The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations*

Thomas S. Dee
Stanford University and NBER

Will Dobbie
Princeton University and NBER

Brian A. Jacob
University of Michigan and NBER

Jonah Rockoff
Columbia University and NBER

April 2016

Abstract

In this paper, we show that the design and decentralized, school-based scoring of New York's high school exit exams – the Regents Examinations – led to the systematic manipulation of test sores just below important proficiency cutoffs. Our estimates suggest that teachers inflate approximately 40 percent of test scores near the proficiency cutoffs. Teachers are more likely to inflate the scores of high-achieving students on the margin, but low-achieving students benefit more from manipulation in aggregate due to the greater density of these students near the proficiency cutoffs. Exploiting a series of reforms that eliminated score manipulation, we find that inflating a student's score to fall just above a cutoff increases his or her probability of graduating from high school by 27 percent. These results have important implications for educational attainment of marginal high school graduates. For example, we estimate that the black-white graduation gap would have been about 5 percent larger in the absence of test score manipulation.

*We are extremely grateful to Don Boyd, Jim Wyckoff, and personnel at the New York City Department of Education and New York State Education Department for their help and support. We also thank Josh Angrist, David Deming, Rebecca Diamond, Roland Fryer, Larry Katz, Justin McCrary, Petra Persson, Crystal Yang, and numerous seminar participants for helpful comments and suggestions. Elijah De la Campa, Kevin DeLuca, Samsun Knight, Sean Tom, and Yining Zhu provided excellent research assistance. Correspondence can be addressed to the authors by e-mail: tdee@stanford.edu [Dee], wdobbie@princeton.edu [Dobbie], bajacob@umich.edu [Jacob], or jonah.rockoff@columbia.edu [Rockoff]. All remaining errors are our own.

In the U.S. and across the globe, educational quality is increasingly measured using standardized test scores. These standardized test results can carry particularly high stakes for both students and educators, with students' scores often determining, at least in part, grade retention, high school graduation, school closures, and teacher and administrator pay. The tendency to place high stakes on student test scores has led to concerns about the reliability and fidelity of standardized test results (e.g., Neal 2013), as well as about outright cheating; the latter evidenced by Jacob and Levitt's (2003) study of testing in Chicago public schools and the 2009 cheating scandal in Atlanta that led to the incarceration of a number of school administrators.[1] However, despite widespread concerns over test validity and the manipulation of scores, we know little about the factors that lead educators to manipulate student test scores (e.g., accountability policies versus individual students traits). Furthermore, there is also little empirical evidence on whether test score manipulation has any long-run consequences for students' educational outcomes and performance gaps by race, ethnicity, and gender.

In this paper, we examine the causes and consequences of test score manipulation in the context of the New York State Regents Examinations, high-stakes exit exams that measure student performance for New York's secondary-school curricula. The Regents Examinations carry important stakes for students, teachers, and schools, based largely on students meeting strict score cutoffs. Moreover, the Regents Examinations were graded locally for most of our sample period (i.e., by teachers in a student's own school), making it relatively straightforward for teachers to manipulate the test scores of students whom they know and whose scores may directly affect them.

In the first part of the paper, we use data from New York City public schools to document sharp discontinuities in the distribution of student scores at proficiency cutoffs, demonstrating that teachers purposefully manipulated Regents scores in order to move marginal students over the performance thresholds. Formal estimates suggest that approximately 6 percent of all Regents exams in core academic subjects were inflated to fall just above the performance cutoffs between the years 2004 and 2010. Teachers also inflated more than 40 percent of Regents exams within the range of scores that were most at risk for manipulation because they would have been just below the cutoffs. However, test score manipulation was reduced by approximately 80 percent in 2011 when the New York State Board of Regents ordered schools to end the longstanding practice of re-scoring exams with scores just below proficiency cutoffs, and disappeared completely in 2012 when the Board ordered that Regents exams be graded by teachers from other schools in a small number of centrally administered locations. These results suggest that both re-scoring policies and local grading are key factors in teachers' willingness or ability to manipulate test scores.

Importantly, we find that the extent of pre-reform manipulation varied systematically across schools and students in a way that suggests the effects of manipulation were highly inequitable.

[1] See https://en.wikipedia.org/wiki/Atlanta_Public_Schools_cheating_scandal. In related work, there is evidence that test-based accountability pressures lead some teachers to narrow their instruction to the tested content (Jacob 2005) and target students who are near performance thresholds (Neal and Schanzenbach 2010). There is also evidence some schools sought to advantageously manipulate the test-taking population following the introduction of test-based accountability (Figlio and Getzler 2002, Cullen and Reback 2002, Jacob 2005).

Black and Hispanic students and students with lower baseline scores and worse behavioral records all benefited more from the manipulation in aggregate due to the greater density of these students near the proficiency cutoffs. However, these same students were all significantly less likely to have a test manipulated conditional on scoring near a proficiency cutoff. We find that approximately 60 percent of the across-race differences were due to across-school variation in manipulation, with the remainder due to within-school variation in the probability of having a score manipulated. Conversely, nearly all of the differential treatment on baseline score and behavioral dimensions can explained by within-school differences in the probability of having a score manipulated.

In the second part of the paper, we estimate the impact of test score manipulation on subsequent student outcomes using the arguably exogenous timing of the decisions by the New York State Board of Regents to end the practice of re-scoring exams and to centralize scoring. Using a difference-in-differences research design, we find that having an exam score manipulated to fall above a performance cutoff increases the probability of graduating from high school by 21.9 percentage points, a 27.4 percent increase from the sample mean.[2] The effects on high school graduation are larger for students eligible for free or reduced price lunch and students with higher baseline test scores, but remain economically and statistically significant for all student subgroups. Estimates from a second empirical strategy that exploits across-school differences in manipulation yield similar results. Taken together with our earlier results showing across-group differences in manipulation, these difference-in-differences estimates suggest that test score manipulation had important, inequitable effects on the graduation outcomes of students in New York City. Indeed, our point estimates suggest that the black-white gap in graduation rates would have increased from 15.6 percentage points to 16.3 percentage points without test score manipulation, and that the fraction of students in our sample graduating from high school would have decreased by 1.2 percentage points without manipulation.

However, we also find that having an exam score manipulated decreases the probability of meeting the requirements for a more advanced high school diploma by 11.0 percentage points, a 50.2 percent decrease from the sample mean. Having a score manipulated also modestly decreases the probability of passing a physical science exam such as Chemistry or Physics, and passing an advanced math sequence that covers topics such as geometry and trigonometry, the two most important requirements for the more advanced diploma, with larger effects when the manipulation occurs on the introductory math and science exams. These results are consistent with the idea that test score manipulation has somewhat heterogeneous effects on human capital accumulation. Students on the margin of dropping out are "helped" by test score manipulation because they are not forced to retake a class that may lead them to leave high school. Conversely, students on the margin of the advanced diploma may be "hurt" by test score manipulation because they are

[2] An important limitation of our difference-in-differences analysis is that we are only able to estimate the effect of eliminating manipulation in partial equilibrium. There may also be important general equilibrium effects of test score manipulation that we are unable to measure using our empirical strategy. For example, it is possible that widespread manipulation may change the way schools teach students expected to score near proficiency cutoffs. It is also possible that test score manipulation can change the signaling value of course grades or a high school diploma.

not pushed to re-learn the introductory material or re-take the introductory class that the more advanced coursework requires.

In the final part of the paper, we present several pieces of evidence to help us shed light on the motivations for test score manipulation prior to the grading reforms. First, we find nearly identical levels of manipulation before the passage of No Child Left Behind (NCLB) and the creation of New York City's own school accountability system. Manipulation was also remarkably similar in schools that are and are not subject to sanctions under either accountability system. Second, we find that a randomized experiment that paid high school teachers for improving several student outcomes, most of which related to Regents exam pass rates, had no effect on test score manipulation. Third, we find evidence of manipulation around both the lowest cutoffs on core exams (which determine eligibility for a basic high school diploma) and for higher cutoffs on elective exams, which provide students with secondary benefits such as taking advanced coursework, automatic admission to some public colleges, or the granting of college credit. We argue that, taken together, these three results suggest that altruism among teachers provide an important motivation for teachers' manipulation of test scores (i.e., helping students avoid any sanctions involved with failing an exam). Interestingly, we also find that a teacher's propensity to manipulate a student's exam is influenced by other student-specific information (e.g., their prior test scores and good behavior).

Our results contribute to an emerging literature that documents both the moral hazard that can be created by test-scoring procedures and the impact of such test manipulation for horizontal equity (e.g., fairness issues raised by differential manipulation) and vertical equity (e.g., the gaps in student outcomes). In early work, Jacob and Levitt (2003) find that test score manipulation occurs in roughly five percent of elementary school classrooms in the Chicago public schools, with the frequency of manipulation responding strongly to relatively small changes in incentives. Outside of the U.S., Lavy (2009) finds that a teacher incentive program in Israel increased teacher effort but did not affect test score manipulation, and Angrist, Battistin, and Vuri (2014) find that small classes increase test score manipulation in Southern Italy due to teachers shirking when they transcribe answer sheets. The paper most closely related to ours is parallel work by Diamond and Persson (2016), who find significant manipulation of test scores just above discrete grade cutoffs in Sweden. Exploiting across-school variation in exposure to test score manipulation, they find that having a score inflated increases educational attainment by 0.5 to 1 year, with larger attainment effects and some evidence of earnings effects for low-skill students. Diamond and Persson (2016) also find no evidence of the negative human capital effects we document in our setting. Finally, in related work, Ebenstein, Lavy, and Roth (forthcoming) find that quasi-random declines in exam scores due to pollution exposure have a negative effect on post-secondary educational attainment and earnings.

This paper is also related to an important literature on bias in local grading. Lavy (2008) uses the difference between non-blind and blind exams in Israel to show teachers grade girls more leniently than boys. Using a similar methodology, Hinnerich, Hoglin, and Johannesson (2011) find no evidence of gender bias in Sweden, while Burgess and Greaves (2013) find that white students were graded more leniently than non-white students in England. Hanna and Linden (2012)

randomly assign student characteristics to exams in India, finding that teachers grade exams from higher-caste students more leniently than lower-caste students. Most recently, Lavy and Sand (2015) show that teachers' grading biases can have important impacts on subsequent achievement and enrollment.

The remainder of the paper is structured as follows. Section I describes the Regents Examinations and their use in student and school evaluations. Section II details the data used in our analysis. Section III describes our empirical design and documents manipulation on the Regents exams. Section IV estimates the impact of manipulation on student outcomes. Section V explores potential motivations for the manipulation of Regents scores, and Section VI concludes.

I. New York Regents Examinations

In 1878, the Regents of the University of the State of New York implemented the first statewide system of standardized, high-stakes secondary school exit exams. Its goals were to assess student performance in the secondary-school curricula and award differentiated graduation credentials to secondary school students (Beadie 1999, NYSED 2008). This practice has continued in New York state to the present day.[3]

A. Regents Examinations and High School Graduation

In recent years, public high school students in New York must meet certain performance thresholds on Regents examinations in five "core" subjects to graduate from high school: English, Mathematics, Science, U.S. History and Government, and Global History and Geography.[4] Regents exams are also given in a variety of other non-core subject areas, including advanced math, advanced science, and a number of foreign languages.[5] Regents exams are administered within schools in January, June, and August of each calendar year, with students typically taking each exam at the end of the corresponding course.

[3]The original Regents exams were administered as high school entrance exams for 8th grade students as early as 1865. These entrance exams were phased out relatively quickly, however. Among other changes over time, New York introduced a new minimum competency test in the late 1970s that students were required to pass in order to graduate from high school. This competency test was replaced in the late 1990s by graduation requirements tied to the more demanding, end-of-course Regents Examinations described in this section (Chudowsky et al. 2002).

[4]The mathematics portion of the Regents exam has undergone a number of changes during our sample period (2004-2013). From 1977-2002, students were required to pass the Sequential Math 1 exam, which covered primarily algebra, to graduate from high school. Sequential Math 2 and Sequential Math 3 were optional math courses available for students wanting to cover more advanced material. From 2003 to 2009, students were required to pass the Math A exam, which covered approximately the same material as the first 1.5 courses in the Sequential Math sequence, to graduate. Compared to Sequential Math 1, Math A had fewer multiple choice questions and more long-answer questions, and included a number of new subjects like geometry and trigonometry. A Math B exam was also available during this period for more advanced students. From the 2009 to the present, the Regents exams reverted back to year long math courses separated into Algebra, Geometry, and Algebra 2. Students are only required to pass this first Algebra course to graduate from high school. There was a year of overlap between the Math A/B exams and the current math exams because while Math A was typically taken by 10th grade students, the first Algebra course under the current system is typically taken by 9th grade students.

[5]New York recently approved alternatives to the Regents Examinations based on comparable performance thresholds in Advanced Placement, International Baccalaureate, and SAT II exams (NYSUT 2010). Information on these alternatives to the Regents exams is not included in our data.

An uncommon and important feature of the Regents exams is that they are graded by teachers from students' own schools. The State Education Department of New York provides explicit guidelines for how the teacher-based scoring of each Regents exam should be organized (e.g., NYSED 2009), which we discuss in greater detail below. After the exams are graded locally at schools, the results are sent to school districts and, ultimately, to the New York State Education Department.

Regents exam scale scores range from 0 to 100. In order to qualify for a "local diploma," the lowest available in New York, students entering high school before the fall of 2005 were required to score at least 55 on all five core examinations. The requirements for a local diploma were then raised for each subsequent entry cohort, until the local diploma was eliminated altogether for students entering high school in the fall of 2008. For all subsequent cohorts, the local diploma has only been available to students with disabilities. In order to receive a (more prestigious) Regents Diploma, students in all entry cohorts were required to score at least 65 on all five core Regents exams.[6] To earn the even more prestigious "Advanced" Regents Diploma, students must also score at least a 65 on additional elective exams in math, science, and foreign language. Appendix Table 1 provides additional details on the degree requirements for each cohort in our sample.[7]

B. The Design and Scoring of Regents Examinations

Regents examinations contain both multiple-choice and open-response or essay questions. For example, the English exam typically includes both traditional multiple-choice questions and open-response questions that ask students to read a passage (e.g., a speech, an informative text with tables or figures, or a literary text) and respond to an essay prompt. Similarly, both the U.S. History and Global History exams include multiple-choice questions, open-response questions, and thematic essays, while the foreign language exams also contain a speaking component. Scoring materials provided to schools include the correct answers to multiple-choice questions, and detailed, subject-specific instructions for evaluating open-response and essay questions.[8]

[6]Beginning with students entering high school in the fall of 2005, eligible students may appeal to graduate with a local or Regents diploma using a score between 62 and 64. Students are eligible to appeal if they have taken the Regents Examination under appeal at least two times, have at least one score between 62 and 64 on this exam, have an attendance rate of at least 95 percent for the most recent school year, have a passing course average in the Regents subject, and is recommended for an exemption by the student's school. In addition, students who are English language learners and who first entered school in the U.S. in grade 9 or above may appeal to graduate with a local diploma if they have taken the required Regents Examination in English language arts at least twice and earned a score on this exam between 55 and 61.

[7]In addition to the important proficiency cutoffs at 55 and 65, there are also less important cutoffs at 75 and 85 scale score points. The 75 point cutoff is sometimes used by New York state community colleges as either a prerequisite or qualification for credit towards a degree, and the 85 point cutoff is often used by high schools as a prerequisite for courses (e.g., Advanced Placement) and by New York State colleges as either a prerequisite or qualification for credit towards a degree. Beginning with students who entered 9th grade in the fall of 2009, an additional accolade of "Annotation of Mastery" also became available for students scoring above 85 on three Regents exams in science or math. While we focus on the relatively more important cutoffs at 55 and 65 in our analysis, there is also visual evidence of a small amount of manipulation at both the 75 and 85 cutoffs.

[8]For the English and social-studies exams, principals are required to designate a scoring coordinator who is responsible for managing the logistics of scoring, assigning exams to teachers, and providing teachers with necessary training. For essay questions, the materials available to support this training include scoring rubrics and pre-scored "anchor papers" that provide detailed commentary on why the example essays merited different scores. For open-

To help ensure consistent scoring, essays are given a numeric rating of one to four by two independent graders. If the ratings are different but contiguous, the final essay score is the average of the two independent ratings. If the ratings are different and not contiguous, a third independent grader rates the essay. If any two of the three ratings are the same, the modal rating is taken. The median rating is taken if each of the three ratings is unique. The number of correct multiple-choice items, the number of points awarded on open-response questions, and the final essay scores are then converted into a final scale score using a "conversion chart" that is specific to each exam. Scale scores range from 0 to 100 on all Regents exams, but typically not all 100 scale scores are possible on any single exam.

During our primary sample period (2003-2004 to 2009-2010), math and science Regents exams with scale scores between 60 and 64 were required to be re-scored, with different teachers rating the open-response questions.[9] Principals at each school also had the discretion to mandate that math and science exams with initial scale scores from 50 to 54 be re-scored. This re-scoring policy is clearly important for our study. Although we find evidence of manipulation in every Regents exam subject area, the policy of re-scoring math and science exams may influence how principals and teachers approach scoring Regents exams more generally. We discuss the impact of this re-scoring policy on our results in greater depth in Section III.C, where we use changes in the Regents re-scoring policies between 2011 and 2013 to examine the mechanisms leading to test score manipulation.

C. Regents Examinations and School Accountability

Beginning in 2002-2003, high schools in New York state have been evaluated under the state accountability system developed in response to the federal No Child Left Behind Act (NCLB). Whether a public high school in New York is deemed to be making Adequate Yearly Progress (AYP) towards NCLB's proficiency goals depends on five measures, all of which are at least partially based on the Regents Examinations.[10] Motivated by perceived shortcomings with NCLB, the NYCDOE

ended questions, the materials include a rubric to guide scoring. A single qualified teacher grades the open-ended questions on the social-science exams. In the math exams, the school must establish a committee of three mathematics teachers to grade the examinations, and no teacher should rate more than a third of the open-ended questions in mathematics. In the science exams, the school must establish a committee of two science teachers to grade the examinations, and no teacher should rate more than a half of the open-ended questions.

[9]Grading guidelines distributed to teachers typically included the following text explaining this policy: "All student answer papers that receive a scale score of 60 through 64 must be scored a second time to ensure the accuracy of the score. For the second scoring, a different committee of teachers may score the student's paper or the original committee may score the paper, except that no teacher may score the same open-ended questions that he/she scored in the first rating of the paper. The school principal is responsible for assuring that the student's final examination score is based on a fair, accurate and reliable scoring of the student's answer paper." See for example: https://www.jmap.org/JMAPRegentsExamArchives/INTEGRATEDALGEBRAEXAMS/0610ExamIA.pdf. Two exceptions to this rule that we are aware of are the Chemistry examination in June 2001, which was only based on multiple choice questions, and the Living Environment exam in June 2001, where exams with scale scores from 62 to 68 were to be re-scored.

[10]First, 95 percent of a school's 12th graders must have taken the Regents Examinations in mathematics and English or an approved alternative (NYSED 2010). Second, the same must be true for all sub-groups with at least 40 students. Third and fourth, a school's performance indices based on the Regents examinations in math and English must meet the statewide objectives for both its overall student population and among accountability sub-groups. The subject-specific performance indices are increasing in the share of students whose scale scores on the Regents

implemented its own accountability system starting in 2006-2007. The central component of the NYCDOE accountability system is the school progress reports, which assigned schools a letter grade, ranging from A to F. For high schools, the school grades assigned through the NYC accountability system also depend heavily on Regents pass rates, particularly pass rates in the core academic subjects that determine high school graduation.[11] We examine the role of these accountability systems in motivating test score manipulation in Section V.

II. Data

We use administrative enrollment and test score data from the New York City Department of Education (NYCDOE).[12] The NYCDOE data contain student-level administrative records on approximately 1.1 million students across the five boroughs of the NYC metropolitan area. The data include information on student race, gender, free and reduced-price lunch eligibility, behavior, attendance, matriculation for all students, state math and English Language Arts test scores for students in grades three through eight, and Regents test scores for high school students. The Regents data provide exam-level information on the subject, month, and year of the test, the scale score, and a school identifier. We have complete NYCDOE data spanning the 2003-2004 to 2012-2013 school years, with Regents test score and basic demographic data available from the 2000-2001 school year.

We also collected the charts that convert raw scores (e.g., number of multiple choice correct,

Examination exceed 55, with students whose scores exceed 65 having twice the impact on this index. Specifically, the performance index equals $100 \cdot$ [(count of cohort with scale scores ≥ 55 + count of cohort with scale scores ≥ 65) / cohort size] (NYSED 2010). Thus, the performance index ranges from 200 when all students have scale scores of 65 or higher to 0 when all students have scale scores below 55. These state-mandated performance objectives increased annually in order to meet NCLB's mandated proficiency goals for the school year 2013-2014. The fifth measure relevant to whether a high school makes AYP under New York's accountability system is whether its graduation rate meets the state standard, which is currently set at 80 percent. Like the other criteria, this standard is also closely related to the Regents Examinations, since eligibility for graduation is determined in part by meeting either the 55 or 65 scale score thresholds in the five core Regents Examinations.

[11] To form the school grades, the NYCDOE calculated performance within three separate elements of the progress report: school environment (15 percent of the overall score), student performance (20-25 percent), and student progress (55-60 percent). The school environment score was determined by responses to surveys of students (in grades 6 and above), parents, and teachers, as well as student attendance rates. For high schools, student performance is measured using the four year graduation rate, the six year graduation rate, a 'weighted' four year graduation rate, and a 'weighted' six year graduation rate. The weighted graduation rates assign higher weights to more advanced diploma types based on the relative level of proficiency and college readiness the diploma indicates. Student progress is measured using a variety of metrics that indicate progress toward earning a high school degree. Most importantly for our analysis, student progress includes the number of passed Regents exams in core subjects. Student progress also depends on a Regents pass rate weighted by each student's predicted likelihood of passing the exam. A school's score for each element (e.g., student progress) is determined both by that school's performance relative to all schools in the city of the same type and relative to a group of peer schools with observably similar students. Performance relative to peer schools is given triple the weight of citywide relative performance. A school's overall score was calculated using the weighted sum of the scores within each element plus any additional credit received. Schools can also receive "additional credit" for making significant achievement gains among students with performance in the lowest third of all students citywide who were Hispanic, Black, or other ethnicities, and students in English Language Learner (ELL) or Special Education programs. See Rockoff and Turner (2010) for additional details on the NYCDOE accountability system.

[12] Appendix A contains all of the relevant information on the cleaning and coding of the variables used in our analysis. This section summarizes the most relevant information from the appendix.

number of points from essays), to scale scores for all Regents exams taken during our sample period. We use these conversion charts in three ways. First, we identify a handful of observations in the New York City data that contain errors in either the scale score or test identifier, i.e., that do not correspond to possible scale scores on the indicated exam. Second, we map raw scores into scale scores for math and science exams so that we can account for predictable spikes in the distribution of scale scores when this mapping is not one to one. Third, we identify scale scores that are most likely to be affected by manipulation around the proficiency cutoffs. See Section III.A for additional details on both the identification of manipulable scores and the mapping of raw to scale scores.

We make several restrictions to our main sample. First, we focus on Regents exams starting in 2003-2004 when tests can be reliably linked to the student enrollment files. We return to tests taken in the school years 2000-2001 and 2001-2002 in Section V to assess manipulation prior to the introduction of NCLB and the NYC school accountability system. Second, we use each student's first exam for each subject to avoid any mechanical bunching around the performance thresholds due to re-taking behavior. In practice, however, results are nearly identical when we include re-tests. Third, we drop August exams, which are far less numerous and typically taken after summer school, but our results are again similar if we use all test administrations. Fourth, we drop students who are enrolled in middle schools, a special education high school, or any other non-standard high school (e.g., dropout prevention schools). Fifth, we drop observations with scale scores that are not possible on the indicated exam (i.e. where there are reporting errors), and all school-exam cells where more than five percent of scale scores are not possible. Finally, we drop special education students, who are subject to a different set of accountability standards during our sample period (see Appendix Table 1), although our results are again similar if we include special education students. These sample restrictions leave us with 1,630,259 core exams from 514,679 students in our primary window of 2003-2004 to 2009-2010. Table 1 contains summary statistics for the resulting dataset, and Appendix A includes additional information on our sample restrictions and the number of observations dropped by each.

III. The Manipulation of Regents Exam Scores

A. Documenting the Extent of Manipulation: Estimates from 2004-2010

We begin by examining bunching in the distribution of core Regents exam scores near the proficiency thresholds at 55 and 65 points for all core Regents exams taken between 2003-2004 and 2009-2010 in Figure 1. We initially focus only on tests during this time period, as exams taken after 2009-2010 are subject to a different set of grading policies which we discuss in see Section III.C.

To construct Figure 1 and all subsequent test score plots, we first collapse the data to the subject-year-month-score level (e.g., Living Environment, June 2004, 77 points). We then make adjustments to account for two mechanical issues that affect the smoothness of the distribution of scale scores. First, we adjust for instances when the number of raw scores that map into each

scale score is not one to one, which causes predictable spikes in the scale score frequency.[13] We also adjust Integrated Algebra and Math A exams for an alternating pattern of spikes in frequency at very low, even raw scores (i.e., 2, 4, 6, etc.) likely due to students who only received credit for a small number of multiple choice questions, worth two scale score points each. For these exams, we average adjacent even and odd scores below 55, which generates total smoothness at this part of the distribution. All of our results are similar but slightly less precise if we do not make these adjustments. Finally, we collapse the adjusted counts to the scale score level and, in Figure 1, plot the fraction of tests in each scale score around the proficiency thresholds, demarcated by the vertical lines at 55 and 65 points.

Figure 1 shows that there are clear spikes around the proficiency cutoffs in the otherwise smooth test score distribution, and the patterns are strongly suggestive of manipulation. The scores immediately below these cutoffs appear less frequent than one would expect from a well-behaved empirical distribution, and the scores at or just above the cutoffs appear more frequent than one would expect. In Appendix Figures 1 and 2, we show that this pattern is still apparent when we examine test scores separately by subject or by year.[14]

To estimate the magnitude and statistical significance of any manipulation in Figure 1 formally, we construct a counterfactual test score distribution using the approach developed by Chetty et al. (2011).[15] First, we fit a polynomial to the counts plotted in the figure, excluding data near the proficiency cutoffs with a set of indicator variables, using the following regression:

$$F_{semt} = \sum_{i=0}^{q} \pi_{iemt} \cdot (\text{Score})^i \; + \sum_{i \in -M_{cemt}, +M_{cemt}} \lambda_{iemt} \cdot \mathbf{1}[\text{Score} = i] \; + \; \varepsilon_{semt} \tag{1}$$

where F_{semt} is the fraction of students with a Regents scale score of s on exam e in month m and year t, Score is the Regents scale score, q is the order of the polynomial, $-M_{cemt}$ denotes the potentially manipulable scores to the left of the proficiency cutoff c, and $+M_{cemt}$ denotes the potentially manipulable scores at or to the right of the proficiency cutoff. We define an estimate of the counterfactual distribution $\{\widehat{F}_{semt}\}$ as the predicted values from (1) omitting the contribution of the indicator variables around the cutoffs: $\widehat{F}_{semt} = \sum_{i=0}^{q} \widehat{\pi}_{iemt} \cdot (Score)^i$. In practice, we estimate $\{\widehat{F}_{semt}\}$ using a sixth-degree polynomial (q=6) interacted with the exam subject e, but constant across years for the same exam subject. Our results are not sensitive to changes in either the polynomial order or whether we allow the polynomial to vary by year or subject.

A key step in estimating equation (1) is identifying the potentially manipulable test scores

[13] For example, on the June 2004 Living Environment Exam, a scale score of 77 points corresponds to either a raw score of 57 or 58 points, while scale scores of 76 or 78 points correspond only to raw scores of 56 or 59 points, respectively. Thus, the frequency of scale score 77 (1,820 exams) is roughly two times higher than the frequency of scale scores of 76 (915) or 78 (917). Our approach is based on the assumption of continuity in underlying student achievement, and thus we adjust the frequencies when raw to scale score mappings are not one to one.

[14] Appendix Figure 2 shows that the amount of manipulation around the 55 cutoff is decreasing over time. This pattern is most likely due to the decreasing importance of the 55 cutoff for graduation over time (see Appendix Table 1). We therefore focus on the 65 cutoff when examining manipulation after 2010.

[15] See Saez (2010), Kleven and Waseem (2013), and Persson (2015) for other examples of "bunching" estimators.

around each cutoff. In other applications of "bunching" estimators, such as constructing counterfactual distributions of taxable income around a kink in marginal tax rates, it has not generally been possible to specify *ex ante* the range of the variable in which manipulation might take place. However, in our case we believe that we are able to identify potentially manipulable or manipulated test scores on both the right and left sides of the proficiency cutoffs based on knowledge of the Regents grading rules. We define a score as manipulable to the left of each proficiency cutoff if it is between 50-54 or 60-64. Recall that math and science exams scored between 60-64 are automatically re-graded during our sample period, with many principals also choosing to re-grade exams scored between 50-54. This range is also consistent with the patterns observed in Figure 1. To the right of each cutoff, we define a score as manipulable differently by subject area. In math and science, it is always possible to award enough additional raw points through partial credit on open-response questions in order to move a student from just below the cutoff to exactly a score of 55 or 65. In contrast, for the exams in English and social studies, a score of exactly 55 or 65 may not always be possible if manipulation is done through changes in scores to essay questions. This is because changes in essay ratings of just one point typically change the scale score by four points. We therefore define a score as manipulable for the math and science exams if it is the closest scale score to the right of the proficiency threshold. For the English and social science exams, we define a score as manipulable to the right of the cutoff if it is within 1 essay point of the proficiency threshold. This differential range by subject is consistent with the patterns observed in Appendix Figure 1. Our estimates are not sensitive to changes in the manipulable score region to either the left or right side of the proficiency cutoffs.

If our demarcation of the manipulable range is accurate, then the unadjusted counterfactual distribution from equation (1) should satisfy the integration constraint, i.e. the area under the counterfactual distribution should equal the area under the empirical distribution. Consistent with this assumption, we find that the missing mass from the left of each cutoff is always of similar magnitude to the excess mass to the right of each cutoff. In contrast, Chetty et al. (2011) must use an iterative procedure to shift the counterfactual distribution from equation (1) to the right of the tax rate kink to satisfy the integration constraint. Given that the integration constraint is satisfied in our context, we use an average of the missing mass and excess mass at each cutoff to increase the precision of our estimates. However, our results are similar if we only use the excess mass to the right of each cutoff.

Let β_{cemt} denote the excess number of test takers who are located to the right of the cutoff c for exam e in month m and year t, or the total amount of manipulation for that cutoff and test administration. Given the setup discussed above, our estimate of the total amount of manipulation for each test administration is $\widehat{\beta}_{cemt} = \frac{1}{2} \cdot \sum_{i \in +M_{cemt}} (F_{set} - \widehat{F}_{semt}) + \frac{1}{2} \cdot \left| \sum_{i \in -M_{cemt}} (F_{set} - \widehat{F}_{semt}) \right| = \frac{1}{2} \cdot \sum_{i \in +M_{cemt}} \widehat{\lambda}_{iemt} + \frac{1}{2} \cdot \left| \sum_{i \in -M_{cemt}} \widehat{\lambda}_{i} \right|$. We also report an estimate of "in-range" manipulation, or the probability of manipulation conditional on scoring just below a proficiency cutoff, which is defined as the excess mass around the cutoff relative to the average counterfactual density in the manipulable score regions: $\widehat{\beta}_{cemt} / \sum_{i \in -M_{cemt}, +M_{cemt}} \widehat{F}_{semt}$. We calculate both total and in-range

manipulation at the cutoff-exam-year level to account for the fact that each test administration potentially has a different set of manipulable scores. In pooled specifications such as that shown in Figure 1, we report the average manipulation across all cutoff-exam-year administrations weighted by the number of exams in each exam-year. In practice, our results are not sensitive to changes in the polynomial order, the manipulable score region, or the way we weight manipulation totals across exams because the manipulation estimates we document are much larger than the changes induced by varying the specification.

We calculate standard errors for our manipulation estimates using a version of the parametric bootstrap procedure developed in Chetty et al. (2011). Specifically, we draw with replacement from the entire distribution of estimated vector of errors $\widehat{\varepsilon}_{set}$ in (1) at the score-exam-test administration level to generate a new set of scale score counts at the exam-test administration level. We then apply the approach described above to calculate 200 new manipulation estimates, and define the standard error as the standard deviation of these 200 bootstrapped estimates.

The dotted line in Figure 1 shows the counterfactual density $\{\widehat{F}_{semt}\}$ predicted using (1), as well as our point estimates and standard errors for manipulation. Using the above parameters, we estimate the average amount of manipulation on the Regents core exams to be 5.8 (se=0.04). That is, we estimate that approximately 6 percent of all Regents core exams between 2004 and 2010 were manipulated to fall just above a proficiency cutoff. Within the range of potentially manipulable scores, we estimate that an average of 44.5 (se=0.26) percent of Regents core exams were manipulated to fall just above a cutoff. These estimates confirm our qualitative conclusion that test scores are far more likely to fall just above a proficiency cutoff than one would expect from a well-behaved empirical distribution. Appendix Table 2 presents results separately for all subjects and test administrations. There is economically and statistically significant manipulation of all Regents core exams in our sample. None of the results suggest that the manipulation documented in Figure 1 is the result of one particular subject or administration.[16]

To provide further evidence that Regents scores near cutoffs were being manipulated, Appendix Figure 3 examines the score distributions for math and English exams taken by New York City students in grades 3 to 8, which also involve high stakes around specific cutoffs but are graded centrally by the state. The distributions for the grade 3 to 8 exams trend smoothly through the proficiency cutoff for these centrally graded exams, and estimates of a discontinuity in the distribution at the proficiency cutoff produce very small point estimates that are statistically insignificant. Thus, there seems to be nothing mechanical about the construction of high stakes tests in New York State that could reasonably have lead to the patterns we see in Figure 1.

[16]The math and science exams tend to have lower levels of manipulation than the English and social science exams. The math and science exams also have fewer open response questions and more multiple choice questions compared to the English and social science exams. The June 2001 Chemistry exam is the only test in our data that consists of multiple-choice questions. Dee et al. (2011) show there is a clear discontinuity in the distribution of Chemistry test scores at 65 points despite the lack of open-response questions. However, the amount of manipulation is significantly less than otherwise similar elective exams. Both sets of results are consistent with the idea that teachers view manipulation on multiple choice items as more costly than open-response items, but not so costly as to eliminate manipulation entirely.

Finally, we note that Dee et al. (2011) find evidence of similar manipulation in state-wide data from the June 2009 administration of the Regents exams and show that the manipulation was not specific to New York City. Unfortunately, these state-wide data do not provide include information on student characteristics and are only available for one year.

B. Heterogenous Treatment of Students

This section examines how test score manipulation differentially impacts students and schools in New York City. Differences in the amount of total manipulation across schools or students have important implications for widely used performance metrics, such as the Regents pass rates used in NCLB and the NYCDOE accountability system. Similarly, differences in the amount of in-range manipulation across students or schools may suggest that local grading results in unfair or biased assessments (e.g., Lavy 2008), particularly given the fact that not all students with scores just below the cutoffs have their scores manipulated. Finally, differences in both total and in-range manipulation may have important implications for across-student and across-school performance, particularly if test score manipulation has significant effects on longer-run academic outcomes such as high school graduation.

We examine how manipulation varies across students and schools in four ways. First, we estimate aggregate results separately by various dichotomous school characteristics. We then separately estimate the amount of manipulation for each high school in our sample. These school-specific estimates shed light on how manipulation varies across both observed and unobserved school characteristics, and, unlike our aggregate results, allow us to explore how both school manipulation varies across continuous school characteristics and what school characteristics are most predictive in a multivariate regression framework. Third, we estimate aggregate results separately by various dichotomous student characteristics. Finally, we use a simple Monte Carlo procedure to examine whether any differences in manipulation across students is due to across- or within-school variation. In Section IV.C, we consider the potential implications of the differential treatments documented below.

Differences Across Schools: Figure 2 reports results separately for mutually exclusive school subgroups. We estimate the school subgroup results using separate sixth order polynomials for each subgroup. We calculate subgroup specific standard errors using the bootstrap procedure described above where we sample only within the subgroup specific vector of errors. We find that schools in the top half of black and Hispanic enrollment (more than 79.2 percent) manipulate 7.0 (se=0.04) percent of all core exams and 43.8 (se=0.22) percent of in-range exams. In schools in the bottom half of black and Hispanic enrollment, 4.6 (se=0.05) percent of all core exams are manipulated and 45.7 (se=0.40) percent of in-range exams are manipulated. We similarly find that schools with higher fractions of students eligible for free or reduced price lunch (more than 60.9 percent) manipulate about 1.6 percentage points more exams than schools with fewer disadvantaged students, but are about 3.1 percentage points less likely to manipulate in-range exams. Finally,

schools whose students had low average 8th grade test scores (below -0.121 standard deviations) manipulate about 3.1 percentage points more exams than schools with higher average achievement, and 2.6 percentage points more likely to manipulate in-range exams.

To explore this issue further, we separately estimate the amount of total and in-range manipulation for each high school in our sample. For each test subject, we split high schools into five equal sized bins based on average Regents score for that test administration. We then estimate the counterfactual distribution for all exams in the test score bin using equation (1). We allow the counterfactual distribution of exams to vary by test quintile to account for the fact that high- and low-achieving schools have different test score distributions. Results are qualitatively similar splitting schools into fewer or more quantiles, or restricting to the subset of very large high schools where we can estimate school-specific counterfactual distributions. After estimating the counterfactual distribution for each subject x average test score quintile, we calculate the total and in-range manipulation for each school using the procedure outlined above. This procedure provides us with a measure of total and in-range manipulation at the school x test x year x month x cutoff level. We limit our analysis to observations with at least 10 students scoring in the manipulable range for the school x year x month x cutoff, which leaves us with 9,392 observations spread across 276 schools ranging from 2004 to 2010.

Appendix Figure 4 plots the distribution of total and in-range manipulation estimates from the above sample collapsed to the school level. Consistent with our results from Figure 1, we find considerable mass around 4 to 5 percent for total manipulation and around 50 percent for in-range manipulation. At the same time, there are many high schools with extremely high or low estimated manipulation. Because each of these individual estimates is measured with error, the distribution shown in Appendix Figure 4 will overstate the true variance of school "effects" in the population (Jacob and Rothstein forthcoming). To recover the true distribution of school effects in our sample, we estimate the following random effects model:

$$Manipulation_{cemth} = \alpha_e + \alpha_m + \alpha_{ct} + \upsilon_h + \varepsilon_{cemth} \tag{2}$$

where $Manipulation_{cemth}$ is the estimated total or in-range manipulation at cutoff c for exam subject e in month m and year t at high school h, α_e are exam subject effects, α_m are month effects, α_{ct} are cutoff x year effects, and υ_h is a random school effect.

Estimates of equation (2) suggest considerable across-school variation in manipulation, particularly for in-range manipulation. The mean school effect υ_h for total manipulation the standard deviation in school effects equal to 1.44 percentage points based on a baseline mean school effect of 3.76 percentage points. For in-range manipulation, however, the standard deviation is 16.6 percentage points on a mean of 47.5 percentage points.[17] These estimates confirm that, even after

[17]Estimates from equation (2) in tabular form are available upon request. Consistent with our earlier results, in-range manipulation is somewhat higher in English and history than math and science. Manipulation rates are also similar across cutoffs in 2004, but over time the prevalence of manipulation increased (decreased) at the 65 (55) cutoff as the high school graduation requirements have changed.

accounting for measurement error, schools differ considerably in their propensity to manipulate test scores.

To examine whether the variation across schools is associated with observable school characteristics, we estimate several extensions of the above random effects model. Table 2 presents results from a series of regressions of equation (2) that include selected school-level observable characteristics. The observations are weighted by the number of in-range exams contributing to the manipulation estimate. Not surprisingly, we find that total manipulation is markedly higher in schools that have higher fractions of students who tend to score closer to the cutoffs: black and Hispanic students, students eligible for free or reduced price lunch, and students with lower 8th grade test scores.

Similarly, in-range manipulation is higher for schools that have a higher fraction of black and Hispanic students, a higher fraction of students eligible for free or reduced price lunch, and lower 8th grade test scores, though the free lunch result is not statistically significant. In contrast to the total manipulation results, only the point estimate on the fraction of black and Hispanic students remains statistically significant when we include all three school characteristics in the regression. The results from this combined specification suggest that a ten percentage point increase in the fraction of black and Hispanic students is associated with a 1.78 (se=0.95) percentage point lower probability of manipulating an in-range test score, a 3.8 percent decrease from the mean in-range manipulation estimate.

Differences Across Students: Figure 3 reports results separately for mutually exclusive student subgroups. Following our school subgroup results, we estimate the student subgroup results using separate sixth order polynomials for each subgroup and calculate subgroup specific standard errors using the bootstrap procedure described above where we sample only within the subgroup specific vector of errors. We find no meaningful differences in the manipulation for female and male students or students eligible and not eligible for free and reduced price lunch. In contrast, Figure 3 shows that 48.1 (se=0.62) percent of in-range core exams are manipulated for white and Asian students, compared to 43.7 (se=0.22) percent for black and Hispanic students. However, because black and Hispanic students have lower Regents scores on average, total manipulation for black and Hispanic students is nearly twice as large as the manipulation for white and Asian students. Thus, although black and Hispanic students are more likely to have a test score near a cutoff, and therefore be at risk for manipulation, white and Asian students are more likely to have their test scores manipulated if they happen to score just below a cutoff. In Panel D of Figure 3, we find that 8.0 (se=0.06) percent of all exams and 44.1 (se=0.27) of in-range exams are manipulated for students with below median 8th grade test scores (below 0.123 standard deviations). In contrast, 4.4 (se=0.05) percent of all exams and 44.9 (se=0.35) percent of in-range exams are manipulated for students with above median 8th grade test scores. Panel E of Figure 3 similarly shows students with either a behavioral violation or more than 20 absences are somewhat more likely to have a manipulated exam in general compared to students with no behavioral or attendance violations, but just over 3 percentage points less likely to have an in-range exam manipulated.

While we can easily measure differences in manipulation between students in different groups,

these will reflect the total of both within- and across-school variation in manipulation. To isolate across-school variation, and thus gauge the magnitude of the within-school component, we use a simple but intuitive Monte Carlo technique. We reassign characteristics randomly among students taking the same exam within each school, keeping the fraction of students with each subgroup designation constant both within schools and across all schools. We then we estimate manipulation for the randomly assigned synthetic subgroups, allowing the sixth order polynomial to be different in each synthetic subgroup. Finally, we repeat this entire process 100 times with different sixth order polynomials in each permutation in order to generate a distribution of total and in-range manipulation based on these synthetic subgroups. By comparing the actual differences across student groups reported in Figure 3 to the synthetic results which, by design, reflect only the across-school differences in manipulation, we can assess the magnitude of any within-school differences in manipulation. However, one limitation of this approach is that reassignment of students will not only lead to differences among students within the manipulable range, but will also influence all students and thus may influence the counterfactual distribution we estimate for a particular subgroup.

We report synthetic subgroup manipulation estimates in Table 3, alongside the actual subgroup estimates. Interestingly, the in-range manipulation advantage for whites and Asians is still present, though reduced by approximately 55 percent, when ethnicity is assigned randomly within schools. These results suggest that much of the white and Asian advantage is due to differences in manipulation across the schools these students attend, but that about 45 percent of the advantage is due to within-school differences in how white and Asian students are treated, conditional on having a score close to the cutoff.

Of course, any within-school difference in manipulation for white and Asian students may driven by other characteristics correlated with ethnicity. Indeed, we find no differences in in-range manipulation across students with high and low baseline test scores when this characteristic is assigned randomly within schools, nor do we see any difference in in-range manipulation when behavior and attendance records are assigned randomly within schools. These results suggests that nearly all of the advantage for students with high baseline test scores or good behavior records is due to within-school differences in the probability of having a test score manipulated.

In summary, we find that students with higher baseline test scores and better behavior with scores near the cutoff are more likely than their within-school peers to have their scores manipulated. This is consistent with teachers using soft information about students' true knowledge of the tested material, or some other measure of merit, when deciding to manipulate a score near the cutoff. We also find that schools serving black and Hispanic students have somewhat lower propensities to manipulate scores near the cutoff, although they have greater fractions of their students with scores in this range. Meanwhile, student gender and poverty are not correlated with the extent of manipulation.

C. The End of Manipulation: Estimates from 2011-2013

On February 12, 2011, the Wall Street Journal published an expose piece regarding manipulation on the Regents exams, including an analysis of state-wide data that reporters had obtained via a FOIA request and shared with the authors of this paper (see Dee et al. 2011 for additional details). The New York Times published a similar story and the results of its own analysis on February 19th, including a statement by a New York State Education Department official that anomalies in the Regents score distribution had been known for some time.[18] In May 2011, the New York State Board of Regents ordered schools to end the longstanding practice of re-scoring math and science exams with scores just below the proficiency cutoffs, and included explicit instructions on June 2011 exams in all subject areas specifying that "schools are no longer permitted to rescore any of the open-ended questions on this exam after each question has been rated once, regardless of the final exam score."[19] In October 2011, the Board of Regents further mandated that teachers would no longer be able to score their own students' state assessments as of the 2012-2013 school year.

In response to the state mandate that exams no longer be graded locally, the NYCDOE implemented a pilot program to grade various January 2012 and June 2012 core exams at centralized locations. Out of the 330 high schools in our sample offering Regents exams in 2012, 27 participated in the pilot program for the January exams, and 164 high schools participated for the June exams. Appendix Table 3 reports summary statistics for students taking a core Regents exam at pilot and non-pilot high schools in 2010-2011 – the year prior to the implementation of the pilot program. Students in pilot schools are more likely to be white and less likely to be Hispanic than students in non-pilot schools. However, there are no statistically significant differences for 8th grade test scores or performance on core Regents exams in the baseline period. None of the results suggest important differences between pilot and non-pilot schools before the implementation of the pilot program.[20] In 2013, all of New York City's high schools began using centralized scoring for all

[18] In 2009, the New York State Comptroller released results from an audit of local scoring practices. The report identified a number of specific shortcomings in scoring procedures, concluding that the oversight by the New York State Education Department was not adequate to assure the accuracy of Regents scores (DiNapoli 2009). The report also made clear that the Education Department had known about widespread scoring inaccuracies from periodic statewide reviews in which trained experts re-scored randomly selected exams from a sample of schools throughout the state. For example, a review of June 2005 exams found that 80 percent of the randomly re-scored exams received a lower score than the original, official score. For 34 percent of the re-scored exams, the difference in scoring was substantial – as much as 10 scale score points. The audit noted that an earlier audit during the 2003-2004 school year also found similar results.

[19] See for example: http://www.nysedregents.org/integratedalgebra/811/ia-rg811w.pdf. A different, perhaps cheaper means to eliminate manipulation, was recommended by Dee et al. (2011): withhold the algorithm that converts raw to scale scores until after schools had submitted student scores to the state. As the algorithm changed from year to year, schools could not easily identify students whose score was just below a cutoff. This recommendation was not taken by the New York State Education Department.

[20] In our discussions with NYCDOE officials, there was no specific formula used to select schools or particular targeting of schools with certain characteristics. We were informed that recruitment for the pilot was driven through high school "networks," i.e., mandatory but self-selected affiliations of 20-25 schools who work together to reduce administrative burdens through economies of scale. We find that network affiliation explains roughly 30 percent of pilot participation using random effects regressions. About half of the high schools in the NYCDOE share their building with another high school, and it is clear that co-located schools exhibited similar participation. Among the roughly one third of high schools that co-located in buildings with four or more high schools, we find that building location explains almost 90 percent of the variation in participation using random effects regressions.

Regents exams as mandated by the state.

In this section, we explore the implications of these swift, widespread, and arguably exogenous changes to the Regents grading policies on the extent of manipulation. Figure 4 plots the empirical distribution of test scores for core Regents exams taken in June between 2009-2010 and 2012-2013, the last year of data available. We focus on June exams to simplify the analysis given the staggered introduction of the pilot program, but results are identical using both January and June test administrations. We plot the results separately by participation in the 2012 pilot program to grade exams centrally. We also calculate manipulation only around the 65 cutoff, as the score of 55 was no longer a relevant cutoff for the vast majority of students in these cohorts (see Appendix Table 1). In June 2009-2010, pilot and non-pilot schools manipulated 72.9 (se=0.99) and 63.3 (se=0.54) percent of in-range exams, respectively.[21] Manipulation dropped to 17.2 (se=0.78) and 13.1 (se=0.39) percent of in-range exams in pilot and non-pilot schools, respectively, when schools were told to stop re-scoring math and science exams below the cutoff in June 2011. Thus, the extent of manipulation was greatly reduced, but clearly not eliminated, when state officials explicitly proscribed the practice of re-scoring exams with scores just below the proficiency cutoffs. Using the variation created by the pilot program, we find a clear role for the centralization of scoring in eliminating score manipulation. In June 2012, manipulation dropped from 17.2 percent to a statistically insignificant -0.8 percent (se=0.55) of in-range exams in pilot schools. Yet it remained fairly steady in non-pilot schools, whose exams were still graded by teachers within the high school, going from 13.1 percent to 12.8 percent (se=0.49). In line with these results, manipulation appears to have been completely eliminated in June 2013, when both pilot and non-pilot schools had adopted centralized grading. Of course, we cannot say with certainty whether centralization by itself would have eliminated manipulation in absence of the state's statements regarding re-scoring math and science exams, since we do not observe high schools operating under these conditions.

Appendix Figure 5 reports results separately for each core exam subject. The results are similar across all subjects, with the exception of Global History. These results are likely due to the fact that only 14 out of 163 schools pilot schools included Global History in the pilot program. In comparison, 159 schools included Integrated Algebra, 149 included Comprehensive English, 139 included Living Environment, and 134 included U.S. History. Our main results are therefore somewhat stronger if we drop Global History or limit the sample for each pilot school to the included subjects.

At the time that state and city policy changes eliminated the practice of score manipulation, it was unclear if this would have important long-term implications for students' academic achievement and attainment. After all, students whose exams would have been manipulated may simply have re-taken and passed the test shortly thereafter. Only now are we able to observe key outcomes, like high school graduation, for the cohorts of students potentially impacted by these policy changes. In the next section, we use these arguably exogenous policy changes to help identify the causal

[21] As can be seen in Appendix Figure 2, in-range manipulation in 2010 across both the 55 and 65 cutoffs remained at roughly 40 percent, in line with prior years. However, manipulation at the 55 cutoff had greatly decreased at this point, as this cutoff was no longer relevant for almost all students taking exams in 2010, while manipulation at the 65 cutoff was quite large.

impact of manipulation. Armed with these estimates, we then gauge the degree to which the longstanding practice of manipulation may have distorted levels and gaps in academic achievement among various groups of students.

IV. The Causal Effect of Test Score Manipulation on Educational Attainment

The swift, widespread elimination of manipulation following changes to Regents grading policies provides us with an identification strategy to estimate the causal impact of test score manipulation on a key longer term outcome, high school graduation, as well as significant intermediate outcomes, such as test re-taking and the number of years of high school education. Specifically, we can compare the outcomes of students with scores close to the cutoff before and after the policy changes. We can also control for any secular shifts in the characteristics of students in the "manipulable range" using prior trends, as well as parallel trends among students with scores just above this range in a differences-in-differences framework. In addition, we present results from a cross-sectional methodology similar to that used by Diamond and Persson (2016) as a secondary strategy to estimate the impact of manipulation on cohorts taking exams several years prior to the reforms. As we explain below, this cross-sectional strategy relies on stronger identification assumptions, but allows us to also examine college enrollment outcomes.

A. Difference-in-Differences Estimates

Our difference-in-differences approach exploits the sharp reduction in score manipulation following New York's policy changes starting in 2011. Intuitively we compare the evolution of outcomes for students with scores just inside the manipulable range, pre- vs. post-reform, to the evolution of outcomes for students with scores just above the manipulable range. The latter group of students helps us establish a counterfactual of what would have happened to the outcomes of students scoring in the manipulable range if the grading reforms had not been implemented. More specifically, we estimate the reduced form impact of the grading reforms using the following specification:

$$y_{isemth} = \alpha_{3h} + \alpha_{3et} + \alpha_{3s} + \alpha_{3s} \cdot \text{Year} + X_i\beta_3 + \gamma_3 \cdot \mathbf{1}[69 \geq \text{Score} \geq 60] \cdot \mathbf{1}[\text{Year} \geq 2011] + \phi_3 \cdot \mathbf{1}[59 \geq \text{Score} \geq 50] \cdot \mathbf{1}[\text{Year} \geq 2011] + \varepsilon_{isemth} \quad (3)$$

where y_{isemth} is the outcome of interest for student i with score s on exam e in month m and year t at high school h, α_{3h} are school effects, α_{et} are exam by year effects, α_s are 10-point Regent score effects, $\alpha_{3s}\cdot$ Year are linear trends in year interacted with Regents score bins to account for the increasingly stringent graduation requirements during this time period (see Appendix Table 1), and X_i includes controls for gender, ethnicity, free lunch eligibility, 8th grade test scores. We also control for the effect of the grading reforms on students scoring between 0-59 as these students are also likely to be affected by the reform if or when they retake the exam. We stack student outcomes across all core Regents exams and adjust our standard errors for clustering at both the student and

school level. Results are similar when we estimate (3) for each exam separately when there is only one observation per student (see Appendix Table 4).[22]

The parameter γ_3 can be interpreted as the differential effect of the reform on students scoring between 60-69 compared to the omitted group of students scoring between 70-100.[23] As the reform eliminated manipulation, we might expect our estimates of γ_3 to be negative for outcomes such as passing the exam and graduating from high school. However, the key identifying assumption is that in the absence of the Regents grading reforms (and conditional on our baseline controls and linear trends within 10-point score bins), any discontinuous change in outcomes for students scoring between 60-69 at the time of the reforms would have been identical to the change for students between 70-100. This assumption would be violated if the implementation of the grading reforms was coincident with unobservable changes in the types of students in each group. Below, we will present several tests in support of our approach.

We begin with a descriptive examination of how mean student outcomes evolved between 2004-2013 for those scoring between 60-69 (i.e. students likely to be affected by test score manipulation) and between 70-100 (i.e. students unlikely to be affected by test score manipulation). Figure 5 presents reduced form estimates from a variant of equation (3) that include the interaction of scoring between 60-69 and each pre-reform year so that we can examine any pre- and post-reform trends. We omit the interaction with 2010 so that all coefficients are relative to that omitted year. We also omit student baseline controls and the linear trends in year interacted with Regents score bins so that the coefficients simply show how outcomes evolve over time for students scoring between 60-69.[24] We also focus on students entering high school between 2003-2004 and 2009-2010 where we observe high school graduation. We measure graduation using an indicator for any diploma type at four years. We do not include GED diplomas in our graduation measure, but we do not examine GED separately as we cannot measure GED reliably in our data.

In Panel A of Figure 5, we examine the fraction of students scoring 65 or above on their first attempt on these exams. There is a sudden drop of around 18 percentage points in the probability of scoring 65 or above for students scoring between 60-69 following the implementation of the grading reforms in 2011, confirming, as we discussed in Section III.C, that the Regents grading reforms significantly decreased test score manipulation. In Panel B of Figure 5, we show that the fall in pass rates also coincides with a sharp increase in test retaking, from around 15 to 18 percent, suggesting that almost all marginal students who failed these Regents exams made a second attempt. In Panel C, we look at the rates of passing the exam within a full calendar year after the

[22]The effect of manipulation on high school graduation is largest for Living Environment, the first exam taken by most students. Effects are also somewhat larger for English and U.S. History, the last exams taken by most students. Note that the number of observations varies by subject in Appendix Table 4 because we do not observe every exam for every student. We observe 4.3 out of 5 core exams for the typical student in our sample.

[23]Results are similar if we drop students score between 0 and 59 or limit the comparison group to students scoring between 70-79.

[24]Appendix Figure 6 presents unadjusted means for students taking the English Language Arts or U.S. History Regents exam for the first time, as these subjects are typically taken in 11th grade as the last two core exams; this allows us to isolate cohorts of test takers likely (un)affected by the grading reforms in the raw data. The results largely follow those from Figure 5.

first attempt. We see a similar sudden drop, but of a smaller magnitude of roughly 6-7 percentage points, suggesting that the majority of these marginal students (but clearly not all of them) were able to pass the exam on a subsequent attempt. We cannot examine re-taking within a full calendar year for the 2013 exams as our data do not extend far enough. Importantly, we also observe clear upwards trends in the probability of scoring 65 or above during the pre-reform period. As discussed above, this upwards trend is consistent with the greater emphasis on the 65 point cutoff during this time period and suggests the inclusion of linear trends in our primary specification.

Panel D of Figure 5 shows that the coefficients on the interactions in 2005-2009 are all small, not statistically different from zero, and there is no trend in the coefficient values during this time period, suggesting similar pre-reform trends for students scoring between 60-69 and between 70-100. In Appendix Figure 6, we show that high school graduation rates were essentially flat for these two groups of students, with no indication of different pre-reform trends. However, starting in 2011, graduation rates suddenly drop by about 3 to 5 percentage points for students scoring between 60-69. Together, the data series shown in Figure 5 strongly suggest that the scoring reforms imposed by New York state had significant impact on students whose scores fell just below the 65 cutoff on the Regents core exams. While most of these students still eventually graduated, 25-30 percent of them appear to have been unable to pass the exam on a subsequent attempt and thus could not graduate from high school.

Regression estimates of equation (3) that also include baseline controls and linear trends are shown in Table 4, pooling all of the core Regents exams. We report the coefficient on the interaction between scoring between 60-69 on an exam and the exam being taken after the grading reforms were implemented in 2011. We also present results with and without school fixed effects, but these controls have very little impact on our estimates. First stage results (Columns 1 and 2) for effects on the probability of scoring 65 or higher are consistent with the patterns observed in Figure 5. The grading reforms are estimated to decrease the probability of scoring 65 or above by 15.9 (se=0.7) percentage points, a substantial decrease from the mean pass rate of 80 percent for students scoring between 60-69 in 2010. Reduced form estimates for high school graduation (Columns 3 and 4) indicate that students scoring between 60-69 are roughly 3.5 percentage points (se=0.4) less likely to graduate high school following the grading reforms.

In Columns 5-6 of Table 4, we present two-stage least squares estimates that provide the local average treatment effect of passing a Regents exam due to test score manipulation. Using this specification, we find that having a score inflated to fall just above the proficiency cutoff increases the probability of graduating from high school by 21.9 percentage points (se=2.9). This is a substantial effect, given an exam-weighted mean graduation rate of 79.8 percent for students in our sample. In other words, consistent with the patterns seen in Figure 5, we estimate that roughly a quarter of "marginal" Regents passers would not have graduated from high school if their scores had not been manipulated.

It is possible that the effects of manipulation were heterogeneous, and in Table 5 we report estimates from our preferred two-stage least squares specification for mutually exclusive student

subgroups. Effects on high school graduation are similar by ethnicity, but we find somewhat larger effects for female students, students from poor households, and students with higher 8th grade test scores. We estimate that manipulation increases the probability that female students graduate high school by 24.4 (se=3.4) percentage points, 5.1 percentage points more than male students. Manipulation also has a 4.4 percentage point larger effect on students eligible for free or reduced price lunch compared to students not eligible for free or reduced price lunch, and 5.9 percentage points larger for students with above median 8th grade test scores compared to students with below median scores.

Appendix Table 5 presents estimates using a variety of specifications and instruments to assess the robustness of our main two-stage least squares results. Column 1 uses the interactions of scoring between 60-69 and year-specific indicators for taking the test between 2011-2013 for a total of three instrumental variables. Column 2 adds an interaction with an indicator for participating in the centralized grading pilot program for a total of six instrumental variables. The estimated effect of manipulation on high school graduation ranges from 17.8 (se=2.5) to 19.1 (se=2.6) percentage points, and none of the point estimates are statistically distinguishable from our preferred estimates in Table 4.

In a further test for potential sources of bias in our main specification, we run placebo regressions where the dependent variable is a fixed student characteristic, rather than a student outcome. These estimates are shown in Panel A of Appendix Table 6. We do find several small but statistically significant "effects" of the reforms on student characteristics. However, if anything these results suggest that our estimates may be slightly biased against finding that the reforms lowered graduation rates for marginal students whose scores were no longer manipulated. We find that students scoring 60-69 following the elimination of re-scoring are approximately 1.5 (se=0.7) percentage points more likely to be eligible for free or reduced price lunch, but also have 8th grade test scores that are 0.053 (se=0.007) standard deviations higher following the grading reforms. To provide an indication of the magnitude of the potential bias due to these differences in baseline characteristics, we also examine differences in predicted high school graduation (using all of the baseline characteristics listed in Panel A of Appendix Table 6). The coefficient indicates that predicted graduation rates are 0.8 percentage points higher (se=0.1) for students scoring 60-69 following the elimination of re-scoring. Thus, while our difference-in-difference regression may contain some specification error, our analysis of student unobservables indicates that the true effect of test score manipulation on graduation rates may be even slightly larger than our two-stage least squares estimates suggest.

Most regents exams are taken well before the end of high school, and failing these exams may effect whether students continue to progress towards graduation or drop out of school. We therefore examine two additional measures of secondary school attainment: the highest grade (from 9 to 12) in which the student is enrolled in NYC public schools and the number of years the student is enrolled in NYC public schools. We select these two measures because they represent two opposing ways to address the issue of grade repetition, i.e., if a student is forced to repeat a grade, does this

repeated year of education represent additional educational attainment? If we measure attainment based on highest grade then repetition does not count as attainment, while if we measure based on years enrolled then repetition counts fully. Results from our preferred two-stage least squares specification on these outcomes are shown in Panel A of Appendix Table 7. We find large effects of manipulation on both of these attainment measures. Having an exam manipulated increases educational attainment by 0.41 grade levels (se=0.04), a 3.4 percent increase from the sample mean, and 0.54 school years (se=0.04), a 13.1 percent increase from the sample mean.[25] These results suggest that manipulation lengthened the extent of secondary education for marginal students, rather than just increase graduation rates for students who were already at the very end of their high school careers.

In Panel B of Appendix Table 7, we examine the quality of the high school degree. Information on diploma type (i.e., Regents vs. Advanced Regents) is not available over the entire sample period, so we measure the quality of the high school degree using indicator variables for whether a student has fulfilled all of the requirements for a specific degree type. Having an exam manipulated increases the probability of meeting the requirements for a Regents diploma, the lowest diploma for most students during this time period, by 33.8 (se=5.3) percentage points, a 54.7 percent increase from the sample mean. However, the probability of meeting the requirements for the Advanced Regents diploma decreases by 11.0 (se=5.2) percentage points, a 50.2 percent decrease from the sample mean. In Panel C, we show that having a score manipulated also modestly decreases the probability of meeting the two most important requirements for an Advanced Regents diploma: passing a physical science exam such as Chemistry or Physics, and passing an advanced math sequence that covers topics such as geometry and trigonometry. In results available upon request, we find that both the Advanced Regents and advanced coursework results are larger in students predicted to be near the margin of Advanced Regents receipt.

To shed further light on the mechanisms underlying this surprising result, Appendix Table 8 presents estimates separately by exam subject. The negative effect of manipulation on meeting the requirements for an Advanced Regents diploma are largest for the Living Environment and Math A/Integrated Algebra exams. In contrast, the effects of meeting the requirements for an Advanced Regents diploma are small and not statistically distinguishable from zero for the English and Social Science exams. These results are consistent with the idea that test score manipulation has somewhat heterogeneous effects. Students on the margin of dropping out are "helped" by test score manipulation because they are not forced to retake a class that may lead them to leave high school. Conversely, students on the margin of an Advanced Regents diploma may be "hurt" by test score manipulation because they are not pushed to re-learn the introductory material or re-take the introductory class that the more advanced coursework requires.[26]

[25] The sample mean is slightly higher than four years because many students in our sample repeat at least one high school grade. Ninth grade is the most commonly repeated grade.

[26] Manipulation on the Global History exam has a positive impact on passing a physical science or advanced math exam. This result may be because Global History is taken before these advanced science and math exams, and re-taking Global History therefore crowds out other courses. There is no effect of manipulation on English or U.S. History exams on passing either advanced science or math, likely because English and U.S. History are taken

B. Across-School Estimates

In addition to our main difference-in-difference results, we also present estimates from a cross-sectional methodology similar to that used by Diamond and Persson (2016).[27] This allows us to examine outcomes for students entering high school before 2005-2006 on whom we have information on the type of high school diploma awarded and college enrollment. While these cohorts are too old to be affected by the Regents grading reforms described above, we can use the variation across schools in the extent of in-range manipulation (from the random effects specification given by equation (2)) to provide estimates of the impact of manipulation on these outcomes. Specifically, we estimate the reduced form impact of attending a high manipulation school using the following specification:

$$y_{isemth} = \alpha_{4h} + \alpha_{4et} + \alpha_{4s} + \alpha_{4s} \cdot \text{Year} + X_i\beta_4 + \gamma_4 \cdot \mathbf{1}[69 \geq \text{Score} \geq 60] \cdot Manipulation_h + \phi_4 \cdot \mathbf{1}[59 \geq \text{Score} \geq 50] \cdot Manipulation_h + \varepsilon_{isemth} \quad (4)$$

where y_{isemth} is the outcome of interest for student i with score s on exam e in month m and year t at high school h, α_{4h} are school effects, α_{et} are exam by year effects, α_s are 10-point Regent score effects, $\alpha_{3s}\cdot$ Year are linear trends in year interacted with Regents score bins to account for the increasingly stringent graduation requirements during this time period (see Appendix Table 1), and X_i includes controls for gender, ethnicity, free lunch eligibility, 8th grade test scores. $Manipulation_h$ is an estimate of in-range manipulation for high school h from the random effects specification as described as Section III.A. To increase the precision of our estimates, we estimate $Manipulation_h$ only exams only around the 65 cutoff in the test administrations where we have information on high school graduation. Results are similar if we use a measure of manipulation $Manipulation_h$ that uses information across both cutoffs or all test administrations. When estimating equation (4), we stack student outcomes across all core Regents exams and adjust our standard errors for clustering at both the student and school level.

The parameter γ_4 can be interpreted as the differential impact of attending a "high" manipulation school for students scoring between 60-69 compared to other students at the same high school. The key identifying assumption is that in the absence of test score manipulation (and conditional on our baseline controls and school fixed effects), outcomes for students scoring between 60-69 would have been identical in high and low manipulation schools. This assumption would be violated if either students scoring between 60-69 at high and low manipulation schools are different in some unobservable way that is correlated with future outcomes, or if high and low manipulation schools

concurrently with the advanced science and math coursework at most high schools.

[27] Diamond and Persson (2016) estimate the impact of test score inflation in Sweden using across-school differences in the range of test scores manipulated interacted with an indicator for whether a student scores in the manipulable range. Thus, in their specification, students scoring in the manipulable range in "low" manipulation schools form the control group for students in the manipulable range in "high" manipulation schools. While our difference-in-differences specification instead uses post-reform data on students scoring in the manipulable range to form the control group for students scoring in the manipulable range prior to the grading reforms, our across-school specification broadly follows the Diamond and Persson (2016) approach.

differ in the way they educate students scoring between 60-69. For example, our approach would be invalid if high manipulation schools also spend more (or less) time educating students expected to score near a proficiency cutoff. Thus, this across-school strategy relies on much stronger identification assumptions than our difference-in-differences specification that also utilizes across-time variation from the Regents grading reforms. Nevertheless, placebo estimates on baseline characteristics and predicted outcomes broadly support our cross-sectional approach as long as school fixed effects are included (see Appendix Table 9).

Table 6 presents results from our secondary empirical strategy, where we use only across-school variation in in-range manipulation to examine outcomes for students entering high school between 2003-2004 and 2005-2006, for whom we observe both diploma type and college enrollment. We instrument for scoring 65 or above using the interaction of school in-range manipulation and scoring between 60-69. First stage results (Columns 2-3) show that a 10 percentage point increase in school-level manipulation increases the probability of scoring 65 or above by 4.0 (se=2.3) percentage points. In our preferred specification with school fixed effects, we find that having a score inflated to fall just above the 65 point proficiency cutoff increases the probability of graduating from high school by a statistically insignificant 4.6 percentage points (se=3.1). We show in Appendix Table 10 that the probability of graduating with a Regents diploma, the lowest diploma type available to students in our preferred difference-in-differences specification, increases by 50.2 (se=10.2) percentage points. In contrast, the probability of graduating with a local or Advanced Regents diploma decreases by 37.4 (se=6.7) and 8.3 (se=7.0) percentage points, respectively. Taken together with our above difference-in-differences estimates, these results suggest that having a score inflated to fall just above the 65 point proficiency cutoff increases the probability of receiving the diploma type associated with that cutoff, while decreasing the probability of receiving either more or less prestigious degrees. Results on advanced course taking also largely follow our difference-in-differences estimates, with manipulation decreasing the probability that a student passes an advanced science or math exam (see Panels B-C of Appendix Table 10).

We also find that having a score manipulated decreases the probability of enrolling in a two-year college by 6.7 (se=3.9) percentage points, but has little impact on the probability of enrolling in a four-year college. Unfortunately, we do not observe college graduation for these cohorts, but results for the number of years in college largely follow our enrollment results.

C. Implications

Our estimates from this section suggest that test score manipulation had economically important long-run effects on students. In light of the differential benefits of manipulation documented in Section III.B, our long-run estimates suggest that test score manipulation also had important distributional effects. To quantify these effects, we multiply the the two-stage least squares estimate of the impact of manipulation from Table 4 by the subgroup-specific total manipulation estimates from Figures 2 and 3.

These back-of-the-envelope calculations suggest that test score manipulation significantly af-

fected relative performance measures in New York City. For example, we estimate that the black-white gap in graduation rates would have increased from 15.6 percentage points to 16.3 percentage points in the absence of test score manipulation, while the graduation gap between high- and low-achieving students would have increased from 25.0 percentage points to 25.8 percentage points.

Our results also have important implications for aggregate graduation rates in New York. Our point estimates suggest that the fraction of students in our sample graduating from high school would have decreased from 76.6 percent to 75.3 percent without test score manipulation. The high school graduation rate is higher in our sample compared to the district as a whole (65.2 percent) because we drop students in special education, students in non-traditional high schools, and students without at least one core Regents score.

While we document important impacts of manipulation on educational attainment, data limitations prevent us from measuring impacts on labor market outcomes. A number of studies estimate significant positive returns to a high school diploma (e.g., Jaeger and Page 1996, Ou 2010, Papay, Willett, and Murnane 2011) and to additional years of schooling around the dropout age (e.g., Angrist and Krueger 1991, Oreopoulos 2007, Brunello et al. 2009). A recent study also finds positive returns to passing the Baccalaureate high school exit exam in France using a regression discontinuity design (Canaan and Mouganie 2015). Conversely, Clark and Martorell (2014) find negligible returns to passing "last chance" high school exit exams in the state of Texas, and Pischke and Von Wachter (2008) find zero returns to additional compulsory schooling in Germany. Our educational attainment effects should be interpreted with this broader body of work in mind.

V. Exploring Potential Explanations for Manipulation

We have shown that test score manipulation was widespread among schools in New York and that the practice had important, inequitable impacts on students' long-run outcomes. Test score manipulation appears to have been facilitated by both a formal policy to re-score math and science exams with scores just below proficiency cutoffs and the decentralized, school-based scoring of exams. In this question, we explore three additional reasons why the system-wide manipulation of Regents exams might have occurred before the grading reforms implemented in 2011.

Test-Based Accountability: There is a large literature documenting how schools may engage in various undesirable behaviors in response to formal test-based accountability systems (e.g., Figlio and Getzler 2002, Cullen and Reback 2002, Jacob 2005, Neal and Schanzenbach 2010, Neal 2013). It is therefore natural to ask whether the implementation of NCLB in the school year 2002-2003 and implementation of New York City's accountability system in 2007-2008, both based heavily on Regents exams, may have driven school staff to manipulate student exam results. Panel A of Figure 6 explores this hypothesis by plotting the distribution of core exams taken between 2001 and 2002, before the implementation of either school accountability system, and exams taken between 2008 and 2010, after the implementation of both accountability systems. Manipulation was clearly prevalent well before the rise of school accountability, with an estimated 60.7 percent (se=0.79) of

in-range exams manipulated before the implementation of these accountability systems, compared to the 44.6 percent (se=0.33) in the years after the implementation of these systems.[28]

To provide additional evidence on this issue, we take advantage of the fact that different schools face more or less pressure to meet the accountability standards during our sample period. Panel B of Figure 6 plots distribution of core exams for schools that did and did not make Adequate Yearly Progress (AYP) in the previous year under the NCLB accountability system, and Panel C of Figure 6 presents results for schools receiving an A or B grade compared to schools receiving a D or F in the previous year under the New York City accountability system. Consistent with our results from Panel A, we find no evidence that test score manipulation varied significantly with pressure from test-based accountability. Schools not meeting AYP manipulate 44.6 percent (se=0.26) of in-range exams, compared to 45.3 percent (se=0.69) for schools meeting AYP. Similarly, schools receiving a D or F from the NYC accountability system manipulate 44.2 percent (se=0.48) of in-range exams, compared to 43.0 percent (se=0.41) for schools receiving an A or B. Thus, we find no evidence that test-based school accountability systems are primary drivers of the manipulation documented above.[29]

Teacher Incentives: A closely related explanation for the system-wide manipulation of Regents exams is that teachers may benefit directly from high test scores even in the absence of accountability concerns. To test whether manipulation is sensitive to teacher incentives in this way, Panel D of Figure 6 plots the distribution of core Regents exam scores for schools participating in a randomized experiment that explicitly linked Regents scores to teacher pay for the 2007-2008 to 2009-2010 school years (Fryer 2013).[30] We find that control schools manipulated 44.7 percent (se=0.4) of in-range exams taken during the experiment, which is higher than our estimate of 41.2 percent (se=0.3) manipulated in treated schools. These results further suggest that manipulation is not driven by formal teacher incentives, at least not as implemented in New York City during this time period.

High School Graduation: A final explanation we consider is that teachers manipulate simply to permit students to graduate from high school, even if it is with the lowest diploma type available to them. To see whether manipulation is driven mainly by a desire just to get students over the bar for high school graduation, we examine the distribution of scores for optional tests that students take to gain greater distinction on their diploma. Appendix Figure 7 plots frequency distributions

[28] Results are similar if we exclude the math core exams that changed from Sequential Math 1 to Math A over this time period. Results are also similar if we exclude both the math and science core exams that required teachers to re-score exams close to the proficiency cutoffs.

[29] Dee et al. (2011) further examine the importance of school accountability using tests taken by 8th grade students. These are typically advanced students who wish to begin fulfilling their high school graduation requirements early. While the tests are still high stakes for students, they play no role in school accountability metrics for middle schools. Consistent with our results from Figure 6, we find that Regents scores are also manipulated for 8th grade students.

[30] The experiment paid treated schools up to $3,000 for every union-represented staff member if the school met the annual performance target set by the DOE. The performance target for high schools depended on student attendance, credit accumulation, Regents exam pass rates in the core subjects, and graduation rates. Fryer (2013) finds no effect of the teacher incentive program on teacher absences or and student attendance, behavior, or achievement. See Fryer (2013) for additional details.

for exams on exams in Chemistry, Physics and Math B (an advanced math exam). On all three exams we see clear patterns consistent with manipulation, particularly at the 65 cutoff, which does not support the idea that the goal of manipulation is mainly geared towards meeting basic graduation requirements. Using information from only the 65 point cutoff, we estimate that 3.4 percent (se=0.05) of these elective Regents exams were manipulated in total, and that 37.4 percent (se=0.3) were manipulated among those with scores within the range just below the cutoff. The latter is only a few percentage points less than the amount of in-range manipulation for core Regents exams.

In sum, these estimates suggest that manipulation was unrelated to the incentives created by school accountability systems, formal teacher incentive pay programs, or concerns about high school graduation. Instead, it seems that the manipulation of test scores may have simply been a widespread "cultural norm" among New York high schools, in which students were often spared any sanctions involved with failing exams, including retaking the test or being ineligible for a more advanced high school diploma. It is of course possible that a more specific cause of the manipulation may be uncovered, but, perhaps due to limitations in our data, we are unable to do so. For example, we do not have information on the specific administrators and teachers responsible for grading each exam. Perhaps with this information, one might be able to identify specific individuals whose behavior drives this practice.

VI. Conclusion

In this paper, we show that the design and decentralized, school-based scoring of New York's high-stakes Regents Examinations led to the systematic manipulation of student test scores just below important performance cutoffs. We find that approximately 40 percent of student test scores near the performance cutoffs are manipulated. Exploiting a series of reforms that sharply reduced test score manipulation, we find that manipulating the test score of a student who would have failed the test by a small margin has a substantial impact on his or her probability of graduating from high school, raising it by approximately 21.9 percentage points or about 27.4 percent. We also find evidence consistent with the manipulation being driven by teachers' desire to help their students receive these benefits of passing an exam, not the recent creation of school accountability systems or formal teacher incentive programs.

Our findings suggest that test score manipulation had important effects on the relative performance of students across and within New York public schools. Our estimates imply, for example, that the black-white gap in graduation rates would have increased from 15.6 percentage points to 16.3 percentage points in the absence of test score manipulation, while the overall graduation rate would have decreased from 76.6 percent to 75.3 percent without test score manipulation.

An important limitation of our analysis is that we are only able to estimate the effect of eliminating manipulation in partial equilibrium. There may also be important general equilibrium effects of test score manipulation that we are unable to measure using our empirical strategy. For example, it is possible that widespread manipulation may change the way schools teach students expected

to score near proficiency cutoffs. It is also possible that test score manipulation can change the signaling value of course grades or a high school diploma. Estimating these impacts remains an important area for future research.

References

[1] Angrist, Joshua, and Alan Krueger. 1991. "Does Compulsory Schooling Attendance Affect Schooling and Earnings?" Quarterly Journal of Economics, 106(4): 979-1014.

[2] Angrist, Joshua, Erich Battistin, and Daniela Vuri. 2014. "In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno." NBER Working Paper 20173.

[3] Beadie, Nancy. 1999. "From Student Markets to Credential Markets: The Creation of the Regents Examination System in New York State, 1864-1890." History of Education Quarterly, 39(1): 1-30.

[4] Brunello, Giorgio, Margherita Fort, and Guglielmo Weber. 2009. "Changes in Compulsory Schooling, Education and the Distribution of Wages in Europe." The Economic Journal, 119(536): 516-539.

[5] Burgess, Simon and Ellen Greaves. 2013. "Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities." Journal of Labor Economics, 31(3): 535-576.

[6] Canaan, Serena, and Mouganie, Pierre. 2015. "Returns to Education Quality for Low-Skilled Students: Evidence from a Discontinuity." Unpublished Working Paper.

[7] Chetty, Raj, John Friedman, Tore Olsen, and Luigi Pistaferri. 2011. "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." The Quarterly Journal of Economics, 126(2): 749-804.

[8] Chudowsky, Naomi, Nancy Kober, Keith Gayler, and Madlene Hamilton. 2002. "State High School Exit Exams: A Baseline Report." Center on Education Policy, Washington DC.

[9] Clark, Damon, and Paco Martorell. 2014. "The Signaling Value of a High School Diploma." Journal of Political Economy, 122(2): 282-318.

[10] Cullen, Julie and Randall Reback. 2002. "Tinkering Toward Accolades: School Gaming under a Performance Accountability System." NBER Working Paper No. 12286.

[11] Dee, Thomas, Brian Jacob, Justin McCrary, and Jonah Rockoff. 2011. "Rules and Discretion in the Evaluation of Students and Schools: The Case of the New York Regents Examinations." Mimeo.

[12] Diamond, Rebecca and Petra Perrson. 2016. "The Long Term Consequences of Grade Inflation." Unpublished Working Paper.

[13] DiNapoli, Thomas. 2009. "Oversight of Scoring Practices on Regents Examinations." Office of the New York State Comptroller. Report 2008-S-151.

[14] Ebenstein, Avraham, Victor Lavy, and Sefi Roth. "The Long Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution." Forthcoming in American Economic Journal: Applied.

[15] Figlio, David and Lawrence Getzler. 2002. "Accountability, Ability and Disability: Gaming the System?" NBER Working Paper No. 9307.

[16] Fryer, Roland. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." Journal of Labor Economics, 31(2): 373-427.

[17] Hanna, Rema, and Leigh Linden. 2012. "Discrimination in Grading." American Economic Journal: Economic Policy, 4(4): 146-168.

[18] Hinnerich, Bjorn, Erik Hoglin and Magnus Johannesson. 2011. "Are Boys Discriminated in Swedish High School?" Economics of Education Review. 30(4): 682-690.

[19] Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: Evidence from School Reform in Chicago." Journal of Public Economics, 89(5-6): 761-769.

[20] Jacob, Brian A. and Steven Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." Quarterly Journal of Economics, 118(3): 843-877.

[21] Jacob, Brian A. and Jesse Rothstein. "The Measurement of Student Ability in Modern Assessment Systems." Forthcoming in Journal of Economic Perspectives.

[22] Jaeger, David A., and Marianne E. Page. 1996. "Degrees Matter: New Evidence on Sheepskin Effects in the Returns to Education." Review of Economics and Statistics, 77(4): 733-739.

[23] Kleven, Henrik and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." Quarterly Journal of Economics, 128(2): 669-723.

[24] Lavy, Victor. 2008. "Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment." Journal of Public Economics, 92: 2083-2105.

[25] Lavy, Victor. 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics." American Economic Review, 99(5): 1979-2011.

[26] Lavy, Victor and Edith Sand. 2015. "On the Origins of Gender Human Capital Gaps: Short and Long Term Consequences of Teachers' Stereotypical Biases." NBER Working Paper No. 20909.

[27] Neal, Derek and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." Review of Economics and Statistics, 92(2): 263-283.

[28] Neal, Derek. 2013. "The Consequences of Using One Assessment System to Pursue Two Objectives." NBER Working Paper 19214.

[29] New York State Education Department. 2008. History of Elementary, Middle, Secondary & Continuing Education. http://www.regents.nysed.gov/about/history-emsc.html, last updated November 25, 2008, accessed January 29, 2011.

[30] New York State Education Department. 2009. Information Booklet for Scoring the Regents Examination in English. Albany, NY.

[31] New York State Education Department. 2010. General Education & Diploma Requirement, Commencement Level (Grades 9-12). Office of Elementary, Middle, Secondary, and Continuing Education. Albany, NY.

[32] New York State United Teachers. 2010. "NYS Education Department Approved Alternatives to Regents Examinations." Research and Education Services, 10-02.

[33] Oreopoulos, Philip. 2007. "Do Dropouts Drop Out Too Soon? Wealth, Health, and Happiness from Compulsory Schooling." Journal of Public Economics, 91 (11-12), 2213-2229.

[34] Ou, Dongshu. 2010. "To Leave or Not to Leave? A Regression Discontinuity Analysis of the Impact of Failing the High School Exit exam." Economics of Education Review, 29(2): 171-186.

[35] Papay, John P., John B. Willett, and Richard J. Murnane. 2011. "Extending the Regression-Discontinuity Approach to Multiple Assignment Variables." Journal of Econometrics, 161(2): 203-207.

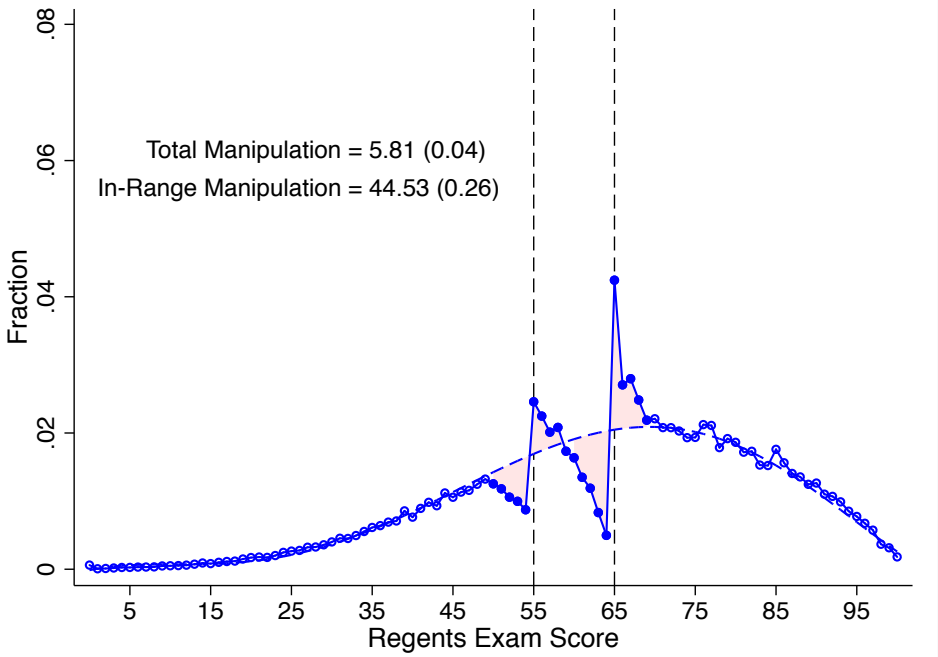
[36] Persson, Petra. 2015. "Social Insurance and the Marriage Market." Institute for Evaluation of Labour Market and Education Policy, Working Paper Series No. 2015:6.

[37] Pischke, Jorn-Steffen, and Till Von Wachter. 2008. "Zero Returns to Compulsory Schooling in Germany: Evidence and Interpretation." The Review of Economics and Statistics, 90(3): 592-598.

[38] Rockoff, Jonah and Lesley Turner. 2010. "Short Run Impacts of Accountability on School Quality." American Economic Journal: Economic Policy, 2(4): 119-147.

[39] Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" American Economic Journal: Economic Policy, 2(3): 180-212.

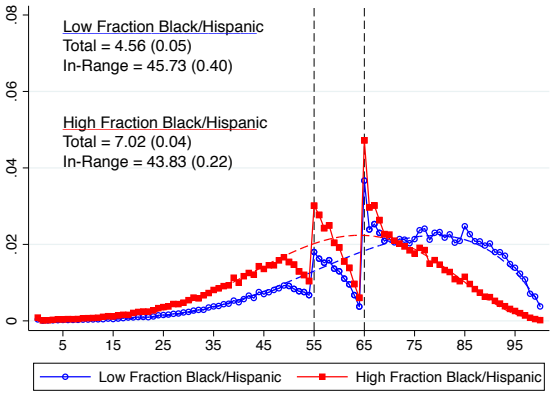
Figure 1
Test Score Distributions for Core Regents Exams, 2004-2010



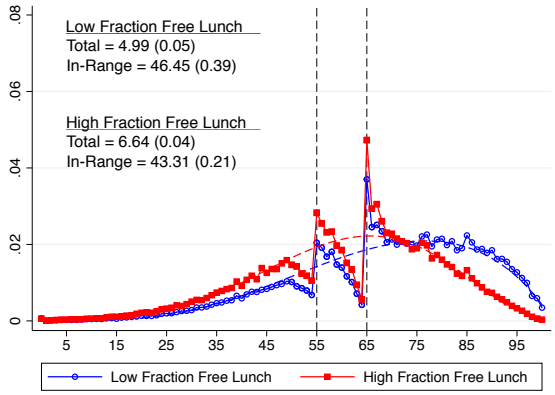
Notes: This figure shows the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Core exams include English Language Arts, Global History, U.S. History, Math A/Integrated Algebra, and Living Environment. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near each cutoff. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

Figure 2
Results by School Characteristics, 2004-2010

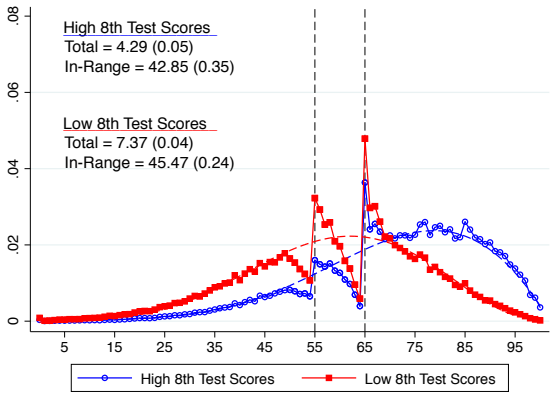
(a) High vs. Low Black/Hispanic Enrollment



(b) High vs. Low Free Lunch Enrollment



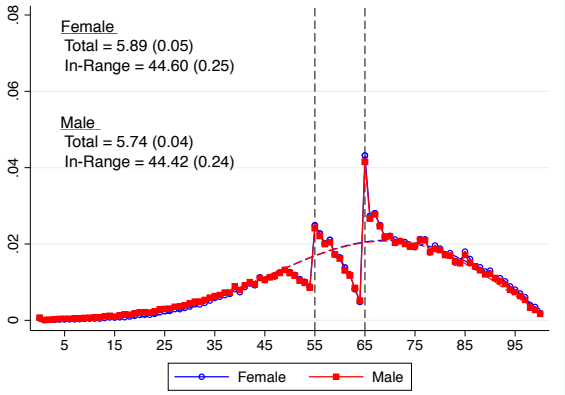
(c) Low vs. High 8th Test Scores



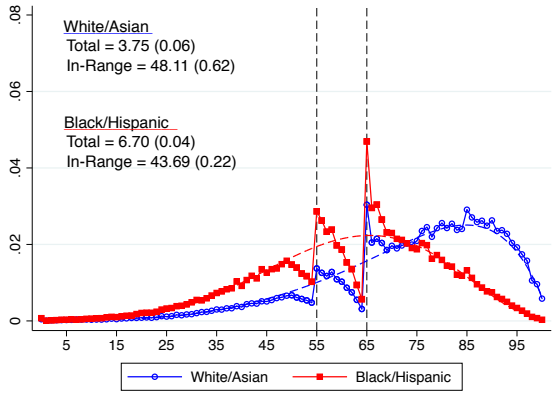
Notes: These figures show the test score distribution for core Regents exams around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Panel (a) considers exams taken in schools in the lowest and highest quartiles of the fraction of black/Hispanic students. Panel (b) considers exams taken in schools in the lowest and highest quartiles of the fraction of free lunch students. Panel (c) considers exams taken in schools in the lowest and highest quartiles of average 8th grade test scores. See the Figure 1 notes for additional details on the sample and empirical specification.

Figure 3
Results by Student Characteristics, 2004-2010

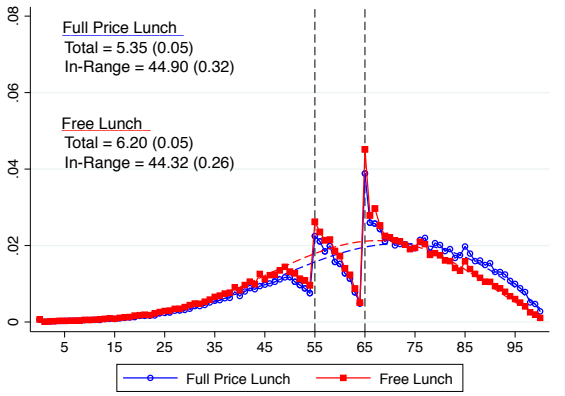
(a) Female vs. Male



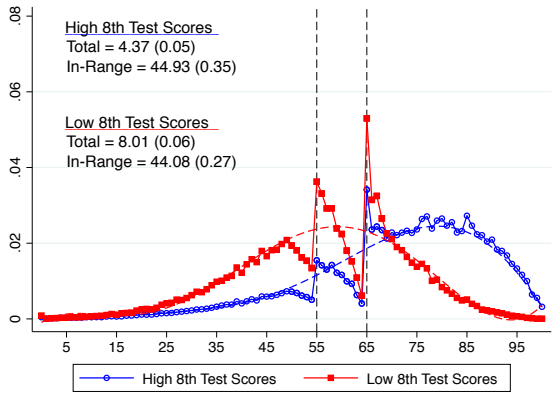
(b) White/Asian vs. Black/Hispanic



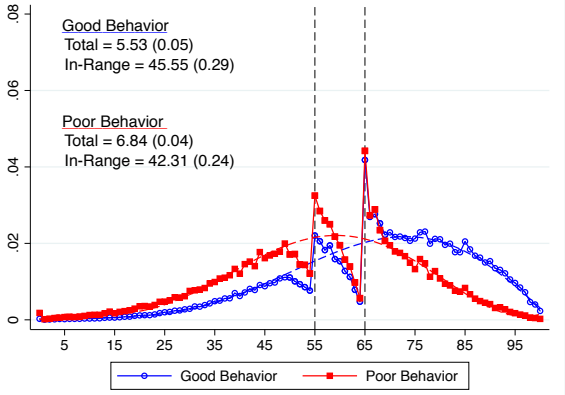
(c) Full Price vs. Free Lunch



(d) Low vs. High 8th Test Scores



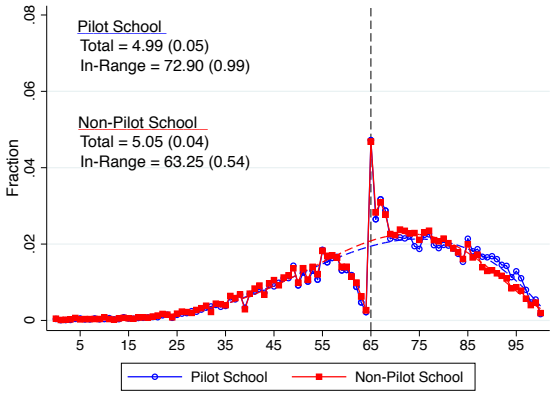
(e) Good vs. Poor Behavior



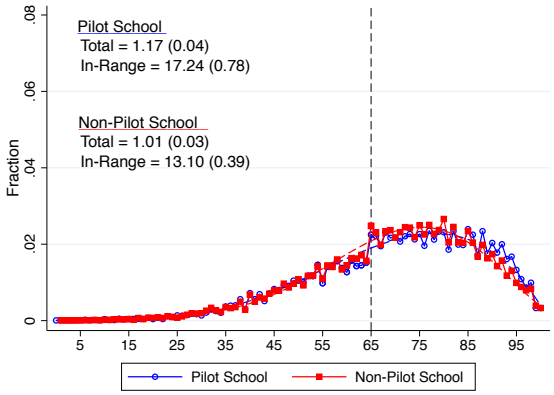
Notes: These figures show the test score distribution for core Regents exams around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Panel (a) considers exams taken by female and male students. Panel (b) considers exams taken by white/Asian and black/Hispanic students. Panel (c) considers exams taken by full price and free or reduced price lunch students. Panel (d) considers exams taken by students in the lower and upper quartiles of the 8th grade test score distribution. Panel (e) considers exams taken by students with both fewer than 20 absences and no disciplinary incidents and students with either more than 20 absences or a disciplinary incident. See the Figure 1 notes for additional details on the sample and empirical specification.

Figure 4
Test Score Distributions Before and After Grading Reforms, 2010-2013

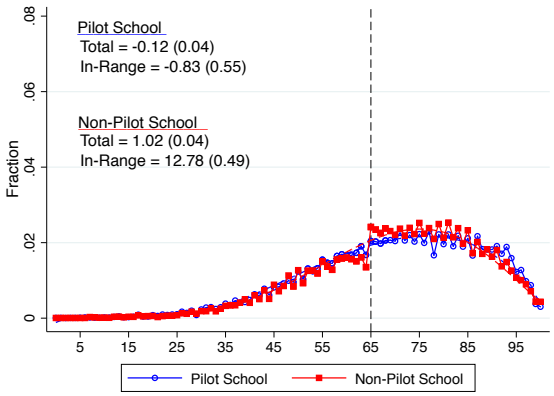
(a) Re-Scoring and Decentralized Grading in All Schools



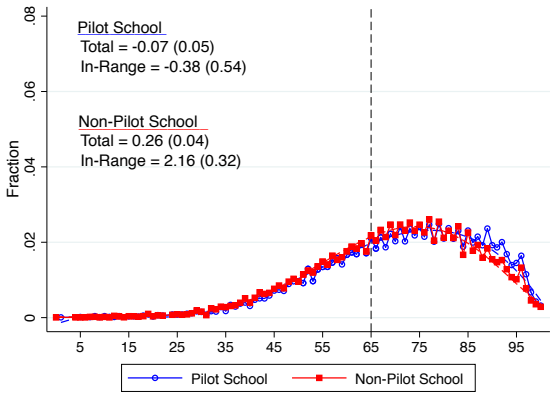
(b) No Re-Scoring and Decentralized Grading in All Schools



(c) No Re-Scoring in All and Pilot of Centralized Grading



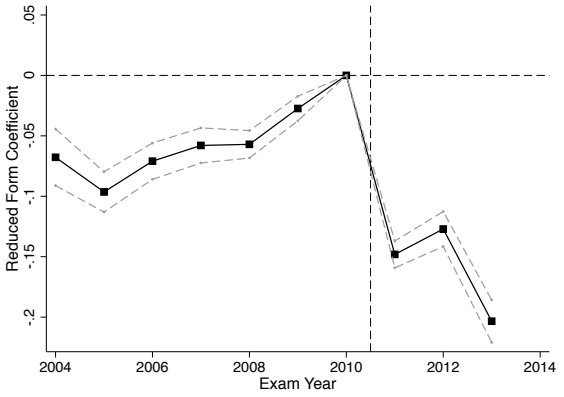
(d) No Re-Scoring and Centralized Grading in All Schools



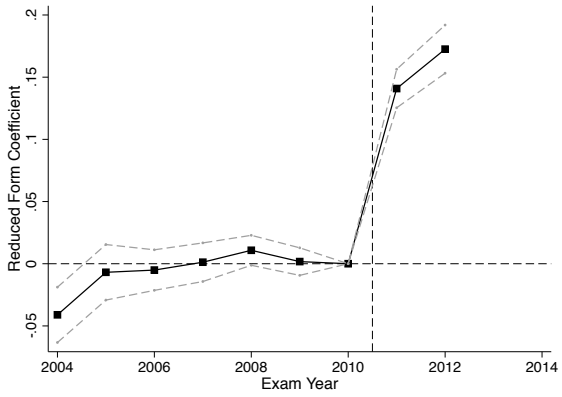
Notes: These figures show the test score distribution around the 65 score cutoff for New York City high school test takers between 2010-2013 in June. Included core exams include English Language Arts, Global History, U.S. History, Integrated Algebra, and Living Environment. Panel (a) considers exams taken in 2010 when re-scoring was allowed and grading was decentralized in both pilot and non-pilot schools. Panel (b) considers exams taken in 2011 when re-scoring was not allowed and grading was decentralized in both pilot and non-pilot schools. Panel (c) considers exams taken in 2012 when re-scoring was not allowed and grading was centralized in pilot schools but decentralized in the non-pilot schools. Panel (d) considers exams taken in 2013 when re-scoring was not allowed and grading was centralized in both pilot and non-pilot schools. See the Figure 1 notes for additional details on the sample and empirical specification.

Figure 5
Regents Grading Reforms and Student Outcomes

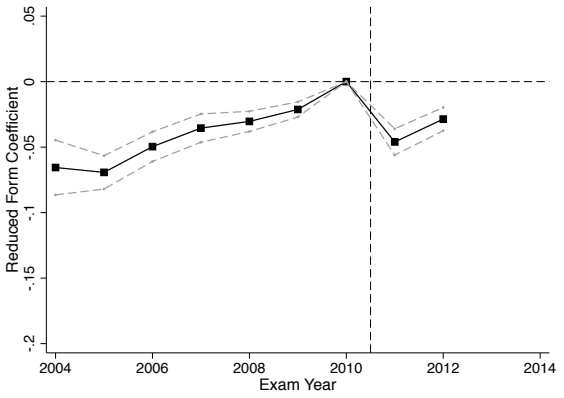
(a) Score 65+ on First Administration



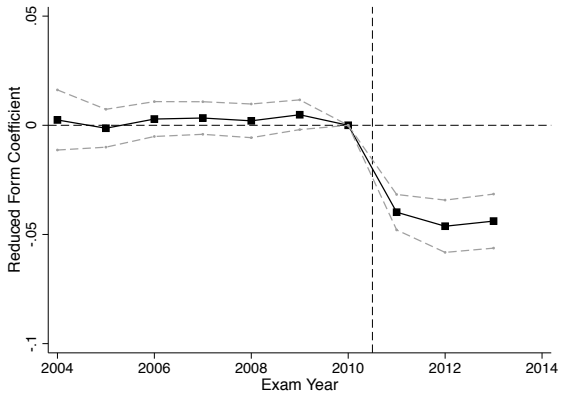
(b) Retake Exam in First Calendar Year



(c) Score 65+ in First Calendar Year



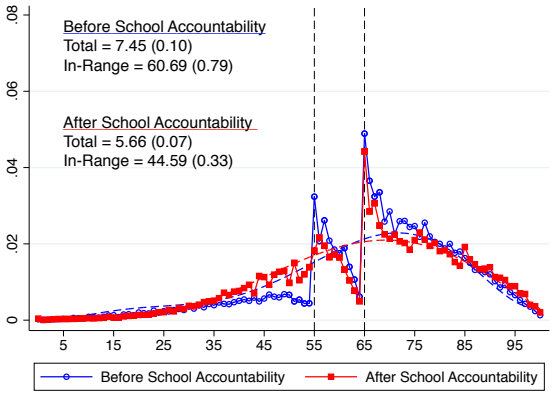
(d) Graduate High School



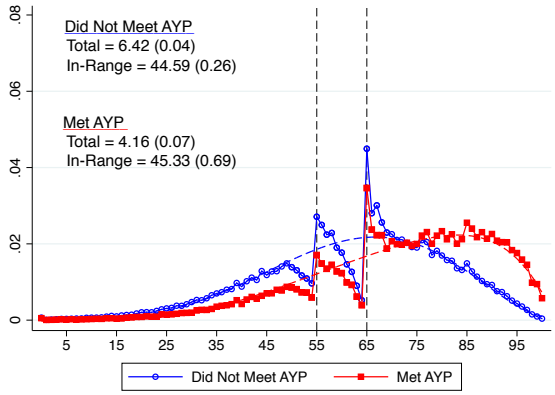
Notes: These figures plot the reduced form impact of the Regents grading reforms on high school graduation. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. We report reduced form results using the interaction of taking the test in the indicated year and scoring between 60-69. We control for an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects, and exam by year-of-test effects. See the Table 4 notes for additional details.

Figure 6
Results by School Accountability Pressure, 2004-2010

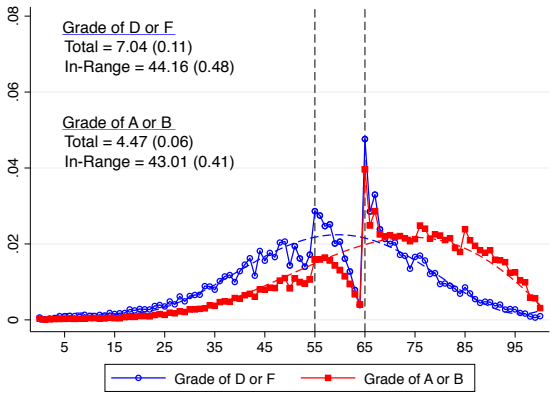
(a) Before and After School Accountability



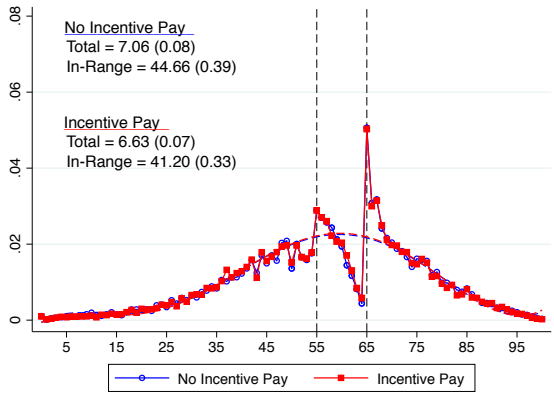
(b) NCLB Adequate Yearly Progress



(c) NYC Accountability Grade



(d) Teacher Incentive Pay Experiment



Notes: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers. Panel (a) plots non-math core exams taken in 2000-2001 before the implementation of NCLB and the NYC Accountability System and in 2008-2010 after the implementation of both accountability systems. Panel (b) plots all core exams for schools that in the previous year did not make AYP under NCLB and schools that did make AYP under NCLB for 2004-2010. Panel (c) plots all core exams for schools that in the previous year received a NYC accountability grade of A or B and schools that received a NYC accountability grade of D or F for 2008-2010. Panel (d) plots all core exams for schools in the control and treatment groups of an experiment that paid teachers for passing Regents scores for 2008-2010. See the Figure 1 notes for additional details on the empirical specification and the data appendix for additional details on the sample and variable definitions.

Table 1
Summary Statistics

	Full Sample	All Exams 0-49	1+ Exam 50-69	All Exams 70-100
Characteristics:	(1)	(2)	(3)	(4)
Male	0.477	0.525	0.477	0.467
White	0.145	0.055	0.096	0.243
Asian	0.165	0.061	0.103	0.285
Black	0.331	0.414	0.387	0.223
Hispanic	0.353	0.461	0.407	0.243
Free Lunch	0.552	0.601	0.589	0.483
8th Grade Test Scores	0.225	-0.714	-0.106	0.906
Core Regents Performance:				
Comprehensive English	69.417	29.892	63.109	85.255
Living Environment	69.622	38.831	63.215	82.725
Math A	69.830	40.453	65.038	84.520
Int. Algebra	66.052	40.830	61.947	79.990
U.S. History	72.496	32.995	65.201	88.994
Global History	67.814	32.560	60.166	86.376
High School Graduation:				
High School Graduate	0.730	0.150	0.685	0.910
Local Diploma	0.261	0.116	0.370	0.101
Regents Diploma	0.270	0.009	0.255	0.352
Advanced Regents Diploma	0.168	0.001	0.039	0.432
College Enrollment:				
Any College	0.502	0.133	0.437	0.690
Years College	1.208	0.242	0.907	1.929
Any Two-Year College	0.188	0.095	0.235	0.125
Years Two-Year College	0.311	0.152	0.383	0.216
Any Four-Year College	0.372	0.051	0.260	0.631
Years Four-Year College	0.898	0.091	0.524	1.714
Students	514,679	36,692	295,301	182,686

Notes: This table reports summary statistics for students in New York City taking a core Regents exam between 2004-2010. High school diploma and college enrollment records are only available for cohorts entering high school before 2004-2005. Enrollment, test score, and high school graduation information comes from Department of Education records. College enrollment information comes from the National Student Clearinghouse. Column 1 reports mean values for the full estimation sample. Column 2 reports mean values for students with all Regents score less than 50. Column 3 reports mean values for students with at least one Regents score between 50 and 69. Column 4 reports mean values for students with all Regents scores 70 or above. See the data appendix for additional details on the sample construction and variable definitions.

Table 2
School Manipulation and School Characteristics

	Total Manipulation				In-Range Manipulation			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Percent Black/Hispanic	0.028***			0.018**	0.228***			0.178*
	(0.006)			(0.008)	(0.077)			(0.095)
Percent Free Lunch		0.014*		−0.004		0.125		−0.019
		(0.007)		(0.007)		(0.084)		(0.080)
8th Test Score Percentile			−0.020***	−0.015***			−0.132***	−0.076
			(0.004)	(0.004)			(0.046)	(0.054)
Observations	10,153	10,153	10,153	10,153	10,153	10,153	10,153	10,153
Dep. Variable Mean	3.703	3.703	3.703	3.703	46.906	46.906	46.906	46.906

Notes: This table reports OLS estimates of school manipulation on school characteristics. School-level manipulation is estimated using all New York City high school test takers between 2004-2010. All specifications above use the number of exams in range of manipulation at each school as weights. See the data appendix for additional details on the sample construction and variable definitions.

Table 3
Student Subsample Results

	Total Manipulation		In-Range Manipulation	
	True Subgroup	Synthetic Subgroup	True Subgroup	Synthetic Subgroup
Panel A: Gender	(1)	(2)	(3)	(4)
Female	5.89	5.74	44.60	44.01
	(0.05)	(0.02)	(0.25)	(0.16)
Male	5.74	5.84	44.42	44.85
	(0.04)	(0.03)	(0.24)	(0.18)
Difference	0.15	−0.10	0.18	−0.84
	(0.06)	(0.05)	(0.42)	(0.34)
Panel B: Ethnicity				
White/Asian	3.75	4.32	48.11	45.90
	(0.06)	(0.03)	(0.62)	(0.31)
Black/Hispanic	6.70	6.45	43.69	43.92
	(0.04)	(0.01)	(0.22)	(0.08)
Difference	−2.95	−2.13	4.42	1.98
	(0.06)	(0.05)	(0.63)	(0.39)
Panel C: Free Lunch Eligibility				
Full Price Lunch	5.35	5.42	44.90	45.19
	(0.05)	(0.02)	(0.32)	(0.19)
Free Lunch	6.20	6.10	44.32	43.88
	(0.05)	(0.02)	(0.26)	(0.13)
Difference	−0.85	−0.68	0.59	1.31
	(0.07)	(0.04)	(0.43)	(0.32)
Panel D: 8th Test Scores				
Above Median 8th Scores	4.37	5.18	44.93	44.42
	(0.05)	(0.02)	(0.35)	(0.16)
Below Median 8th Scores	8.01	6.80	44.08	44.38
	(0.06)	(0.03)	(0.27)	(0.20)
Difference	−3.64	−1.62	0.84	0.04
	(0.08)	(0.05)	(0.43)	(0.36)
Panel E: Behavior and Attendance				
Good Attendance/Behavior	5.53	5.57	45.55	44.48
	(0.05)	(0.01)	(0.29)	(0.10)
Poor Attendance/Behavior	6.84	6.56	42.31	44.30
	(0.04)	(0.05)	(0.24)	(0.31)
Difference	−1.31	−0.99	3.25	0.18
	(0.06)	(0.06)	(0.38)	(0.41)

Notes: This table reports subsample estimates of test score manipulation by student characteristics. Columns 1 and 3 report results using actual student characteristics. Columns 2 and 4 report results with randomly assigned synthetic student characteristics. We hold the the fraction of students with each characteristic constant within each school when creating synthetic subgroups. See the text for additional details.

Table 4
Effect of Test Score Manipulation on High School Graduation

	First Stage		Reduced Form		2SLS	
	(1)	(2)	(3)	(4)	(5)	(6)
Graduate High School	−0.158***	−0.159***	−0.032***	−0.035***	0.204***	0.219***
	(0.007)	(0.007)	(0.004)	(0.004)	(0.029)	(0.029)
Observations	1,696,873	1,696,873	1,696,873	1,696,873	1,696,873	1,696,873
Dep. Variable Mean	0.712	0.712	0.798	0.798	0.798	0.798
Student Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	No	Yes	No	Yes	No	Yes

Notes: This table reports estimates of test score manipulation on student outcomes. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. Columns 1-2 report first stage results from a regression of an indicator for scoring 65+ on the first administration on the interaction of taking the test between 2011-2013 and scoring between 60-69. Columns 3-4 report reduced form results using the interaction of taking the test between 2011-2013 and scoring between 60-69. Columns 5-6 report two-stage least squares results using the interaction of taking the test between 2011-2013 and scoring between 60-69 as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects interacted with year-of-test, and exam by year-of-test effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Table 5
High School Graduation Effects by Student Subgroup

	Male	Female	Black/ Hispanic	White/ Asian	Free Lunch	Full Price Lunch	Low 8th Score	High 8th Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Graduate High School	0.193***	0.244***	0.205***	0.189***	0.227***	0.183***	0.116***	0.175***
	(0.037)	(0.034)	(0.028)	(0.052)	(0.030)	(0.052)	(0.033)	(0.044)
Observations	799,273	897,587	1,130,205	549,830	1,022,187	674,504	644,048	755,658
Dep. Variable Mean	0.768	0.824	0.754	0.889	0.797	0.798	0.674	0.917
Student Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports two-stage least squares estimates of the effect of test score manipulation by student subgroup. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. We use the interaction of taking the test between 2011-2013 and scoring between 60-69 as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects interacted with year-of-test, and exam by year-of-test effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

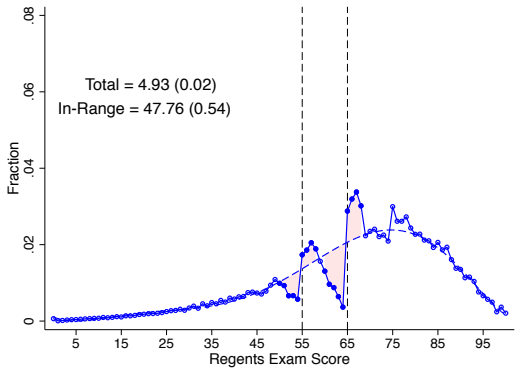
Table 6
Effect of Test Score Manipulation on Student Outcomes using Across-School Variation

	Sample Mean	First Stage		Reduced Form		2SLS	
Panel A: High School Graduation	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Graduate High School	0.756	0.399***	0.400***	0.028**	0.018	0.070**	0.046
	(0.430)	(0.023)	(0.023)	(0.012)	(0.013)	(0.030)	(0.031)
Panel B: College Enrollment							
Any College	0.505	0.399***	0.400***	−0.018	−0.020	−0.044	−0.049
	(0.500)	(0.023)	(0.023)	(0.017)	(0.014)	(0.042)	(0.036)
Any Two-Year College	0.189	0.399***	0.400***	−0.020	−0.027*	−0.050	−0.067*
	(0.392)	(0.023)	(0.023)	(0.016)	(0.016)	(0.040)	(0.039)
Any Four-Year College	0.350	0.399***	0.400***	−0.001	0.003	−0.001	0.006
	(0.477)	(0.023)	(0.023)	(0.015)	(0.013)	(0.038)	(0.032)
Observations	587,116	587,116	587,116	587,116	587,116	587,116	587,116
Student Controls	–	Yes	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	–	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	–	No	Yes	No	Yes	No	Yes

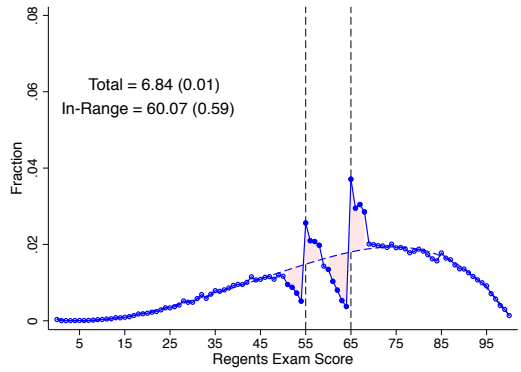
Notes: This table reports estimates of test score manipulation on student outcomes that use across-school variation in manipulation. The sample includes students entering high school between 2003-2004 and 2005-2006 and taking core Regents exams between 2004-2010. Columns 2-3 reports first stage results from a regression of an indicator for scoring 65+ on the first administration on the interaction of school in-range manipulation and scoring between 60-69. Columns 4-5 reports reduced form results using the interaction of school in-range manipulation and scoring between 60-69. Columns 6-7 reports two-stage least squares results using the interaction of school in-range manipulation and scoring between 60-69 as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects interacted with year-of-test, and exam by year-of-test effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Figure 1
Results by Subject, 2004-2010

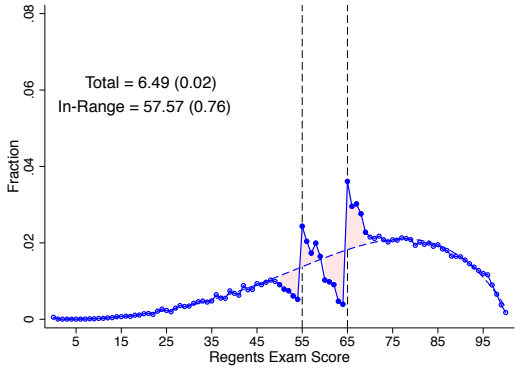
(a) English Language Arts, 2004-2010



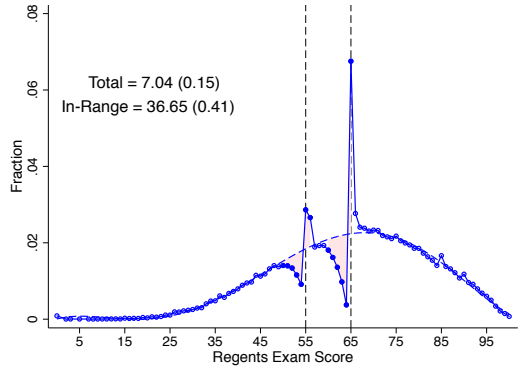
(b) Global History, 2004-2010



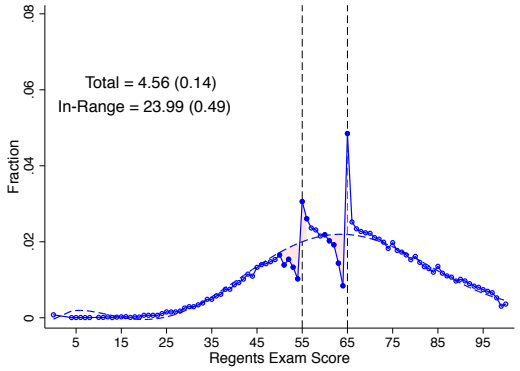
(c) U.S. History, 2004-2010



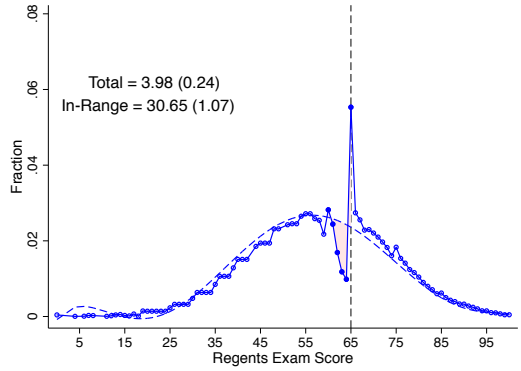
(d) Living Environment, 2004-2010



(e) Math A, 2004-2008



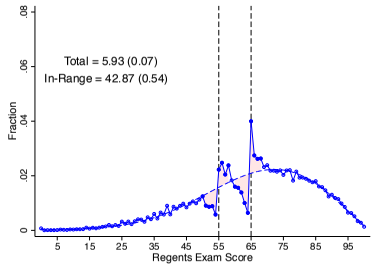
(f) Integrated Algebra, 2008-2010



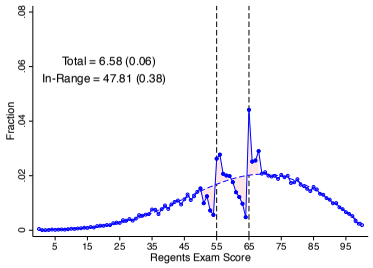
Notes: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near each cutoff. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

Appendix Figure 2
Results by Year

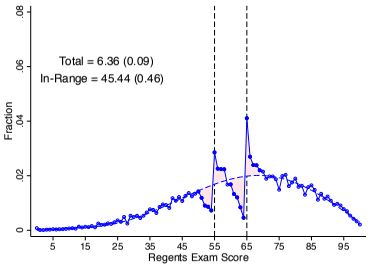
(a) 2004 Core Exams



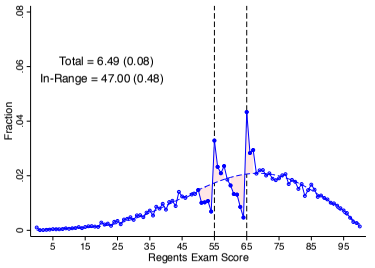
(b) 2005 Core Exams



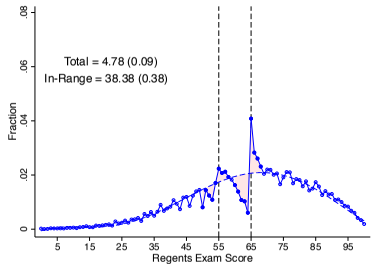
(c) 2006 Core Exams



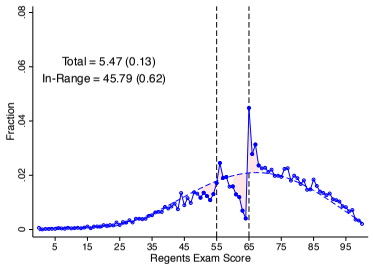
(d) 2007 Core Exams



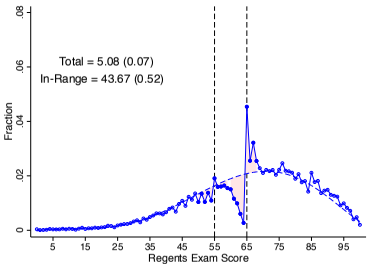
(e) 2008 Core Exams



(f) 2009 Core Exams



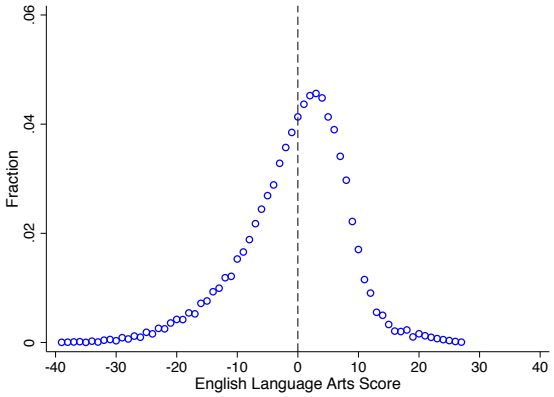
(g) 2010 Core Exams



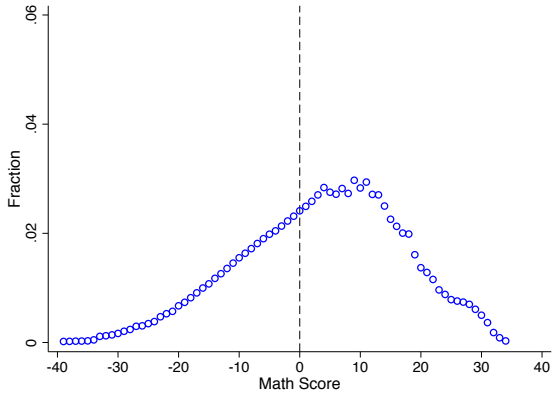
Notes: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near each cutoff. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

Appendix Figure 3
Test Score Distributions for Centrally Graded Exams in Grades 3-8

(a) English Language Arts



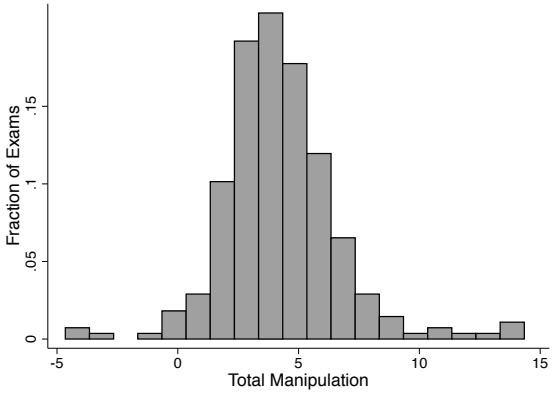
(b) Math



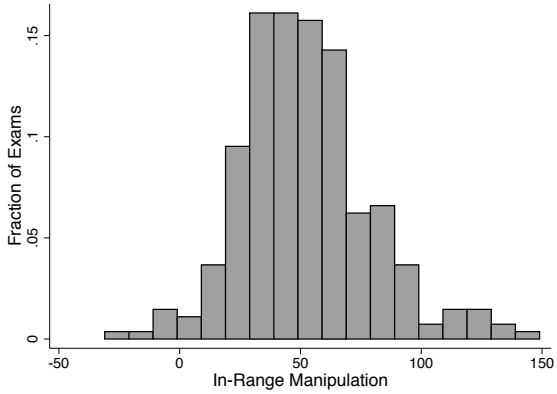
Notes: These figures show the test score distribution around the proficiency score cutoff for New York City grade 3-8 test takers between 2004-2010. Each point shows the fraction of test takers in a score bin. See the data appendix for additional details on the variable definitions.

Appendix Figure 4
Distribution of School Manipulation Estimates, 2004-2010

(a) Total Manipulation



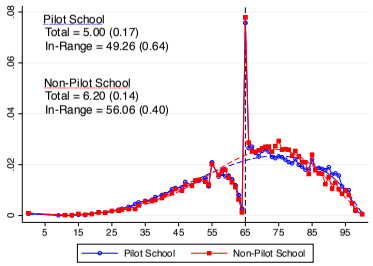
(b) In-Range Manipulation



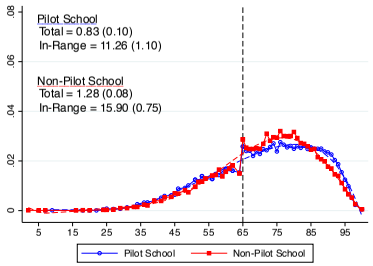
Notes: These figures show the distribution of school x exam x cutoff manipulation estimates for core Regents exams around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Panel (a) is total manipulation estimates. Panel (b) is in-range manipulation estimates. See the text for additional details on the sample and empirical specification.

Appendix Figure 5
Test Score Distributions Before and After Grading Reforms by Subject, 2010-2013

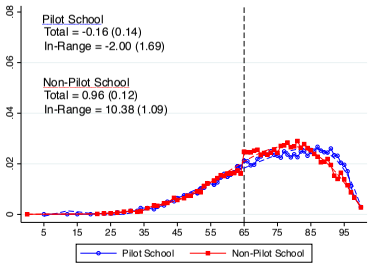
(a) 2010 Living Env.



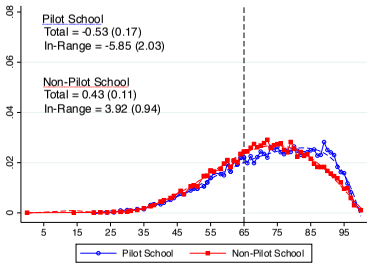
(b) 2011 Living Env.



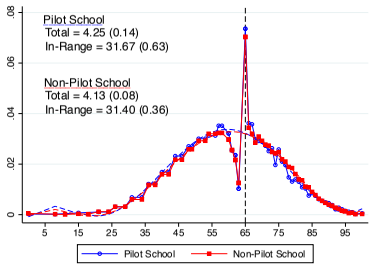
(c) 2012 Living Env.



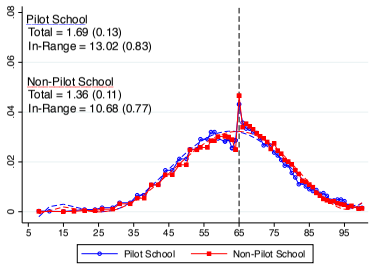
(d) 2013 Living Env.



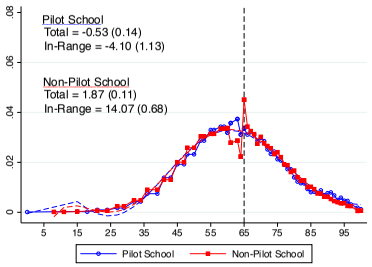
(a) 2010 Int. Algebra



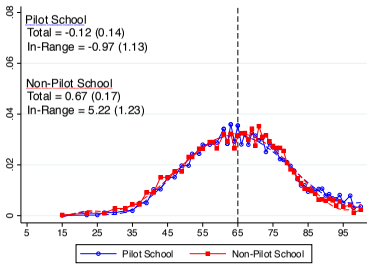
(b) 2011 Int. Algebra



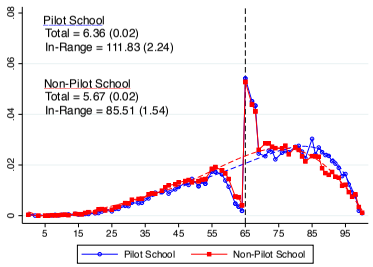
(c) 2012 Int. Algebra



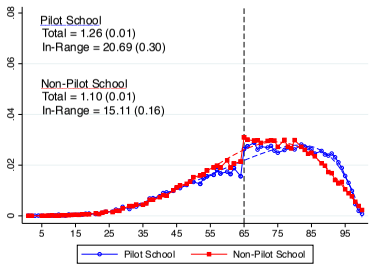
(d) 2013 Int. Algebra



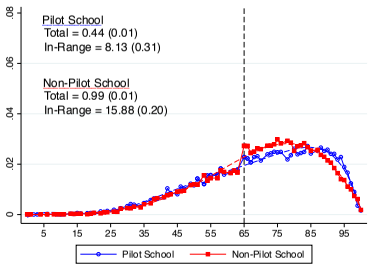
(a) 2010 Global History



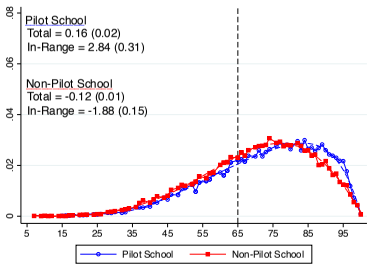
(b) 2011 Global History



(c) 2012 Global History

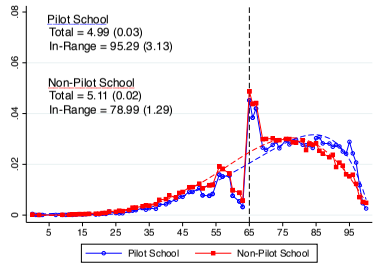


(d) 2013 Global History

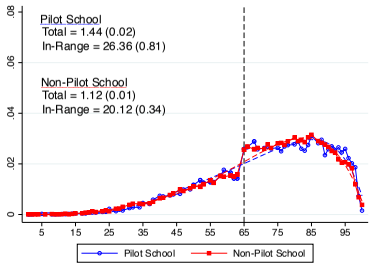


Appendix Figure 5 Continued
Test Score Distributions Before and After Grading Reforms by Subject, 2010-2013

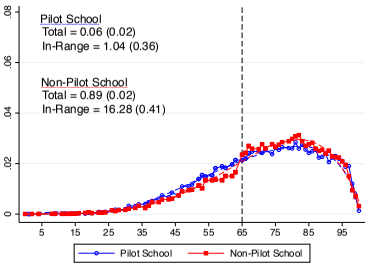
(a) 2010 U.S. History



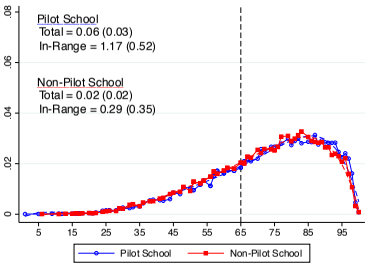
(b) 2011 U.S. History



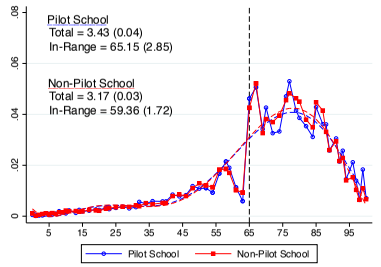
(c) 2012 U.S. History



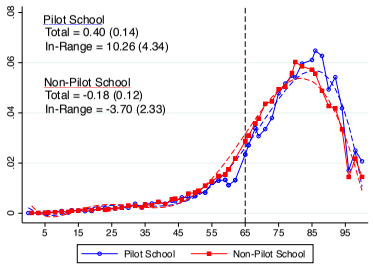
(d) 2013 U.S. History



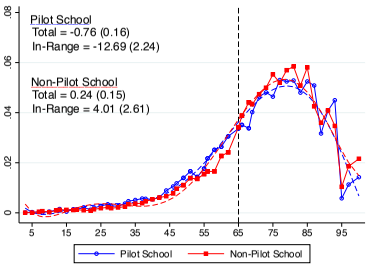
(a) 2010 Comp. English



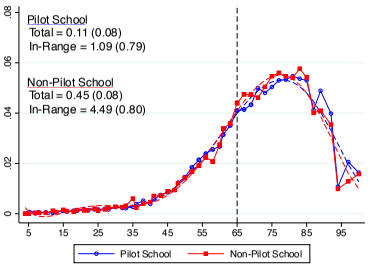
(b) 2011 Comp. English



(c) 2012 Comp. English



(d) 2013 Comp. English

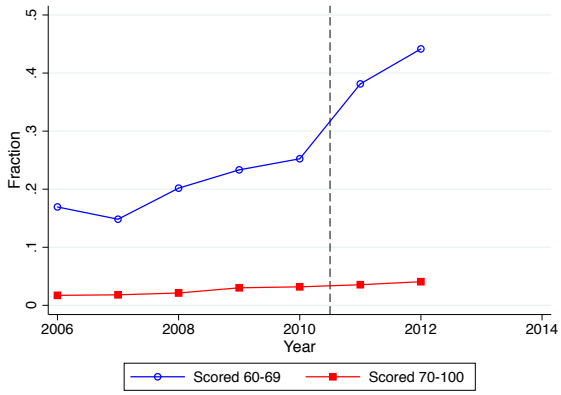


Notes: These figures show the test score distribution around the 65 score cutoff for New York City high school test takers between 2010-2013 in June. Included core exams include English Language Arts, Global History, U.S. History, Integrated Algebra, and Living Environment. Panel (a) considers exams taken in 2010 when re-scoring was allowed and grading was decentralized in both pilot and non-pilot schools. Panel (b) considers exams taken in 2011 when re-scoring was not allowed and grading was decentralized in both pilot and non-pilot schools. Panel (c) considers exams taken in 2012 when re-scoring was not allowed and grading was centralized in pilot schools but decentralized in the non-pilot schools. Panel (d) considers exams taken in 2013 when re-scoring was not allowed and grading was centralized in both pilot and non-pilot schools. See the Figure 5 notes for additional details.

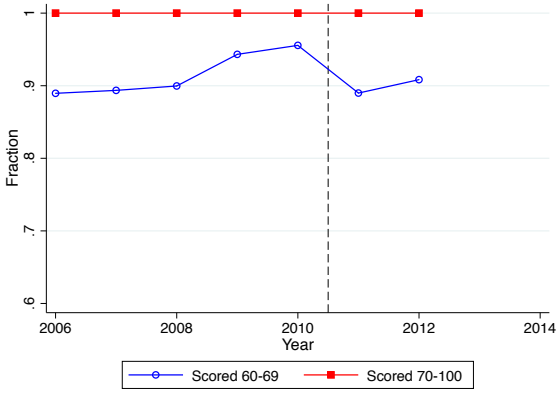
Appendix Figure 6
Regents Grading Reforms and Student Outcomes in Raw Data

(a) Score 65+ on First Administration

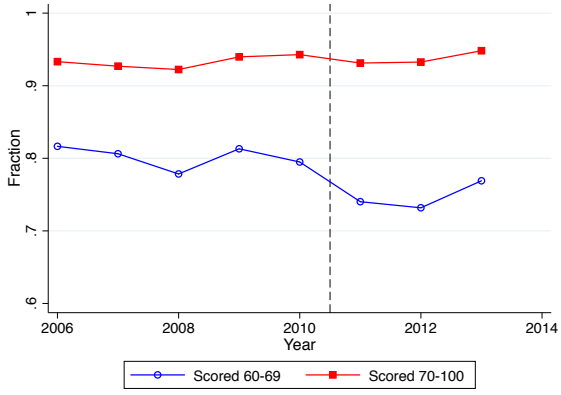
(b) Retake Exam in First Calendar Year



(c) Score 65+ in First Calendar Year

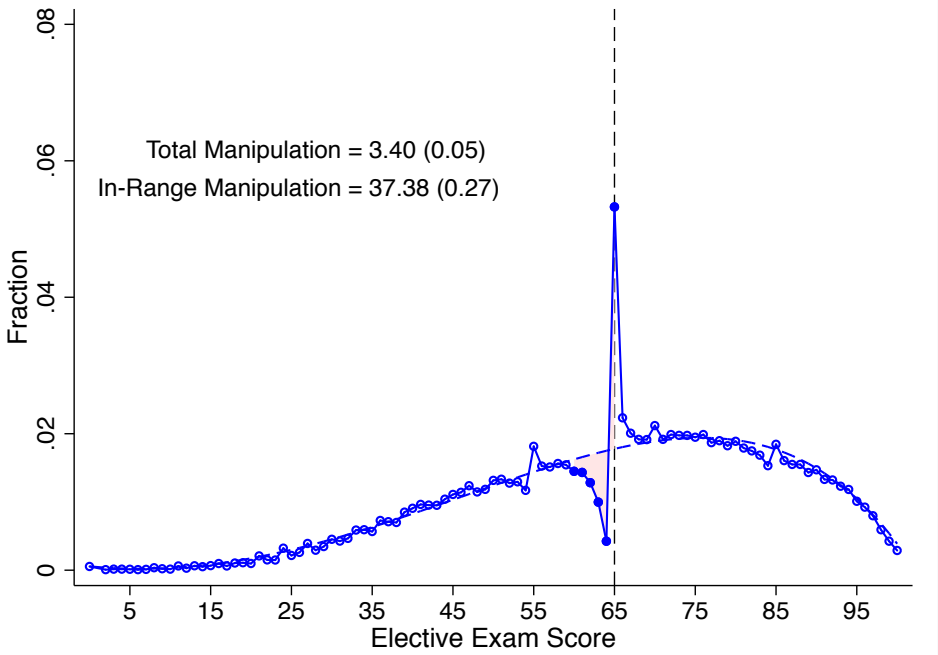


(d) Graduate High School



Notes: These figures plot student outcomes before and after the elimination of Regents re-scoring in 2011 and the de-centralization of Regents scoring in 2012 and 2013. The sample includes students taking Comprehensive English or U.S. History in 11th grade between 2004-2013. We stack outcomes for students taking both exams.

Appendix Figure 7
Results for Elective Regents Exams, 2004-2010



Notes: This figure shows the test score distribution around the 65 score cutoff for New York City high school test takers between 2004-2010. Included elective exams include Chemistry, Math B, and Physics. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near the cutoff. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of the cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

Appendix Table 1
Regents Exam Requirements by Diploma Type and Cohort

Year of 9th Grade Entry	Local Diploma	Regents Diploma	Advanced Regents Diploma
Fall 2001-2004	55+ in 5 core subjects	65+ in 5 core subjects	65+ in 5 core subjects; 65+ in 1 Adv. Math, 1 Physical Science, 1 Language
Fall 2005	65+ in 2 core subjects, 55+ in 3 core subjects		
Fall 2006	65+ in 3 core subjects, 55+ in 2 core subjects		
Fall 2007	65+ in 4 core subjects, 55+ in 1 core subjects		65+ in 5 core subjects; 65+ in Adv. Math Sequence, 1 Physical Science, 1 Language
Fall 2008-present	Available only to Disabled Students		

Notes: The five core Regents-Examination subjects are English, Mathematics, Science, U.S. History and Government, Global History and Geography. Students who have 10 credits of Career and Technical Education (CTE) or Arts classes are exempt from the Language requirement of the Advanced Regents Diploma.

Appendix Table 2
Estimates by Test Subject x Year x Month

	Comp. English	Living Env.	Math A	U.S. History	Global History	Int. Algebra
January 2004:	(1)	(2)	(3)	(4)	(5)	(6)
Total Manipulation	5.23		6.02			
	(0.02)		(0.14)			
In-Range Manipulation	52.77		31.49			
	(0.65)		(0.41)			
	20330		22777			
June 2004:						
Total Manipulation	6.34	7.62	6.01	5.89	6.05	
	(0.02)	(0.20)	(0.14)	(0.01)	(0.01)	
In-Range Manipulation	60.44	35.64	31.51	52.10	52.82	
	(0.71)	(0.53)	(0.42)	(0.72)	(0.49)	
	26686	41878	31168	38107	48935	
January 2005:						
Total Manipulation	5.41		4.73			
	(0.01)		(0.12)			
In-Range Manipulation	51.60		28.57			
	(0.59)		(0.43)			
	23870		24483			
June 2005:						
Total Manipulation	7.58	6.42	4.99	6.73	7.38	
	(0.01)	(0.18)	(0.14)	(0.02)	(0.01)	
In-Range Manipulation	72.36	32.75	26.19	59.17	64.28	
	(0.83)	(0.52)	(0.47)	(0.81)	(0.59)	
	24066	43587	31914	35419	47461	
January 2006:						
Total Manipulation	4.28		3.63			
	(0.02)		(0.14)			
In-Range Manipulation	42.71		19.07			
	(0.52)		(0.53)			
	27822		28211			
June 2006:						
Total Manipulation	5.95	6.91	4.71	6.75	7.76	
	(0.01)	(0.18)	(0.16)	(0.02)	(0.01)	
In-Range Manipulation	58.00	35.39	22.18	60.71	68.05	
	(0.69)	(0.49)	(0.54)	(0.84)	(0.64)	
	24483	41356	28272	36808	47148	
January 2007:						
Total Manipulation	5.76		4.82			
	(0.02)		(0.14)			
In-Range Manipulation	54.91		25.21			
	(0.64)		(0.47)			
	29954		27695			
June 2007:						
Total Manipulation	6.02	6.51	3.40	7.03	7.36	
	(0.02)	(0.18)	(0.16)	(0.02)	(0.01)	
In-Range Manipulation	57.45	33.21	16.00	63.10	65.00	
	(0.67)	(0.52)	(0.60)	(0.89)	(0.61)	
	22404	40941	27254	37699	44552	
January 2008:						

Total Manipulation	3.24		3.60			
	(0.02)		(0.12)			
In-Range Manipulation	32.32		21.91			
	(0.39)		(0.49)			
	27930		26362			
June 2008:						
Total Manipulation	4.02	5.55	3.17	5.65	6.33	2.54
	(0.02)	(0.12)	(0.16)	(0.02)	(0.01)	(0.19)
In-Range Manipulation	40.10	36.27	14.99	50.49	56.33	19.59
	(0.49)	(0.40)	(0.61)	(0.70)	(0.54)	(1.11)
	23617	42081	18048	38293	44954	34186
January 2009:						
Total Manipulation	3.91					5.15
	(0.01)					(0.19)
In-Range Manipulation	38.14					39.83
	(0.45)					(0.71)
	27548					10491
June 2009:						
Total Manipulation	4.15	7.83		7.51	6.52	3.97
	(0.01)	(0.18)		(0.01)	(0.01)	(0.20)
In-Range Manipulation	40.45	39.96		65.42	57.09	30.46
	(0.48)	(0.44)		(0.89)	(0.54)	(0.89)
	23695	41259		39469	43283	39511
January 2010:						
Total Manipulation	3.70					3.80
	(0.02)					(0.13)
In-Range Manipulation	35.32					38.92
	(0.41)					(0.56)
	27098					13955
June 2010:						
Total Manipulation	3.72	8.50		5.84	6.46	4.00
	(0.02)	(0.18)		(0.02)	(0.01)	(0.20)
In-Range Manipulation	35.51	43.53		51.90	56.71	30.51
	(0.41)	(0.40)		(0.73)	(0.53)	(0.89)
	22772	41473		37435	42705	34131

Notes: This table reports manipulation around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the Figure 1 notes for additional details on the empirical specification and the data appendix for additional details on the sample and variable definitions.

Appendix Table 3
Comparison of Pilot and Non-Pilot High Schools

	Pilot Schools	Non-Pilot Schools	Difference
Characteristics:	(1)	(2)	(3)
Male	0.484	0.466	0.018
White	0.197	0.107	0.090**
Asian	0.206	0.204	0.003
Black	0.276	0.302	−0.026
Hispanic	0.315	0.383	−0.068*
Free Lunch	0.651	0.699	−0.048
8th Grade Test Scores	0.326	0.291	0.036
Core Regents Performance:			
Comprehensive English	76.890	75.215	1.675
Living Environment	74.932	74.567	0.365
Int. Algebra	68.795	69.484	−0.689
U.S. History	77.513	76.542	0.971
Global History	72.184	70.781	1.403
Students	54,852	73,416	

Notes: This table reports summary statistics for students in New York City taking a core Regents exam in 2010-2011. Column 1 reports mean values for students enrolled in a school that is in the distributed scoring pilot program. Column 2 reports mean values for students not enrolled in a school that is in the distributed scoring pilot program. Column 3 reports the difference in means with standard errors clustered at the school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table 4
Difference-in-Differences Results by Subject

	Living Env.	Math A/ Algebra	Global History	English	U.S. History
	(1)	(2)	(3)	(4)	(5)
Graduate High School	0.394***	0.197**	0.195***	0.252***	0.263***
	(0.077)	(0.086)	(0.028)	(0.072)	(0.046)
Observations	308,100	373,432	338,926	379,097	297,318
Dep. Variable Mean	0.814	0.746	0.794	0.804	0.842
Student Controls	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes

Notes: This table reports two-stage least squares estimates of the effect of test score manipulation by subject. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. We use the interaction of taking the test between 2011-2013 and scoring between 60-69 as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects, cohort effects, year-of-test effects, and school effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table 5
Robustness of Difference-in-Differences Results

	2SLS	
	(1)	(2)
Graduate High School	0.191***	0.178***
	(0.026)	(0.025)
Observations	1,696,873	1,696,873
Dep. Variable Mean	0.798	0.798
Year-Specific Interaction	Yes	Yes
Pilot School Interaction	No	Yes

Notes: This table reports two-stage least squares estimates of the effect of test score manipulation using different instrumental variables for scoring 65+ on the first administration. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. Column 1 uses the interactions of scoring between 60-69 and year-specific indicators for taking the test between 2011-2013 as instruments. Column 2 uses the interactions of scoring between 60-69 and year-specific indicators for taking the test between 2011-2013 and an indicator for attending a school in the distributed grading pilot program as instruments. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects, cohort effects, year-of-test effects, and school effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table 6
Difference-in-Differences Placebo Estimates

	Sample Mean	Reduced Form	
Panel A: Characteristics	(1)	(2)	(3)
Male	0.471	0.003	0.003
	(0.499)	(0.006)	(0.005)
White	0.145	0.005	0.001
	(0.352)	(0.005)	(0.003)
Asian	0.180	0.002	0.002
	(0.384)	(0.005)	(0.004)
Black	0.316	−0.004	0.003
	(0.465)	(0.008)	(0.004)
Hispanic	0.352	−0.002	−0.003
	(0.478)	(0.007)	(0.004)
Free Lunch	0.602	0.018**	0.015**
	(0.489)	(0.008)	(0.007)
8th Grade Test Scores	0.271	0.058***	0.053***
	(0.819)	(0.008)	(0.007)
Panel B: Predicted Outcomes			
Predicted Graduation	0.798	0.009***	0.008***
	(0.118)	(0.001)	(0.001)
Observations	1,696,873	1,696,873	1,696,873
Student Controls	–	No	No
Year x Score Trends	–	Yes	Yes
School Fixed Effects	–	No	Yes

Notes: This table reports placebo estimates of test score manipulation on student characteristics. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. Columns 2-3 report reduced form results using the interaction of taking the test between 2011-2013 and scoring between 60-69. All specifications include an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects interacted with year-of-test, and exam by year-of-test effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table 7
Difference-in-Differences Results for Additional Outcomes

	Sample Mean	First Stage		Reduced Form		2SLS	
Panel A: Attainment Measures	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Years Enrolled in High School	4.095	−0.158***	−0.159***	−0.090***	−0.085***	0.565***	0.536***
	(0.579)	(0.007)	(0.007)	(0.006)	(0.006)	(0.041)	(0.040)
Highest Enrolled Grade	11.838	−0.158***	−0.159***	−0.064***	−0.065***	0.405***	0.408***
	(0.530)	(0.007)	(0.007)	(0.005)	(0.005)	(0.036)	(0.036)
Panel B: Diploma Requirements							
Regents Requirements Met	0.617	−0.158***	−0.159***	−0.052***	−0.054***	0.327***	0.338***
	(0.486)	(0.007)	(0.007)	(0.010)	(0.009)	(0.058)	(0.053)
Adv. Regents Requirements Met	0.219	−0.158***	−0.159***	0.018**	0.017**	−0.114**	−0.110**
	(0.414)	(0.007)	(0.007)	(0.009)	(0.008)	(0.057)	(0.052)
Panel C: Advanced Science and Math Exams							
Pass Physical Science Exam	0.488	−0.158***	−0.159***	0.012*	0.009*	−0.079*	−0.057*
	(0.500)	(0.007)	(0.007)	(0.007)	(0.006)	(0.044)	(0.031)
Pass Advanced Math Sequence	0.257	−0.158***	−0.159***	0.006	0.004	−0.036	−0.024
	(0.437)	(0.007)	(0.007)	(0.008)	(0.007)	(0.050)	(0.045)
Observations	1,696,873	1,696,873	1,696,873	1,696,873	1,696,873	1,696,873	1,696,873
Student Controls	–	Yes	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	–	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	–	No	Yes	No	Yes	No	Yes

Notes: This table reports estimates of test score manipulation on additional attainment outcomes. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. Columns 2-3 report first stage results from a regression of an indicator for scoring 65+ on the first administration on the interaction of taking the test between 2011-2013 and scoring between 60-69. Columns 4-5 report reduced form results using the interaction of taking the test between 2011-2013 and scoring between 60-69. Columns 6-7 report two-stage least squares results using the interaction of taking the test between 2011-2013 and scoring between 60-69 as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects interacted with year-of-test, and exam by year-of-test effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table 8
Difference-in-Differences Results by Subject for Additional Outcomes

	Living Env.	Math A/ Algebra	Global History	English	U.S. History
Panel A: Attainment Measures	(1)	(2)	(3)	(4)	(5)
Highest Enrolled Grade	0.707***	0.904***	0.336***	0.410***	0.187***
	(0.105)	(0.166)	(0.033)	(0.077)	(0.040)
Years Enrolled in High School	0.969***	1.869***	0.284***	0.248***	−0.016
	(0.117)	(0.247)	(0.036)	(0.089)	(0.056)
Panel B: Diploma Requirements					
Regents Requirements Met	1.337***	0.749***	0.402***	0.304***	0.164**
	(0.186)	(0.216)	(0.043)	(0.100)	(0.069)
Adv. Regents Requirements Met	−0.424***	−0.891***	0.049	−0.125	−0.018
	(0.151)	(0.184)	(0.050)	(0.081)	(0.058)
Panel C: Advanced Science and Math Exams					
Pass Physical Science Exam	−0.133	−0.292**	0.066*	−0.059	−0.051
	(0.093)	(0.128)	(0.038)	(0.067)	(0.044)
Pass Advanced Math Sequence	−0.241**	−0.798***	0.091*	0.009	0.068
	(0.113)	(0.177)	(0.048)	(0.070)	(0.045)
Observations	308,100	373,432	338,926	379,097	297,318
Dep. Variable Mean	0.814	0.746	0.794	0.804	0.842
Student Controls	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes

Notes: This table reports two-stage least squares estimates of the effect of test score manipulation by subject. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. We use the interaction of taking the test between 2011-2013 and scoring between 60-69 as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects, cohort effects, year-of-test effects, and school effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table 9
Placebo Estimates using Across-School Variation

	Sample Mean	Reduced Form	
Panel A: Characteristics	(1)	(2)	(3)
Male	0.467	−0.0601**	−0.0220
	(0.499)	(0.0253)	(0.0176)
White	0.136	0.0559	0.0445***
	(0.343)	(0.0405)	(0.0171)
Asian	0.142	−0.0108	0.0002
	(0.349)	(0.0337)	(0.1149)
Black	0.349	0.0078	−0.0336*
	(0.477)	(0.0303)	(0.0191)
Hispanic	0.364	−0.0509**	−0.0093
	(0.481)	(0.0258)	(0.1579)
Free Lunch	0.529	−0.0303	−0.0097
	(0.499)	(0.0324)	(0.0162)
8th Grade Test Scores	0.156	0.0563	−0.0099
	(0.712)	(0.0567)	(0.0318)
Panel B: Predicted Outcomes			
High School Graduation	0.756	0.0113	−0.0004
	(0.120)	(0.0104)	(0.0056)
Any College	0.505	0.0132	0.0008
	(0.112)	(0.0108)	(0.0052)
Any Two-Year College	0.189	−0.0011	0.0006
	(0.035)	(0.0024)	(0.0014)
Any Four-Year College	0.350	0.0154	0.0004
	(0.143)	(0.0130)	(0.0065)
Observations	587,116	587,116	587,116
Student Controls	–	Yes	Yes
Year x Score Trends	–	Yes	Yes
School Fixed Effects	–	No	Yes

Notes: This table reports placebo estimates that use across-school variation in manipulation. The sample includes students entering high school between 2003-2004 and 2005-2006 and taking core Regents exams between 2004-2010. Columns 2-3 report reduced form results using the interaction of school in-range manipulation and scoring between 60-69. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects interacted with year-of-test, and exam by year-of-test effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table 10
Across-School Results for Additional Outcomes

	Sample Mean	First Stage		Reduced Form		2SLS	
Panel A: Attainment Measures	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Years Enrolled in High School	4.045	0.399***	0.400***	0.013	0.011	0.033	0.026
	(0.691)	(0.023)	(0.023)	(0.018)	(0.019)	(0.045)	(0.047)
Highest Enrolled Grade	11.764	0.399***	0.400***	0.046***	0.036**	0.116***	0.091**
	(0.631)	(0.023)	(0.023)	(0.016)	(0.017)	(0.040)	(0.044)
Panel B: Diploma Type							
Local Diploma	0.265	0.399***	0.400***	−0.165***	−0.149***	−0.414***	−0.374***
	(0.442)	(0.023)	(0.023)	(0.030)	(0.029)	(0.071)	(0.067)
Regents Diploma	0.329	0.399***	0.400***	0.196***	0.201***	0.492***	0.502***
	(0.470)	(0.023)	(0.023)	(0.050)	(0.046)	(0.113)	(0.102)
Advanced Regents Diploma	0.161	0.399***	0.400***	−0.003	−0.033	−0.008	−0.083
	(0.368)	(0.023)	(0.023)	(0.037)	(0.029)	(0.094)	(0.070)
Panel C: Diploma Requirements							
Regents Requirements Met	0.504	0.399***	0.400***	0.189***	0.180***	0.474***	0.451***
	(0.500)	(0.023)	(0.023)	(0.035)	(0.028)	(0.083)	(0.066)
Adv. Regents Requirements Met	0.169	0.399***	0.400***	−0.017	−0.032	−0.043	−0.081
	(0.375)	(0.023)	(0.023)	(0.032)	(0.027)	(0.080)	(0.066)
Panel D: Advanced Science and Math Exams							
Pass Physical Science Exam	0.428	0.399***	0.400***	0.028	−0.020	0.071	−0.050
	(0.495)	(0.023)	(0.023)	(0.025)	(0.025)	(0.063)	(0.063)
Pass Advanced Math Sequence	0.197	0.399***	0.400***	−0.016	−0.028	−0.039	−0.069
	(0.398)	(0.023)	(0.023)	(0.033)	(0.027)	(0.082)	(0.066)
Observations	587,116	587,116	587,116	587,116	587,116	587,116	587,116
Student Controls	–	Yes	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	–	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	–	No	Yes	No	Yes	No	Yes

Notes: This table reports estimates of test score manipulation on student outcomes that use across-school variation in manipulation. The sample includes students entering high school between 2003-2004 and 2005-2006 and taking core Regents exams between 2004-2010. Column 2 reports first stage results from a regression of an indicator for scoring 65+ on the first administration on the interaction of school in-range manipulation and scoring between 60-69. Column 3 reports reduced form results using the interaction of school in-range manipulation and scoring between 60-69. Column 4 reports two-stage least squares results using the interaction of school in-range manipulation and scoring between 60-69 as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects interacted with year-of-test, and exam by year-of-test effects. Standard errors are clustered at both the student and school level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix A: Data Appendix

This appendix contains all of the relevant information on the cleaning and coding of the variables used in our analysis.

A. Data Sources

Regents Scores: The NYCDOE Regents test score data are organized at the student-by-test administration level. Each record includes a unique student identifier, the date of the test, and test outcome. These data are available for all NYC Regents test takers from the 1998-1999 to 2012-2013 school years.

Enrollment Files: The NYCDOE enrollment data are organized at the student-by-year level. Each record includes a unique student identifier and information on student race, gender, free and reduced-price lunch eligibility, school, and grade. These data are available for all NYC K-12 public school students from the 2003-2004 to 2012-2013 school years.

State Test Scores: The NYCDOE state test score data are organized at the student-by-year or student-by-test administration level. The data include scale scores and proficiency scores for all tested students in grades three through eight. When using state test scores as a control, we standardize scores to have a mean of zero and a standard deviation of one in the test-year.

Graduation Files: The NYCDOE graduation files are organized at the student level. For cohorts entering high school between 1998-1999 and 2004-2005, the graduation data include information on the receipt of local, Regents, and advanced Regents diplomas. For cohorts entering high school between 2005-2006 to the present, the data do not include diploma specific information in each year. We therefore measure high school graduation using an indicator for receiving any of the three diploma types offered by New York City in the four years from high school entry. GED diplomas and diplomas awarded after four years are not included in our graduation measure.

National Student Clearinghouse Files: National Student Clearinghouse (NSC) files at the student level are available for cohorts in the graduation files entering high school between 1998-1999 and 2004-2005. The NSC is a non-profit organization that maintains enrollment information for 92 percent of colleges nationwide. The NSC data contain information on enrollment spells for all covered colleges that a student attended, though not grades or course work. The NYCDOE graduation files were matched to the NSC database by NSC employees using each student's full name, date of birth, and high school graduation date. Students who are not matched to the NSC database are assumed to have never attended college, including the approximately four percent of requested records that were blocked by the student or student's school. See Dobbie and Fryer (2013) for additional details.

NCLB Adequate Yearly Progress: Data on Adequate Yearly Progress come from the New York State Education Department's Information and Reporting Services. These data are available from 2004-2011.

NYC School Grades: Data on school grades come from the NYCDOE's School Report Cards. These data are available from 2008-2012.

Regents Raw-to-Scale Score Conversion Charts: Raw-to-scale-score conversion charts for all Regents exams were downloaded from www.jmap.org and www.nysedregents.org. We use the raw-to-scale-score conversion charts to mark impossible scale scores, and to define which scale scores are manipulable. Specifically, we define a score as manipulable if it is within 2 raw points (or 1 essay point) above the proficiency threshold. To the left of each proficiency cutoff, we define a scale score as manipulable if it is between 50-54 or 60-64.

B. Sample Restrictions

We make the following restrictions to the final dataset used to produce our main results documenting manipulation:

1. We only include "core" Regents exams taken after 2003-2004. Exams taken before 2003-2004 cannot be reliably linked to student demographics. The core Regents exams during this time period include: Integrated Algebra (from 2008 onwards), Mathematics A (from 2003-2008), Living Environment, Comprehensive English, US History and Global History. These exams make up approximately 75 percent of all exams taken during our sample period. Occasionally we extend our analysis to include the following "elective" Regents exams: Math B, Chemistry, and Physics. We do not consider foreign language exams due, in part, to the lack of score conversion charts for these years. We also do not consider Sequential Math exams, which are taken before 2003. We also focus on exams taken in the regular test period. This restriction drops all core exams taken in August and the Living Environment, U.S. History, and Global History exams taken in January. We also drop all elective exams taken in January and August. However, the patterns we describe in the paper also appear in the these test administrations. Following this first set of sample restrictions, we have 2,472,197 exams in our primary window of 2003-2004 to 2009-2010.

2. Second, we drop observations with scale scores that are not possible scores for that given exam. This sample restriction leaves us with 2,455,423 remaining exams.

3. Third, we only consider a student's first exam in each subject to avoid any mechanical bunching around the performance thresholds due to re-taking behavior. This sample restriction leaves us with 1,977,915 remaining exams.

4. Fourth, we drop students who who are enrolled in a non-high schools, special education schools, and schools with extremely low enrollments. This sample restriction leaves us with 1,821,458 remaining exams.

5. Fifth, we drop all exams originating from schools where more than five percent of core exam scores contain reporting errors. This is to eliminate schools with systematic mis-grading. This sample restriction leaves us with 1,728,551 remaining exams.

6. Finally, we drop special education students who are held to different accountability standards during our sample period (see Appendix Table 1). This sample restriction leaves us with 1,630,284 remaining exams.

C. Adjustments to Raw Frequency Counts

We create the frequency counts of each exam using the following four step process:

1. First, we collapse the test-year-month-student level data to the test-year-month-scaled score level, gathering how many students in a given test-year-month achieve each scaled score.

2. Second, we divide this frequency of students-per-score by the number of raw scores that map to a given scaled score in order to counter the mechanical overrepresentation of these scaled scores. We make one further adjustment for Integrated Algebra and Math A exams that show regular spikes in the frequency of raw scores between 20-48 due to the way multiple choice items are scored. We adjust for these mechanical spikes in the distribution by taking the average of adjacent even and odd scores between 20-48 for these subjects.

3. Third, we collapse the adjusted test-year-month-scaled score level data to either the test-scaled score or just scaled score level using frequency weights.

4. Finally, we express these adjusted frequency counts as the adjusted fraction of all test takers in the sample to facilitate the interpretation of the estimates.

D. Misc. Data Cleaning

Test Administration Dates: We make two changes to the date of test administration variable. First, we assume that any Math A exams taken in 2009 must have been taken in January even if the data file indicates a June administration, as the Math A exam was last administered in January of 2009. Second, we assume that any test scores reported between January and May could not have been taken in June. We therefore assume a January administration in the same year for these exams. Finally, we drop any exams with corrupted or missing date information that can not be inferred.

Duplicates Scores: A handful of observations indicate two Regents scores for the same student on the same date. For these observations, we use the max score. Results are identical using the min or mean score instead.