, 1956, p. 9). Understanding why schools and their leaders may behave by satisficing is further explained by Simon's (1955) theory of bounded rationality. Simon (1956) argues that actors can rarely take advantage of all the information provided to them, and instead make choices about how much and which information of which to take advantage. Considering these concepts, school leaders may have an overwhelming amount of information at their disposal, to the point where they may not be able act on all of it. Thus, leaders may only use some of that information in deciding which proverbial levers to pull to impact student learning. In this paper, I examine the possibility that the comparatively limited information in the Student Achievement categorical rating demonstrably causes some schools and their leaders to make changes that lead to positive academic outcomes in the form of test score growth.

While the actual behaviors of school leaders are not observed in this study, there are numerous ways extant literature has established schools' responses to accountability reforms. For example, Shipps and White (2009) examine the differences in school principal behavior before and after increased accountability policies in New York City. They found that principals paid closer attention to bureaucratic expectations and market-style accountability, each of which are directly connected to the New York City progress reports.

Bureaucratic expectations are inherently part of school progress reports in that they are developed and shared by the Department of Education; they are treated as reviews of schools' performance for parents and school staff alike. Further, all forms of standardized progress reports inherently align with market ideology (Engel, 2000) in that they suggest intra-district comparisons and competition. In this system, even though schools are scored at least in part against their own achievement goals, each school is still given a label on a consistent and comparable metric against other schools in the city. While not every student can choose their elementary school, and thus elementary schools may not fall cleanly into the market phenomenon described above, the reports still provide information on how a school is doing in direct comparison to its peers. School leaders may use comparatively lower Student Achievement ratings, then, as signals to change their behaviors in ways that are different from schools assigned higher ratings.

Existing literature also demonstrates ways schools and their leaders respond to accountability pressure in unequal ways across student groups, further suggesting a satisficing approach. For example, Booher-Jennings (2005) uncovered the use of educational triage in response to Texas' accountability system, in which teachers and administrators diverted resources to attend to students close to the threshold of passing (i.e. "bubble kids") and students that were known to count for the school's accountability rating. A similar set of circumstances could be relevant in New York City. Schools are commonly judged by their percentage of students meeting proficiency (NYCDOE, 2019). Similar to the findings in Booher-Jennings (2005), schools in New York City may also be practicing educational triage and addressing some of these subgroups differentially based on their Student Achievement rating.

METHODS

DATA/SAMPLE

To answer research questions on the impact of school report card grades, I used the "Student Achievement" ratings from all Elementary, Middle, and Kindergarten through eighth grade schools (n = 1091) in the New York City Department of Education from 2014–15 through 2018–19 school years. These years were selected because these were the first years the new reports were used and include all available years of data at the time of writing (with an exception for 2016–17 described below). As noted before, the Student Achievement rating is on a four-level categorical scale and built from a 1–5 continuous measure known as the Student Achievement Score. This continuous metric is built from a complex formula taking into account student achievement, future credit accumulation, and performance relative to a NYCDOE-assigned target. These data create my assignment (score) and treatment (rating) variables.

For my outcome variables, I construct a variety of grade-level test score growth metrics for grades 3 through 8. In New York State, every 3rd through 8th grade student completes an annual Math and English (ELA) exam each Spring. These data were downloaded from the NYCDOE website in Excel format and merged with the quality report data by a NYCDOE-assigned school ID number and year. Schools without test scores for both years, generally new or closed schools, were excluded. Similarly charter schools, who have different reporting requirements, were also excluded. These test scores are reported as collapsed at the grade by subject by school level, from 2014 through 2019; I then use them to construct year-to-year growth measures for each grade-subject-school. Excluded from the analysis is growth from the 2016–17 to 2017–18 school years, as New York State revised the exam in the 2017–18 school year to reduce the number of days tested and substantively changed the scaling of exam scores. As a result, the scores from 2017–18 are comparable to the following year (2018–19), but not prior years. These metrics are described below.

DESCRIPTIVE STATISTICS

In Tables 1 and 2, I present descriptive data. Table 1 presents the distribution of Student Achievement ratings, grouped both by school-year (one observation for each school-year combination) and grade-subject (one observation for each grade-subject combination within each school-year). Of note is the uneven distribution across the rating categories. There are few (n=55) schools that received the “Not Meeting Target” rating label across all years, less than 2%. Further, the majority of schools received a “Meeting Target” rating; just over half. The remaining schools were roughly evenly distributed between “Approaching Target” and “Exceeding Target.” Also of note were the relatively consistent percentages across the school-year and grade-subject breakdowns, which suggests no substantive differences within the grades served between schools with different Student Achievement rating categories.

There are a few key observable differences between schools receiving different Student Achievement ratings; in Table 2, I present key variables that highlight some of those differences. For example, schools that received lower scores tended to have a larger percentage of students of color. Schools that received lower scores tend to have slightly higher percentages of students with disabilities, higher Economic Need Index scores, and more students chronically absent. Schools with higher Student Achievement ratings tend to have more experienced principals, although not more experienced teachers.

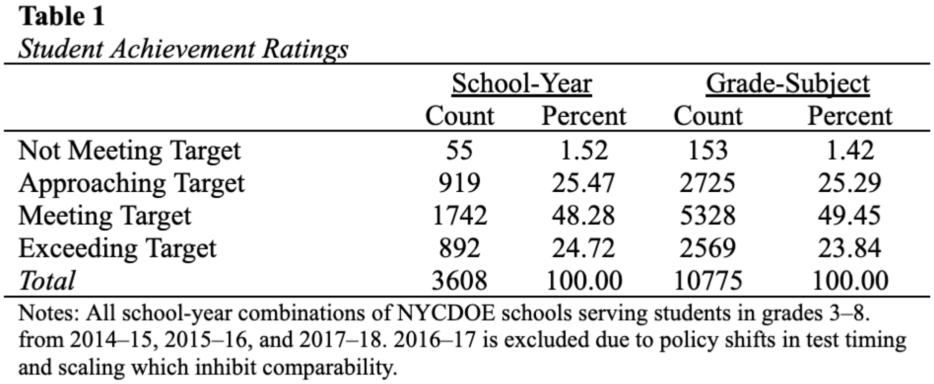

Table 1
Student Achievement Ratings

	School-Year		Grade-Subject	
	Count	Percent	Count	Percent
Not Meeting Target	55	1.52	153	1.42
Approaching Target	919	25.47	2725	25.29
Meeting Target	1742	48.28	5328	49.45
Exceeding Target	892	24.72	2569	23.84
Total	3608	100.00	10775	100.00

Notes: All school-year combinations of NYCDOE schools serving students in grades 3–8. from 2014–15, 2015–16, and 2017–18. 2016–17 is excluded due to policy shifts in test timing and scaling which inhibit comparability.

OUTCOME MEASURES

My outcomes of interest are generally described as test-related growth in the year following the assignment of a given Student Achievement rating. Table 3 presents outcome averages for each of the four Student Achievement rating levels, grouped into two panels by subject area. In each panel, the first row represents growth after receiving the specified Student Achievement rating. The subsequent rows represent other test-related outcomes of interest.

While I neither individually examine the behaviors of an individual school and its

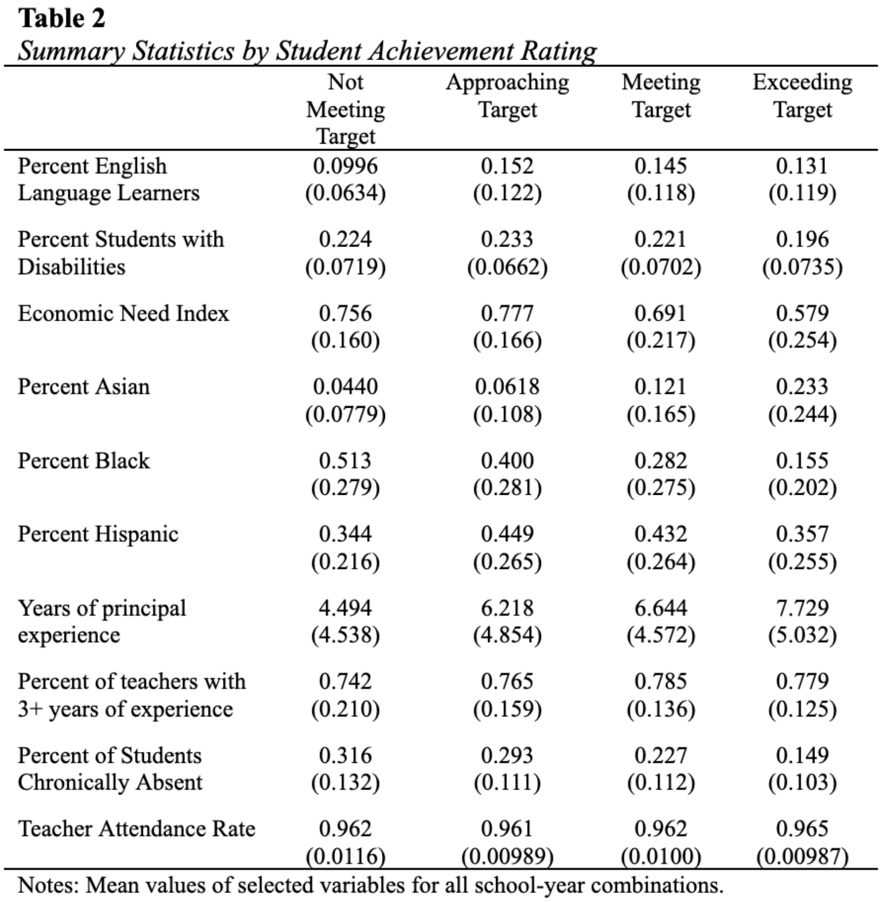

Table 2

Summary Statistics by Student Achievement Rating

	Not Meeting Target	Approaching Target	Meeting Target	Exceeding Target
Percent English Language Learners	0.0996 (0.0634)	0.152 (0.122)	0.145 (0.118)	0.131 (0.119)
Percent Students with Disabilities	0.224 (0.0719)	0.233 (0.0662)	0.221 (0.0702)	0.196 (0.0735)
Economic Need Index	0.756 (0.160)	0.777 (0.166)	0.691 (0.217)	0.579 (0.254)
Percent Asian	0.0440 (0.0779)	0.0618 (0.108)	0.121 (0.165)	0.233 (0.244)
Percent Black	0.513 (0.279)	0.400 (0.281)	0.282 (0.275)	0.155 (0.202)
Percent Hispanic	0.344 (0.216)	0.449 (0.265)	0.432 (0.264)	0.357 (0.255)
Years of principal experience	4.494 (4.538)	6.218 (4.854)	6.644 (4.572)	7.729 (5.032)
Percent of teachers with 3+ years of experience	0.742 (0.210)	0.765 (0.159)	0.785 (0.136)	0.779 (0.125)
Percent of Students Chronically Absent	0.316 (0.132)	0.293 (0.111)	0.227 (0.112)	0.149 (0.103)
Teacher Attendance Rate	0.962 (0.0116)	0.961 (0.00989)	0.962 (0.0100)	0.965 (0.00987)

Notes: Mean values of selected variables for all school-year combinations.

leaders, nor track individual students in and out of these levels, school-level measures of proficiency across years can be a good measure for student performance and a proxy for administrator behavior. New York State also converts student exam scores to a 1 to 4 scale to indicate level of proficiency for each student; category 1 is the lowest, category 2 follows, and categories 3 and 4 are each considered proficient. As schools may be interested in improving subsets of student scores, I construct growth metrics for numbers of students in the following categories: 1 (lowest), category 2 ("bubble") and category 3 / 4 (proficient). The city-defined definitions for each of these categories is presented in Table 4.

The importance of the signs of these metrics is worth discussing specifically as they are not uniformly interpreted across categories; a negative "growth" in the lowest category, for example, means a school had less students in the lowest category (in a given grade-subject) than in the prior year — what most would consider a good thing, despite the negative numeric change. A positive growth in the proficient category, though, means a school has more students achieving proficiency (again in a given grade-subject) than in the prior year — also a good thing.

In Table 5, I present four possibilities for various combinations of signs of three performance category outcome measures at four hypothetical schools. Schools A and B

Table 3

Outcome Measures by Student Achievement Rating

	Student Achievement Rating			
	Not Meeting Target	Approaching Target	Meeting Target	Exceeding Target
Math				
Score Growth (in points)	2.007 (7.375)	1.014 (8.906)	0.0444 (8.630)	-0.0921 (7.889)
Change in lowest score category (in pp)	-5.124 (13.85)	-2.387 (13.61)	-0.123 (11.79)	0.526 (8.892)
Change in "bubble" category (in pp)	0.916 (10.64)	-0.175 (10.70)	-1.163 (9.461)	-0.965 (8.469)
Change in proficiency (in pp)	4.208 (10.32)	2.561 (10.54)	1.285 (11.03)	0.439 (10.74)
English Language Arts (ELA)				
Score Growth (in points; $\mu = 406$)	1.279 (8.630)	2.587 (8.425)	2.428 (7.900)	1.845 (7.359)
Change in lowest score category (in pp)	-1.942 (14.46)	-2.988 (12.65)	-2.144 (10.62)	-0.901 (7.993)
Change in "bubble" category (in pp)	-1.264 (11.27)	-0.691 (10.60)	-1.532 (9.451)	-1.846 (8.591)
Change in proficiency (in pp)	3.206 (11.99)	3.679 (10.49)	3.676 (10.78)	2.747 (10.65)

Notes: The first row in each panel represents raw score growth in test points between consecutive years. Mean score ≈ 300; sd ≈ 15 for both exams, corrected for between year differences in scaling. Rows 2–4 in each panel are measured in percentage point change in number of students in listed categories.

Table 4

Scoring Levels and Distribution for New York City Elementary Exams

	Description	% ELA	% Math
Level 1	Students performing at this level are well below proficient in standards for their grade. They demonstrate knowledge, skills, and practices that are considered insufficient for the expectations at this grade.	24.18%	29.75%
Level 2	Students performing at this level are below proficient in standards for their grade. They demonstrate knowledge, skills, and practices that are considered partial but insufficient for the expectations at this grade.	32.31%	28.04%
Level 3	Students performing at this level are proficient in standards for their grade. They demonstrate knowledge, skills, and practices that are considered sufficient for the expectations at this grade	27.04%	20.51%
Level 4	Students performing at this level excel in standards for their grade. They demonstrate knowledge, skills, and practices that are considered more than sufficient for the expectations at this grade.	16.46%	20.19%

Note: Descriptions from NYCDOE (2019). Percentages are weighted averages; grades 3-8, 2014-2019.

present the most obvious interpretations. At school A, the proportion of students in the lowest category increases, while the number of proficient students decreases. This is a school that is not showing improvement, regardless of the difference in level 2 students. At school B, the lowest category decreases while the proficient category increases. This is a school that is clearly improving; whether the students are moving out of the lowest category into level 2 (or "bubble") or proficient category is certainly important, but with this current data we have no way of knowing if that's the case. Schools C and D are slightly more complicated; in school C, we see increases in both the lowest and proficient category, suggesting that students are being pulled from the bubble to both extremes, suggesting heterogeneity by prior performance level. Conversely at school D students are leaving both the lowest and proficient category, congregating at the bubble, again suggesting heterogeneity, although with notably different results. As students clump at the bubble category, this could indicate School D is increasing scores of its lowest students while suppressing proficiency.

EMPIRICAL STRATEGY

To examine the causal impact of a particular categorical rating, I echo approaches from previous scholarship on school accountability grades, namely those conducted on New York City data (i.e., Rockoff & Turner, 2010; Winters & Cowen, 2012). Ideally, to replicate a fully controlled trial, I would examine the same school under two different conditions; for example, one in which they receive a label of "Approaching Target," and one in which they receive a label of "Meeting Target." For obvious reasons, this is not possible; schools only receive one score/rating each year, and schools have already received these labels. Further, it would be difficult and unethical to randomly assign something so important in

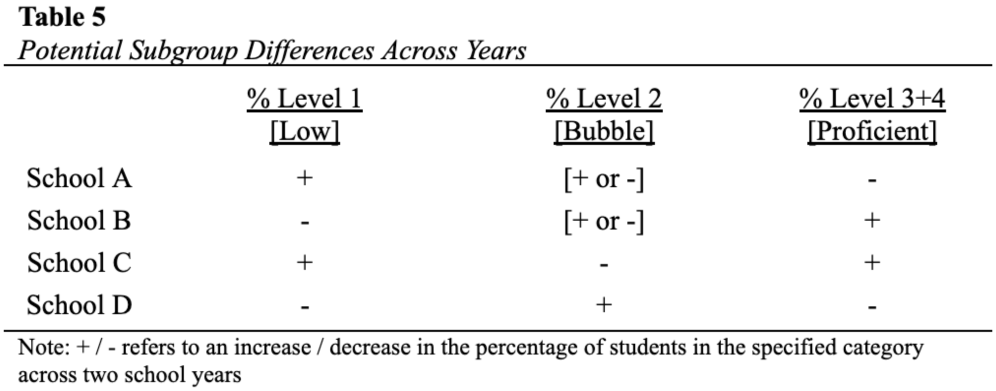

Table 5
Potential Subgroup Differences Across Years

	% Level 1 [Low]	% Level 2 [Bubble]	% Level 3+4 [Proficient]
School A	+	[+ or -]	-
School B	-	[+ or -]	+
School C	+	-	+
School D	-	+	-

Note: + / - refers to an increase / decrease in the percentage of students in the specified category across two school years

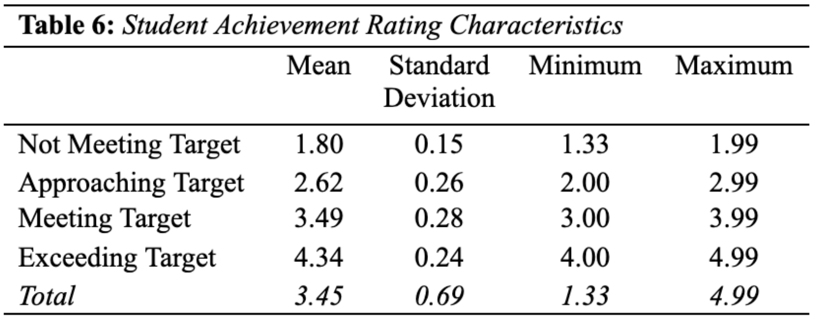

Table 6: *Student Achievement Rating Characteristics*

	Mean	Standard Deviation	Minimum	Maximum
Not Meeting Target	1.80	0.15	1.33	1.99
Approaching Target	2.62	0.26	2.00	2.99
Meeting Target	3.49	0.28	3.00	3.99
Exceeding Target	4.34	0.24	4.00	4.99
Total	*3.45*	*0.69*	*1.33*	*4.99*

assignment inherent in a carefully controlled experiment. However, there are ways, with some assumptions, to create (almost) as-good-as randomization.

In a regression discontinuity ("RD") approach, the underlying notion is that observations close to the left and right of any given cutoff are essentially statistically identical based on their close proximity on the assignment variable which determines their categorical label, which Lee and Lemieux (2010) described as the "Local Randomization" assumption (p. 295). While in a controlled trial, treatment is assigned based on strict randomization, here treatment is assigned to those close to the cut point in what is assumed to be a near-random way.

This approach leverages the discontinuous treatment assignment mechanism built into the School Quality Reports. Table 6 and Figure 1 demonstrate this assignment mechanism clearly. The Student Achievement Score (referred to from here on as "score") is generated on a 1 to 5 continuous scale, and depending on this score, schools are assigned one of four Student Achievement Ratings (referred to from here on as "rating"). Note the lack of overlap between the rating categories; the maximum for each category is exactly .01 below the whole number threshold for the next category. Plot point sizes in Figure 1 highlight the cluster in the middle two ratings, and the few schools gathered on the extremes.

In this case, the continuous student achievement score concretely determines the categorical rating, but schools close to the predefined cut point (for example, 2.99 vs. 3.00) are so close that they are essentially randomly distributed on either side of the cut, meaning the difference between receiving a rating of "Approaching Target" and "Meeting Target" is essentially random. Thus, creating localized regression models around the cut point can estimate the causal impact of treatment; in this case, treatment is defined by receiving a particular rating relative to another.

The models implemented are of the following form:

$$E_{(y+1)sgc}-E_{ysgc}=\beta_0+\gamma T_{ys}+\beta_i(S_{ys})+\beta_j T_{ys}(S_{ys})+\beta_k X_{ys}+\mu_y+\varepsilon_{ys} \quad (1)$$

where $E_{(y+1)sgc}$ represents an outcome metric for year y+1, or one year following the assignment of a Student Achievement rating, in school *s*, grade *g*, and content area *c* (either Math or English), while E_{ysgc} represents the same metric for the year the rating was assigned. Together, the left side of equation (1) represents growth in a specified outcome. T_{ys} represents a dummy variable for receiving a lower rating at a given school in a given year for a specified cut point between two ratings; for example, T_{ys}=1 if a school received an "Approaching" rating and T_{ys}=0 if the school received a "Meeting," if examining the "Approaching vs. Meeting" cut point.[1] S_{ys} is a vector which represents the continuous Achievement Score for a given year/school and its quadratic term, $T_{ys}(S_{ys})$ represents a

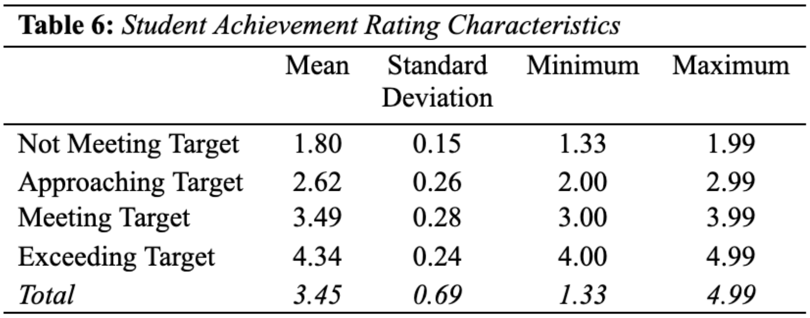

Table 6: *Student Achievement Rating Characteristics*

	Mean	Standard Deviation	Minimum	Maximum
Not Meeting Target	1.80	0.15	1.33	1.99
Approaching Target	2.62	0.26	2.00	2.99
Meeting Target	3.49	0.28	3.00	3.99
Exceeding Target	4.34	0.24	4.00	4.99
Total	*3.45*	*0.69*	*1.33*	*4.99*

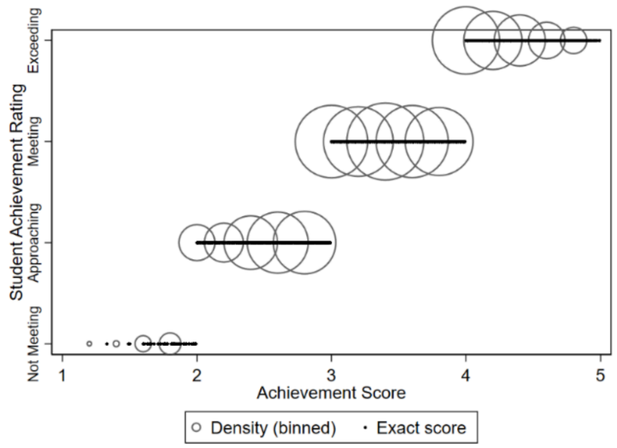

Figure 1: *Achievement Score and Achievement Rating*

vector of interactions between treatment and Achievement Score, allowing for differing coefficients on either side of the cut point. X_{ys} represents a vector of school-year covariates and μ_y is a year fixed-effect; these terms are added in later models. Lastly, is an idiosyncratic error term. Finally, γ is the parameter of interest, and given the assumptions of the regression discontinuity design, represents the causal impact of being just assigned a particular label relative to one Student Achievement rating higher.

ESTABLISHING THE VALIDITY OF THE RD IDENTIFICATION STRATEGY

An important preliminary check for internal validity is to assess the possibility of manipulation at the cut point (Lee & Lemieux, 2010). Because the assignment variable — achievement score — is assumed to be continuous, there should be little evidence of significant jumps anywhere along the spectrum, but specifically not at the cut points. If there were to be a jump at the cut point, it might signify and unobservable manipulation to the assignment variable at the cut point, violating a core assumption of the regression discontinuity approach, and thus rendering our analysis inaccurate. To see potential evidence visually, a histogram is the most appropriate choice, and presented in Figure 2. There are three distinct areas, defined by the two cut points of the achievement score. Schools that were designated "Not Meeting" or "Approaching" the target have been classified as "Below Target", while the "Meeting Target" and "Exceeding Target" labels are directly from the achievement rating. The cut point lines are presented in red for convenience as well.

Upon simple visual analysis, while the cut at 4 seems to not be an issue, there does appear to be a small jump from "Below Target" to "Meeting Target" where Achievement Score equals 3. This is potentially statistically problematic; if there is manipulation happening to move scores from immediately below the cut point to immediately above, this would violate a core assumption of RD and render any inference based on the RD inaccurate. However, considering the nature of the School Quality reports and their underlying statistics, it would be difficult for any real manipulation to take place. First, it is impossible to predict and manipulate the wide range of scores that will eventually be used to calculate an achievement score and therefore rating. While administrators may have

[1] While it may seem more intuitive to code this in the reverse, I chose to code treatment in this way as to more intuitively interpret the effect of the lower score relative to the higher score, given the theory that schools may be motivated in particular by a lower score.

had access to their scores before they were finalized, it would still have been difficult to manipulate scores after tests were concluded and shift scores in one direction or the other. Finally, as I address more fully in my limitations section, any manipulation of this sort may actually lead to underestimating the effects at that cut point, given the results presented below.

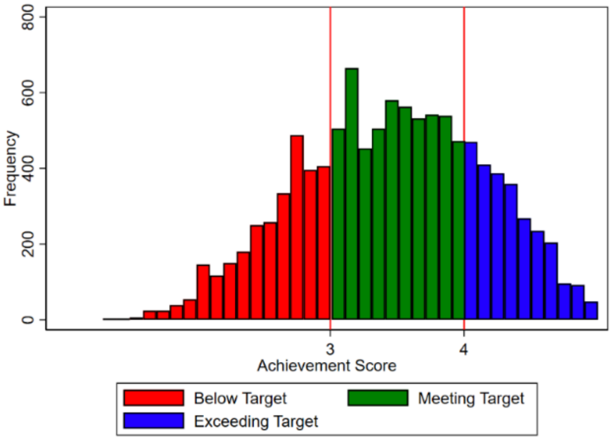

Figure 2: *Histogram of achievement ratings*

As a final check of the regression discontinuity assumptions, I present a parallel analysis using a covariate as an outcome measure as suggested by Lee and Lemieux (2010) in Table 7. If there was notable manipulation or some other difference in the groups on either side of the cut points, this difference may be revealed by differences in covariates, which, based on the assumptions of regression discontinuity, should be similar across both sides of the cut point. Table 5 presents results from RDRobust (Calonico et al., 2017) for all school-level covariates used later in analysis at both cut points: percent of English language learners, percent of students in special education, economic need index (calculated by NYCDOE to represent schoolwide economic need), a variety of race percentages, principal and teacher experience, and student and teacher attendance. Columns 1–3 indicate balance at the Approaching vs. Meeting cut point, while columns 4–6 indicate balance at the Meeting vs. Exceeding cut point. The first columns (1 and 4) utilize data-driven bandwidth selections, while the remaining use a predefined smaller and larger bandwidth. There should not be any significant results in these tests; if there were, it would signify a discontinuity in one of our covariates, violating the local randomization assumption, and would suggest that there was a statistical difference between the two groups close to the cut point. As suspected, the estimates are small in magnitude, indicating little difference, and only a small handful are significant, and only at the p= .05 level. Indeed, given the large number of statistical test results being presented in this table (60), it is not surprising that some may appear significant. This helps reinforce (yet not necessarily fully confirm) the original assumption of local randomization around the two cut points.

RESULTS

In the following section, I discuss the findings of the regression discontinuity design. First, I explain visual differences at the cut point using binned plots with local linear specifications mapped on for ease of interpretation, finding that being just assigned (i.e., assignment based on being just past the cut point) a rating of "Approaching" has a positive impact on some math-related outcomes. Those results do not appear to be

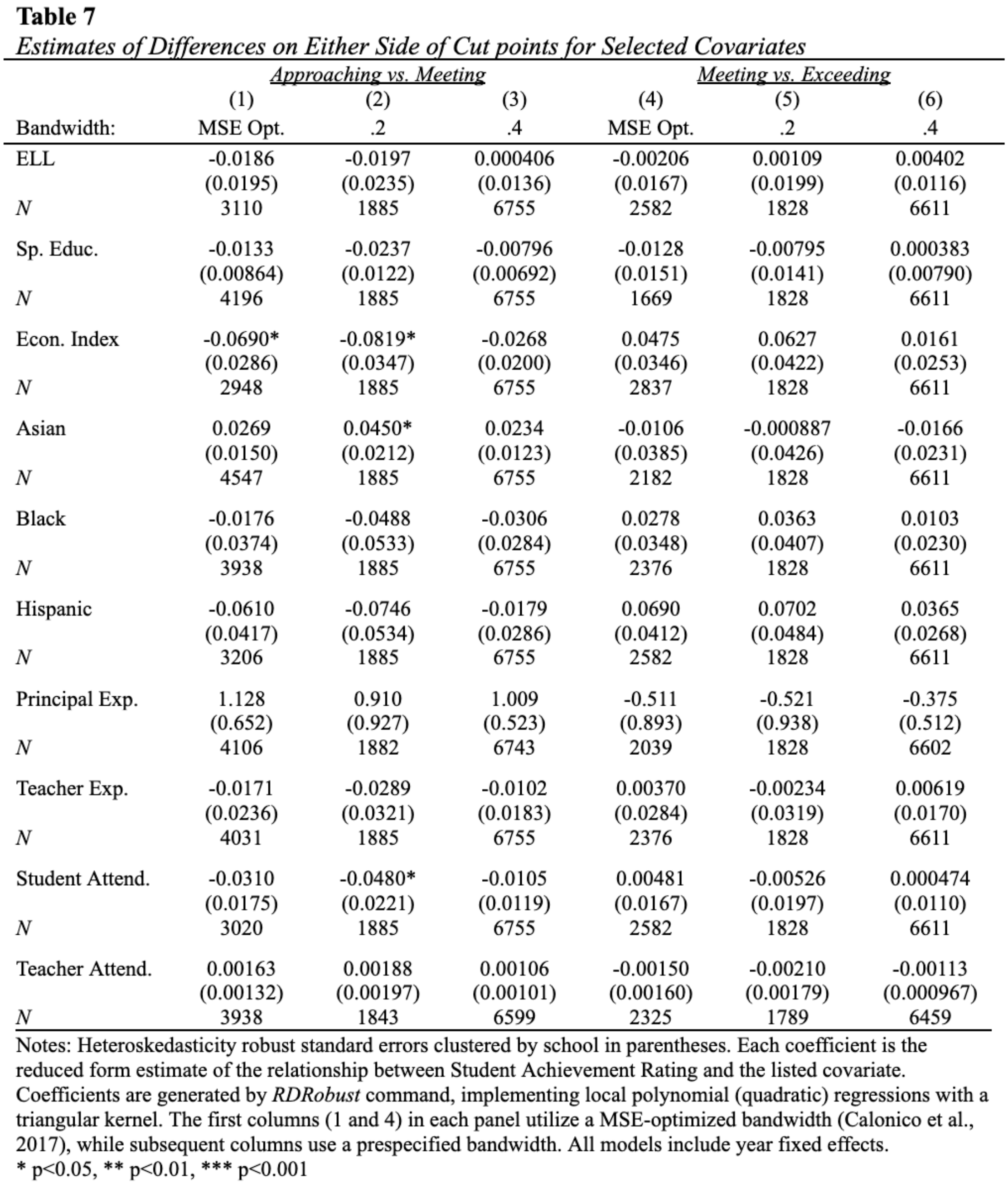

Table 7

Estimates of Differences on Either Side of Cut points for Selected Covariates

	Approaching vs. Meeting			*Meeting vs. Exceeding*		
	(1)	(2)	(3)	(4)	(5)	(6)
Bandwidth:	MSE Opt.	.2	.4	MSE Opt.	.2	.4
ELL	-0.0186	-0.0197	0.000406	-0.00206	0.00109	0.00402
	(0.0195)	(0.0235)	(0.0136)	(0.0167)	(0.0199)	(0.0116)
N	3110	1885	6755	2582	1828	6611
Sp. Educ.	-0.0133	-0.0237	-0.00796	-0.0128	-0.00795	0.000383
	(0.00864)	(0.0122)	(0.00692)	(0.0151)	(0.0141)	(0.00790)
N	4196	1885	6755	1669	1828	6611
Econ. Index	-0.0690*	-0.0819*	-0.0268	0.0475	0.0627	0.0161
	(0.0286)	(0.0347)	(0.0200)	(0.0346)	(0.0422)	(0.0253)
N	2948	1885	6755	2837	1828	6611
Asian	0.0269	0.0450*	0.0234	-0.0106	-0.000887	-0.0166
	(0.0150)	(0.0212)	(0.0123)	(0.0385)	(0.0426)	(0.0231)
N	4547	1885	6755	2182	1828	6611
Black	-0.0176	-0.0488	-0.0306	0.0278	0.0363	0.0103
	(0.0374)	(0.0533)	(0.0284)	(0.0348)	(0.0407)	(0.0230)
N	3938	1885	6755	2376	1828	6611
Hispanic	-0.0610	-0.0746	-0.0179	0.0690	0.0702	0.0365
	(0.0417)	(0.0534)	(0.0286)	(0.0412)	(0.0484)	(0.0268)
N	3206	1885	6755	2582	1828	6611
Principal Exp.	1.128	0.910	1.009	-0.511	-0.521	-0.375
	(0.652)	(0.927)	(0.523)	(0.893)	(0.938)	(0.512)
N	4106	1882	6743	2039	1828	6602
Teacher Exp.	-0.0171	-0.0289	-0.0102	0.00370	-0.00234	0.00619
	(0.0236)	(0.0321)	(0.0183)	(0.0284)	(0.0319)	(0.0170)
N	4031	1885	6755	2376	1828	6611
Student Attend.	-0.0310	-0.0480*	-0.0105	0.00481	-0.00526	0.000474
	(0.0175)	(0.0221)	(0.0119)	(0.0167)	(0.0197)	(0.0110)
N	3020	1885	6755	2582	1828	6611
Teacher Attend.	0.00163	0.00188	0.00106	-0.00150	-0.00210	-0.00113
	(0.00132)	(0.00197)	(0.00101)	(0.00160)	(0.00179)	(0.000967)
N	3938	1843	6599	2325	1789	6459

Notes: Heteroskedasticity robust standard errors clustered by school in parentheses. Each coefficient is the reduced form estimate of the relationship between Student Achievement Rating and the listed covariate. Coefficients are generated by *RDRobust* command, implementing local polynomial (quadratic) regressions with a triangular kernel. The first columns (1 and 4) in each panel utilize a MSE-optimized bandwidth (Calonico et al., 2017), while subsequent columns use a prespecified bandwidth. All models include year fixed effects.
* p<0.05, ** p<0.01, *** p<0.001

present for ELA outcomes, nor for being just assigned the "Meeting" rating. Then, I present numeric estimates that bolster the conclusions evident in the graphical approach.

VISUAL RESULTS

Figure 3 demonstrates the four math outcomes of interest at the "Approaching vs. Meeting" cut point. There are apparent discontinuities in all four of the graphs presented; each graph suggests that just being assigned a rating of "Approaching" improves math-related outcomes. First, in graph A, it appears that there is increased score growth below the cut point, suggesting being assigned a lower rating causes a positive impact on math score growth for schools just below the cut point when compared to their similar peers just above the cut point. Similarly, graph D suggests that when schools just below the cut point are assigned a lower rating, they have a greater increase in their percentage of students in the overall category of proficiency in comparison to their peers just above the cut point.

Graph B (figure 3) shows an opposite visual pattern, as in being just at the lower rating suggests a negative impact relative to schools just above the cut point. However, this may be consistent evidence of improvement regardless of differing signs. Graph B suggests that schools just below the cut point have a larger decrease in the number of students in the lowest category, which would generally be interpreted as overall improvement. Similar to graph A, the positive difference in graphs C and D suggests that schools just below the cut point both increased the number of students in the bubble category (level 2) as well as students in the Proficient category, when compared to schools just above the cut point. However, the evidence in graph C is the least conclusive visually.

Figure 4 demonstrates the four ELA outcomes of interest at the “Approaching vs. Meeting” cut point. These figures appear to be less conclusive than their counterparts in Figure 3 in that the discontinuities are less pronounced. Still, in graphs A and B there appears to be evidence of improvement for schools in both score growth and movement of students out of the lowest performance category just below the cut point receiving a rating of Approaching in comparison to similar schools just above the cut point.

Figure 5 presents the same math-related outcomes as Figure 3 but shifts the perspective to the Meeting vs. Exceeding cut point. In comparison to Figure 3, there do not appear to be as notable discontinuities here. A weak argument may be made that graphs A and B present discontinuities; interestingly, the signs of these discontinuities are opposite of the prior evidence, perhaps suggesting that schools just above the higher cut point improved as a function of the higher rating. Still, that claim would need to be confirmed by additional evidence presented below. Figure 6 presents the same ELA-related outcomes as Figure 4, but at the Meeting vs. Exceeding cut point. Similar to math-related results in Figure 5, there may be a weak argument for differences present in graphs A and B, again with different signs than at the prior cut points, but that evidence should be interpreted cautiously and only if verified by additional analyses that follow.

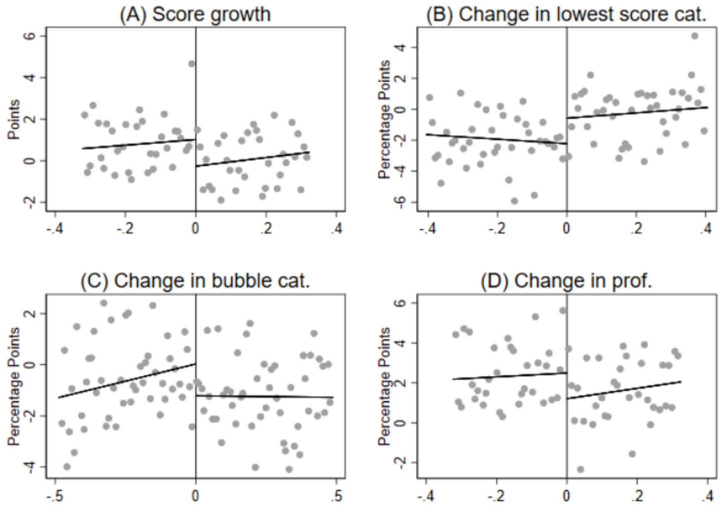

Figure 3: Reduced form impact of Approaching vs. Meeting on math outcomes

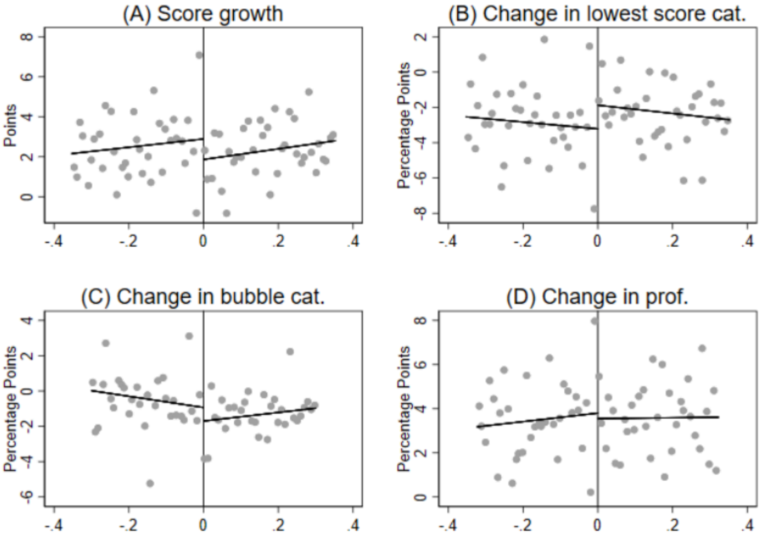

Figure 4: Reduced form impact of Approaching vs. Meeting on ELA outcomes

SENSITIVITY TO BANDWIDTH AND FUNCTIONAL FORM

The results from regression discontinuity designs can be sensitive to the choice of bandwidth; in other words, depending on how one defines the range of scores for which the schools are essentially similar, the analysis may be biased or imprecise. Indeed, the choice of bandwidth is a limitation to the regression discontinuity approach; choosing to widen the bandwidth to improve precision (by including more data points) inherently adds bias to point estimates (Lee & Lemieux, 2010). To address this, Lee and Lemieux (2010) suggest exploring the sensitivity of results with a variety of bandwidths and a variety of orders of polynomial, although recent evidence (Gelman & Imbens, 2019) suggest limiting higher order polynomials to second order (quadratic). Ideally, these differing specifications should provide similar estimates of the treatment in both magnitude and sign, revealing a rough approximation of the "true" causal effect.

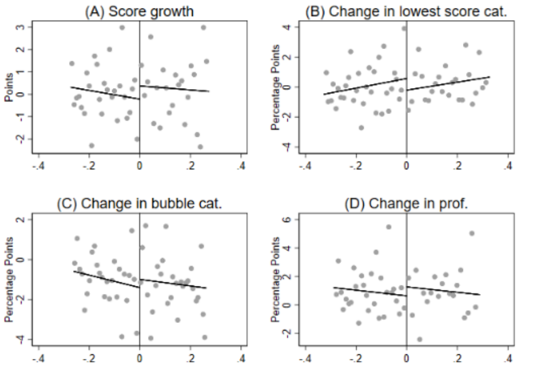

Figure 5: Reduced form impact of Meeting vs. Exceeding on math outcomes

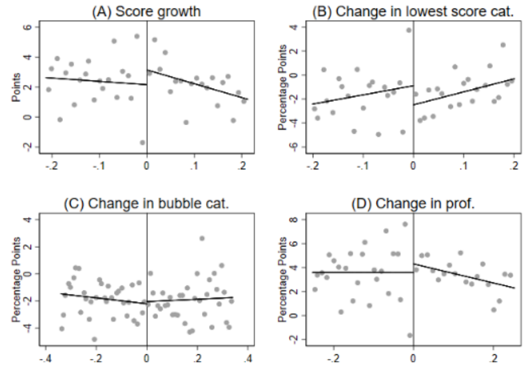

Figure 6: Reduced form impact of Meeting vs. Exceeding on ELA outcomes

I present estimates for a variety of bandwidths and polynomial specifications in Tables 8 and 9. Presented in each cell is , the coefficient on the "treatment," which is defined as being either just below or above the cut point; in Table 6, treatment is just receiving an "Approaching" rating, while in Table 7, treatment is just receiving a "Meeting" rating.

The estimates presented, then, are the causal impact of being just below the cut point, in comparison to otherwise similar schools, for each of the given outcome measures. While there is no expectation that the estimates should be exactly the same across bandwidths and polynomial orders, similarities between the variety of specifications, as well as the visual evidence presented in the earlier figures, may provide a preponderance of evidence of a true causal impact.

The estimates in table 8 are split by math and ELA results in the top and bottom panels, respectively. Rows 1, 2, and 4 each suggest similar results to the visual evidence provided in figure 3; while the magnitudes are not identical, the fact that multiple bandwidths and specifications lead to statistically significant increases in score growth should be taken together as a preponderance of evidence. There is ample evidence, then, that being just below the cut point and receiving an Approaching rating causes increases in math score growth, decreases in the percentage of students in the lowest proficiency category, and increases in the percentage of students scoring proficient, relative to similar schools just beyond the cut point receiving a Meeting rating. There does not seem to be substantial evidence that there is an impact on the "bubble" student category. Conversely, in examining the bottom panel for ELA estimates, there does not appear to be statistically significant evidence of differences at the Approaching vs. Meeting rating, the exception being suggestive evidence of differences in score growth. This is not surprising considering the less substantial visual evidence presented in figure 4.

In Table 9, I present a similar set of estimates for the higher Meeting vs. Exceeding cut point; the top panel is Math outcomes and the bottom panel for ELA. These results confirm the visual evidence in figures 5 and 6; there does not appear to be much evidence of an impact of just receiving the Meeting rating, with the exception being row 2 in the lower panel for ELA. These results weakly suggest that just being rated Meeting may cause an increase in the percentage of students in the lowest category of ELA performance relative to schools just above the cut point receiving a rating of Exceeding.

ROBUSTNESS CHECKS AND ADDITIONAL LIMITATIONS

One mechanism to potentially increase precision is the addition of covariates. The addition of covariates should strictly not shift the magnitude or direction of the results, only the precision, and if the addition of covariates does in fact shift the results the implication is that there was either a manipulation issue or a specification issue (Lee & Lemieux, 2010). In Table 10, I present results for both Math and English scores at the Approaching vs. Meeting cut point with a collection of school-level covariates added (see Table 2 for the comprehensive list of covariates). I choose to omit further results for Meeting vs. Exceeding as the preferred specification was not statistically significant. The results in Table 10 resemble the results in Table 8, as they should, including the relatively weak, suggestive evidence that there may be an impact on ELA score growth.

There are some additional considerations and limitations that must be addressed. The first is regarding the strength and significance of the conclusions; while there is ample evidence that receiving a lower label increased future growth in test scores, particularly for math exams, the results are by no means wholly conclusive. Because there are only three sets of paired years data, with more data the conclusions would be more robust and perhaps more precise.

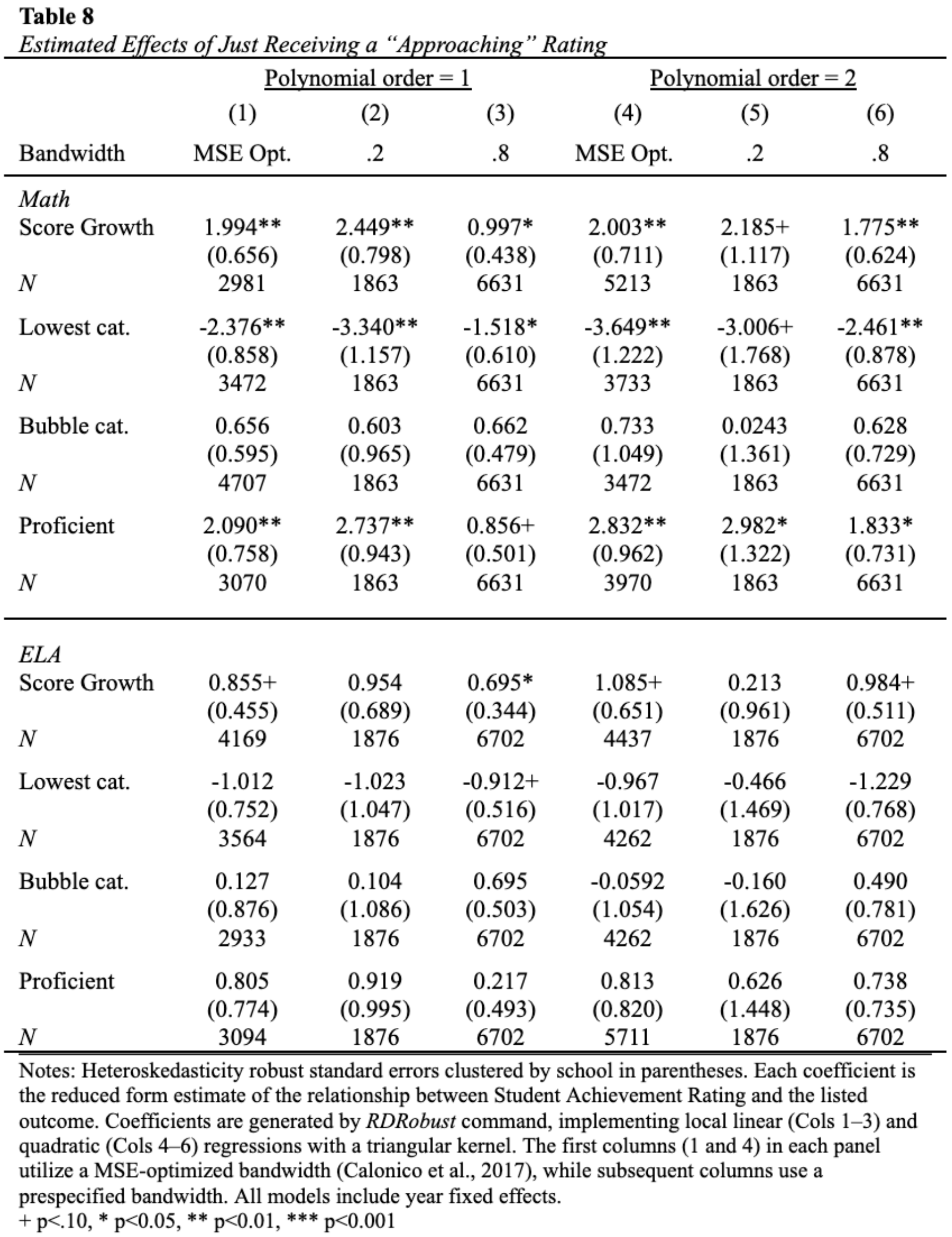

Table 8

Estimated Effects of Just Receiving a "Approaching" Rating

	Polynomial order = 1			Polynomial order = 2		
	(1)	(2)	(3)	(4)	(5)	(6)
Bandwidth	MSE Opt.	.2	.8	MSE Opt.	.2	.8
Math						
Score Growth	1.994**	2.449**	0.997*	2.003**	2.185+	1.775**
	(0.656)	(0.798)	(0.438)	(0.711)	(1.117)	(0.624)
N	2981	1863	6631	5213	1863	6631
Lowest cat.	-2.376**	-3.340**	-1.518*	-3.649**	-3.006+	-2.461**
	(0.858)	(1.157)	(0.610)	(1.222)	(1.768)	(0.878)
N	3472	1863	6631	3733	1863	6631
Bubble cat.	0.656	0.603	0.662	0.733	0.0243	0.628
	(0.595)	(0.965)	(0.479)	(1.049)	(1.361)	(0.729)
N	4707	1863	6631	3472	1863	6631
Proficient	2.090**	2.737**	0.856+	2.832**	2.982*	1.833*
	(0.758)	(0.943)	(0.501)	(0.962)	(1.322)	(0.731)
N	3070	1863	6631	3970	1863	6631
ELA						
Score Growth	0.855+	0.954	0.695*	1.085+	0.213	0.984+
	(0.455)	(0.689)	(0.344)	(0.651)	(0.961)	(0.511)
N	4169	1876	6702	4437	1876	6702
Lowest cat.	-1.012	-1.023	-0.912+	-0.967	-0.466	-1.229
	(0.752)	(1.047)	(0.516)	(1.017)	(1.469)	(0.768)
N	3564	1876	6702	4262	1876	6702
Bubble cat.	0.127	0.104	0.695	-0.0592	-0.160	0.490
	(0.876)	(1.086)	(0.503)	(1.054)	(1.626)	(0.781)
N	2933	1876	6702	4262	1876	6702
Proficient	0.805	0.919	0.217	0.813	0.626	0.738
	(0.774)	(0.995)	(0.493)	(0.820)	(1.448)	(0.735)
N	3094	1876	6702	5711	1876	6702

Notes: Heteroskedasticity robust standard errors clustered by school in parentheses. Each coefficient is the reduced form estimate of the relationship between Student Achievement Rating and the listed outcome. Coefficients are generated by *RDRobust* command, implementing local linear (Cols 1–3) and quadratic (Cols 4–6) regressions with a triangular kernel. The first columns (1 and 4) in each panel utilize a MSE-optimized bandwidth (Calonico et al., 2017), while subsequent columns use a prespecified bandwidth. All models include year fixed effects.
+ p<.10, * p<0.05, ** p<0.01, *** p<0.001

The second and more important consideration is the issue of potential manipulation at the Approaching vs. Meeting cut. Visually, the histogram appears as if there may be a "jump" at the cut point, which ideally should not be the case; there will of course be some idiosyncratic lumpiness throughout the distribution but seeing a particular "jump" at the cut point suggests there may be schools manipulating their scores right at the cut point to move from just below to just above. Further, there is a policy-related chance that manipulation was happening. Because schools had access to the data used to calculate the Student Achievement score ahead of the report's publication, they could have theoretically calculated their (future) Student Achievement scores relative to the cut point prior to their official label assignment. Knowing their label assignment, they could have attempted to interfere with students' testing or attempt to manually edit their data to account for the potential lower rating.

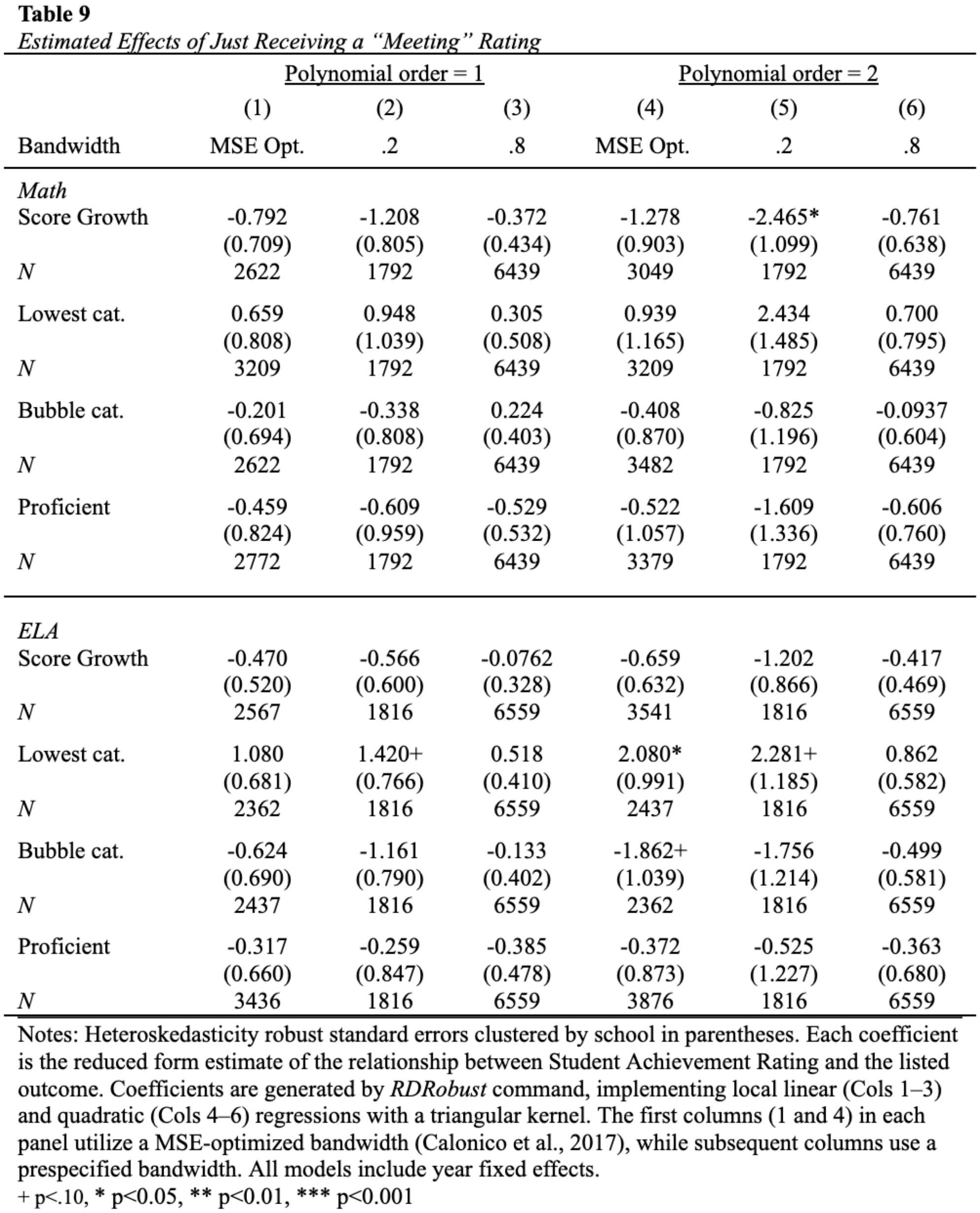

Table 9

Estimated Effects of Just Receiving a "Meeting" Rating

	Polynomial order = 1			Polynomial order = 2		
	(1)	(2)	(3)	(4)	(5)	(6)
Bandwidth	MSE Opt.	.2	.8	MSE Opt.	.2	.8
Math						
Score Growth	-0.792	-1.208	-0.372	-1.278	-2.465*	-0.761
	(0.709)	(0.805)	(0.434)	(0.903)	(1.099)	(0.638)
N	2622	1792	6439	3049	1792	6439
Lowest cat.	0.659	0.948	0.305	0.939	2.434	0.700
	(0.808)	(1.039)	(0.508)	(1.165)	(1.485)	(0.795)
N	3209	1792	6439	3209	1792	6439
Bubble cat.	-0.201	-0.338	0.224	-0.408	-0.825	-0.0937
	(0.694)	(0.808)	(0.403)	(0.870)	(1.196)	(0.604)
N	2622	1792	6439	3482	1792	6439
Proficient	-0.459	-0.609	-0.529	-0.522	-1.609	-0.606
	(0.824)	(0.959)	(0.532)	(1.057)	(1.336)	(0.760)
N	2772	1792	6439	3379	1792	6439
ELA						
Score Growth	-0.470	-0.566	-0.0762	-0.659	-1.202	-0.417
	(0.520)	(0.600)	(0.328)	(0.632)	(0.866)	(0.469)
N	2567	1816	6559	3541	1816	6559
Lowest cat.	1.080	1.420+	0.518	2.080*	2.281+	0.862
	(0.681)	(0.766)	(0.410)	(0.991)	(1.185)	(0.582)
N	2362	1816	6559	2437	1816	6559
Bubble cat.	-0.624	-1.161	-0.133	-1.862+	-1.756	-0.499
	(0.690)	(0.790)	(0.402)	(1.039)	(1.214)	(0.581)
N	2437	1816	6559	2362	1816	6559
Proficient	-0.317	-0.259	-0.385	-0.372	-0.525	-0.363
	(0.660)	(0.847)	(0.478)	(0.873)	(1.227)	(0.680)
N	3436	1816	6559	3876	1816	6559

Notes: Heteroskedasticity robust standard errors clustered by school in parentheses. Each coefficient is the reduced form estimate of the relationship between Student Achievement Rating and the listed outcome. Coefficients are generated by *RDRobust* command, implementing local linear (Cols 1–3) and quadratic (Cols 4–6) regressions with a triangular kernel. The first columns (1 and 4) in each panel utilize a MSE-optimized bandwidth (Calonico et al., 2017), while subsequent columns use a prespecified bandwidth. All models include year fixed effects.
+ p<.10, * p<0.05, ** p<0.01, *** p<0.001

Manipulating test scores to move from one side of the cut point to another seems like an unlikely and unwieldy task for an administrator, however. The administrator would need both a significant amount of time to develop the calculations, as well as plausible rationale to manually adjust scores. Further, a savvier administrator would likely manipulate their grade to be higher. Assuming a savvy administrator would also lead to higher test scores, this manipulation would in fact bias results in the opposite direction; those schools that were manipulated to be just beyond the cut point should see more growth in comparison their otherwise similar schools just below the cut point. Any potential manipulation, then, would suggest the results are actually larger than presented here. Adding further analysis to those specific cases immediately past the cut certainly would be beneficial in the future.

An additional potential issue to be considered is that of schools sliding back and forth across the cut point. Because test scores are included in subsequent years' Student

Achievement scores, there is a possibility that the gains described above of just receiving a rating of Approaching will push a given school into the Meeting category the following year, making the school a control school. While the year fixed effect in the model addresses any between-year dependencies by limiting comparisons to within-year, it does not wholly address the issue from an interpretive standpoint. Because the gains are local to the cut point, if schools are simply sliding back and forth across the cut point, the results are far less meaningful, suggesting any impacts are both short-term and immediately reversed.

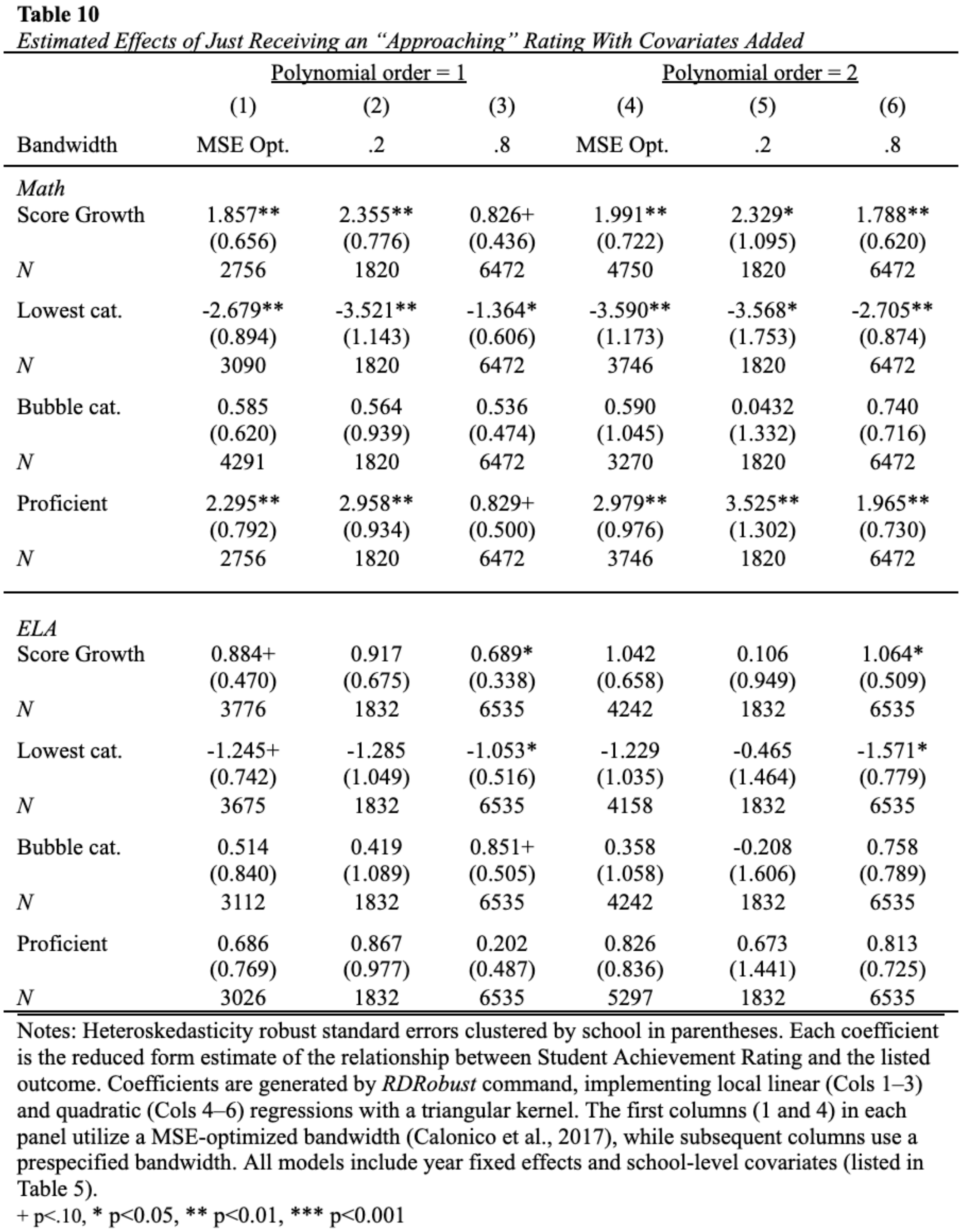

Table 10

Estimated Effects of Just Receiving an "Approaching" Rating With Covariates Added

	Polynomial order = 1			Polynomial order = 2		
	(1)	(2)	(3)	(4)	(5)	(6)
Bandwidth	MSE Opt.	.2	.8	MSE Opt.	.2	.8
Math						
Score Growth	1.857**	2.355**	0.826+	1.991**	2.329*	1.788**
	(0.656)	(0.776)	(0.436)	(0.722)	(1.095)	(0.620)
N	2756	1820	6472	4750	1820	6472
Lowest cat.	-2.679**	-3.521**	-1.364*	-3.590**	-3.568*	-2.705**
	(0.894)	(1.143)	(0.606)	(1.173)	(1.753)	(0.874)
N	3090	1820	6472	3746	1820	6472
Bubble cat.	0.585	0.564	0.536	0.590	0.0432	0.740
	(0.620)	(0.939)	(0.474)	(1.045)	(1.332)	(0.716)
N	4291	1820	6472	3270	1820	6472
Proficient	2.295**	2.958**	0.829+	2.979**	3.525**	1.965**
	(0.792)	(0.934)	(0.500)	(0.976)	(1.302)	(0.730)
N	2756	1820	6472	3746	1820	6472
ELA						
Score Growth	0.884+	0.917	0.689*	1.042	0.106	1.064*
	(0.470)	(0.675)	(0.338)	(0.658)	(0.949)	(0.509)
N	3776	1832	6535	4242	1832	6535
Lowest cat.	-1.245+	-1.285	-1.053*	-1.229	-0.465	-1.571*
	(0.742)	(1.049)	(0.516)	(1.035)	(1.464)	(0.779)
N	3675	1832	6535	4158	1832	6535
Bubble cat.	0.514	0.419	0.851+	0.358	-0.208	0.758
	(0.840)	(1.089)	(0.505)	(1.058)	(1.606)	(0.789)
N	3112	1832	6535	4242	1832	6535
Proficient	0.686	0.867	0.202	0.826	0.673	0.813
	(0.769)	(0.977)	(0.487)	(0.836)	(1.441)	(0.725)
N	3026	1832	6535	5297	1832	6535

Notes: Heteroskedasticity robust standard errors clustered by school in parentheses. Each coefficient is the reduced form estimate of the relationship between Student Achievement Rating and the listed outcome. Coefficients are generated by *RDRobust* command, implementing local linear (Cols 1–3) and quadratic (Cols 4–6) regressions with a triangular kernel. The first columns (1 and 4) in each panel utilize a MSE-optimized bandwidth (Calonico et al., 2017), while subsequent columns use a prespecified bandwidth. All models include year fixed effects and school-level covariates (listed in Table 5).

+ p<.10, * p<0.05, ** p<0.01, *** p<0.001

Figure 7 provides a descriptive picture of this potential issue by mapping the treatment/control status of all schools within the preferred bandwidth of year 1 (2015's report card) at the Approaching/Meeting cut point. In the leftmost column are size-weighted markers for schools just below and above the Approaching/Meeting cut point on the 2015 report card. The middle column filters those schools by their rating, if within bandwidth, the following year. Because some schools move out of the bandwidth altogether in the following year, the markers in 2016 do not sum to their respective markers in 2015. Finally, a similar split is demonstrated between 2016 and 2018 (the next analytic year). Of the 157 treatment schools within the analytic bandwidth labeled Approaching in 2015, only 11 moved up to Meeting in 2016 and back again to Approaching in 2018. Similarly, of the 220 "control" schools, only 6 move back to Approaching and subsequently up to Meeting once again, suggesting any problematic sliding back and forth across the cut point is limited.

While the descriptive picture above suggests only a limited impact of "sliding" back and forth across the cut point, there are legitimate policy reasons why the impact might be limited as well. While the measured outcome, test scores, are a part of future Student Achievement scores and ratings, it is not the only measure; there is significant noise in the assignment variable as it is constructed from a variety of metrics (including test scores, attendance, school surveys, etc.).

Last, there may be a concern that prior treatment, including prior year ratings from this system or the prior system, may present an identification issue. For that to be the case, prior years' treatment would need to be endogenously associated with the treatment above and beyond the forcing variable, which is unlikely to be the case. That is to say, even with prior intervention based on prior treatment, which may move schools up (or down) on the continuous forcing variable, there's nothing to suggest that movement would be different immediately surrounding the cut point. While prior treatment may impact some schools, those schools are likely to be distributed across the cut point in a given year, and year fixed-effects ensures only within-year comparisons are estimated.

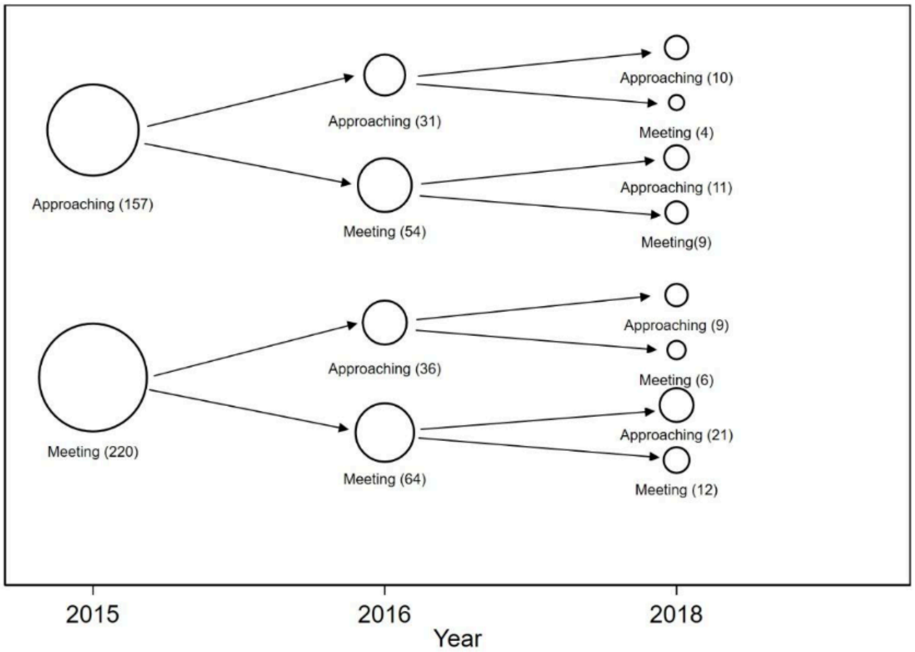

Figure 7: *Within-bandwidth distribution of 2015 sample schools*

DISCUSSION

The results suggest two key causal implications of receiving an "Approaching" Student Achievement rating for schools near the cut point in comparison to schools near the cut point receiving a rating of Meeting. First, schools just receiving an Approaching rating causes greater growth in the following year's math scores, while the evidence is not nearly as strong for ELA. This is consistent with prior evidence of that strict accountability implementation increases in math more dramatically than reading (Hansen et al., 2018).

Second, schools just below the cut point appear to be more effective than otherwise similar schools at moving students out of the lowest category of math score, as well as more effective than otherwise similar schools at moving students into the proficient category of math score. While there is little evidence for the bubble students, this very well may be because schools just below the cut point are moving students both in and out of the bubble, masking any real impact or difference between the two groups of schools despite the progress being made. In conjunction, the evidence I present above suggests that, at least for math, schools and their leaders are responding specifically to the categorical Student Achievement rating above and beyond any information presented by the continuous numeric achievement score.

While the evidence above may indicate a lower categorical Student Achievement rating causes increased math test score growth in the following year for schools near the Approaching/Meeting cut point, it does not address why this happens. There are a variety of potential explanations with policy implications. For example, it may be that for schools just below the cut point there is a differential motivating factor that leads to different tactics leading to increased achievement and/or focus on test scores in the following year relative to their peers, a form of satisficing (Simon, 1956). Alternatively, schools who just barely reach the threshold for a higher score may see this achievement as sufficient relative to their peers and place less emphasis on test scores the following year (another form of satisficing). Again, this is important from a policy perspective because it suggests that schools and their personnel react positively to negative information about their institution, even if that information is comparatively marginal (i.e., puts them just below the cut point). Further, it suggests that despite the availability of the fully continuous Student Achievement score, schools are reacting to the categorical ratings and not numeric score, otherwise there would likely be no discontinuous result.

IMPLICATIONS

FOR POLICY AND PRACTICE

I first consider these findings in the context of the broader education accountability movement. There was a marked increase in available information from No Child Left Behind and other national policies (e.g., Every Student Succeeds Act, Elementary and Secondary Education Act, etc.) that require additional information, namely testing, to be collected by states and municipalities. These findings indicate that while incredibly detailed, individual level data are collected from numerous standardized examinations and/or other data systems, the broad, school-level categorical data causes schools to make some sort of change leading to differences in future test scores. Whether this is because schools are making meaningful changes or not is not determined here; rather, I show evidence above that at least something different is happening as a function of the categorical rating that may have been difficult to calculate without NCLB's systematic collection of data. Still, while the increase in information ushered in by new accountability policies might be celebrated, it is unclear whether schools have the capacity to use that information effectively; that is, taking advantage of its full level of detail. Therefore, while collapsing data down to a more digestible chunk may have its benefits, in these results there is evidence that it perhaps causes limited response when more comprehensive response may be more beneficial.

While it is dangerous to make broad policy recommendations stemming from a single study or perspective, it is worth considering the impact providing clear and distinct information has on schools, again considering if that information (the Student Achievement rating) is a rough approximation of a more subtle yet just as easily available metric (the Student Achievement score). That is, the evidence above suggests that schools are less likely to respond to a continuous measure and more likely to respond to a categorical one. Perhaps, then, district accountability offices might see more efficiently distributed impacts if they create and distribute more simple, categorical measures of school quality to induce positive changes, especially for schools that are close to given cut points. These could include or expand upon the six sub-areas currently in the NYCDOE School Quality Snapshot (NYCDOE, 2018).

A second implication is that there is a need for district-level support for schools and their data teams. While the New York City Department of Education intended to create a more detailed, nuanced look at school quality, the evidence above suggests the response was similar to the old, "one-dimensional" (NYCDOE, 2015) report cards. If schools at the cut points are responding only to the ratings and not the more detailed information contained in the continuous Student Achievement score, that may be because of a lack of knowledge or resources for doing so. School leaders and teachers would perhaps benefit, then, from additional training carried out by district-provided experts in data analysis. If district leaders could develop processes for schools and their leaders to use the more complex and detailed information, perhaps there could be positive impacts across the spectrum and not just at the cut points.

FOR FUTURE RESEARCH

To advance understanding of school report cards and school accountability systems writ large, researchers should extend these analyses to other outcome variables. For instance, they should examine less traditional outcomes besides test scores, some of which are accessible via publicly shared data on the NYCDOE website. First might be survey-related outcomes, including results from parents and teachers; NYCDOE conducts an annual school climate survey (NYCDOE, 2019) that asks parents, teachers, and older (high school) students questions regarding the functioning of their school including evaluating leadership, school culture, and safety. There are a variety of potential outcome variables of interest embedded in the survey, including shifts in trust for principals based on prior rating, or different parental perspectives on a school based on prior rating.

Also, this analysis could be extended to high schools, which would inherently lead to another compelling application. Because New York City has a system of "universal choice" (see Abdulkadiroğlu et al., 2005 for a summary of the system), students have access to these scores prior to making their school application decisions. Perhaps, then, there are causal impacts on not only what happens at a given school, but who chooses to attend; does a lower label cause different students to apply to a given school in comparison to otherwise similar schools with a higher (or lower) label? Each of these potential extensions could be explored in the future.

CONCLUSIONS

The RD analyses suggests that receiving a lower categorical Student Achievement rating on a school accountability report may causally increase test score growth on Math exams for those schools who are close to the cut point, as well as decrease the number of students in the lowest proficiency category while increasing the number of students in scoring proficient. While these results are similar to prior work, there are a few key differences; while prior work (Rockoff & Turner, 2010; Winters & Cowen, 2012) demonstrated a causal impact of low grades, they did so at a time when specific sanctions including financial considerations and choice implications were associated with the report cards. Further, the former report cards were scaled only partially, and to a large group of peer schools. Thus, the prior measured impacts may or may not necessarily have been directly attributable to the report card itself; rather, the sanctions, choice threats, or relative performance to peers may have been motivating factors. Indeed, Hanushek and Raymond (2005) note that shifts from simple public-facing accountability to a system involving consequences had positive impacts on student achievement.

In comparison, the results presented here do not necessarily come attached to a consequential system; there were no such threats associated with a low score at the time. In fact, the reports themselves were designed to be more holistic and inspire a more diverse set of changes (NYCDOE, 2015). This strengthens the argument that the rating itself is causing the shift in score growth. While there are a variety of potential explanations for such a phenomenon, the fact that the rating appears to have a causal impact in and of itself provides important information for those working in and around school accountability: even if detailed information is available, the act of labeling a school with a particular rating can have an impact on its own.

REFERENCES

Abdulkadiroğlu, A., Pathak, P. A., & Roth, A. E. (2005). The new york city high school match. *American Economic Review, 95*(2), 364–367.

Booher-Jennings, J. (2005). Below the Bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal, 42*(2), 231–268. https://doi.org/10.3102/00028312042002231

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). rdrobust: Software for regression discontinuity designs. *The Stata Journal, 17*(2).

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis, 24*(4), 305–331.

Chakrabarti, R. (2007). *Vouchers, public school response, and the role of incentives* (Staff Report No. 306). Federal Reserve Bank of New York. http://eric.ed.gov/?id=ED517702

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9), 1045–1057.

Corcoran, S., & Pai, G. (2013). *Unlocking New York City's High School Progress Report*. New Visions for Public Schools. https://research.steinhardt.nyu.edu/scmsAdmin/media/users/ggg5/Unlocking_NYCs_High_School_Progress_Report_Corcoran_Pai.pdf

Darville, S. (2014, October 1). Read Chancellor Fariña's speech outlining the city's new school rating system, but no other changes. *Chalkbeat*. https://chalkbeat.org/posts/ny/2014/10/01/read-chancellor-farinas-speech-outlining-the-citys-new-school-rating-system/

Dee, T. S., & Jacob, B. (2011). The impact of no Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418–446. https://doi.org/10.1002/pam.20586

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2016). School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics, 98*(5), 848–862.

Deming, D. J., & Figlio, D. (2016). Accountability in US education: Applying lessons from k–12 experience to higher education. *Journal of Economic Perspectives, 30*(3), 33–56. https://doi.org/10.1257/jep.30.3.33

Engel, M. (2000). *The Struggle for Control of Public Education: Market Ideology Vs. Democratic Values*. Temple University Press.

Figlio, D. N., & Lucas, M. E. (2004). What's in a grade? School report cards and the housing market. *American Economic Review, 94*(3), 591–604. https://doi.org/10.1257/0002828041464489

Friesen, J., Javdani, M., Smith, J., & Woodcock, S. (2012). How do school ‘report cards’ affect school choice decisions? *Canadian Journal of Economics/Revue Canadienne d’économique, 45*(2), 784–807. https://doi.org/10.1111/j.1540-5982.2012.01709.x

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business and Economic Statistics, 37*(3), 447-456.

Gootman, E., & Medina, J. (2007, November 6). 50 New York Schools Fail Under Rating System. *The New York Times*. https://www.nytimes.com/2007/11/06/education/06reportcards.html

Hamilton, L. S., & Koretz, D. M. (2002). *Tests and their use in test-based accountability systems* (p. 39). RAND Corporation.

Hansen, M., Levesque, E., Valant, J., & Quintero, D. (2018). *The 2018 Brown Center report on American education: How well are American students learning?* Washington, DC: The Brookings Institution.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management, 24*(2), 297–327. https://doi.org/10.1002/pam.20091

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5–6), 761–796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Lee, D. S., & Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature, 48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3–16. https://doi.org/10.3102/0013189X031006003

NYCDOE. (2015). *Chancellor Fariña Announces More Students Are Graduating College and Career Ready As Part of First Annual ‘School Quality’ Reports Release*. https://www.schools.nyc.gov/about-us/news/announcements/contentdetails/2014/11/10/chancellor-fariña-announces-more-students-are-graduating-college-and-career-ready-as-part-of-first-annual-school-quality-reports-release

NYCDOE. (2018). *Educator Guide - New York City DOE School Quality Reports*. Retrieved from https://infohub.nyced.org/docs/default-source/default-document-library/2018-19-educator-guide-ems---11-13-2019.pdf

NYCDOE. (2019). *NYC School Survey*. Retrieved from https://www.schools.nyc.gov/about-us/reports/school-quality/nyc-school-survey

Rockoff, J., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal. Economic Policy, 2*(4), 119–147. http://dx.doi.org.ezproxy.lib.uconn.edu/10.1257/pol.2.4.119

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal. Economic Policy, 5*(2), 251–281. http://dx.doi.org.ezproxy.lib.uconn.edu/10.1257/pol.5.2.251

Shipps, D., & White, M. (2009). A New Politics of the Principalship? Accountability-Driven Change in New York City. *Peabody Journal of Education, 84*(3), 350–373. https://doi.org/10.1080/01619560902973563

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics, 69*(1), 99. https://doi.org/10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129.

Wall, P. (2014, October 1). Under gentler rating system, schools will no longer be ranked or graded. *Chalkbeat*. https://chalkbeat.org/posts/ny/2014/10/01/under-gentler-rating-system-schools-will-no-longer-be-ranked-or-graded/

Winters, M. A., & Cowen, J. M. (2012). Grading New York: Accountability and student proficiency in America's largest school district. *Educational Evaluation and Policy Analysis, 34*(3), 313–327. https://doi.org/10.3102/0162373712440039