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

August 2017

Abstract

We show that the design and decentralized scoring of New York's high school exit exams – the Regents Examinations – led to systematic manipulation of test scores just below important proficiency cutoffs. Exploiting a series of reforms that eliminated score manipulation, we find heterogeneous effects of test score manipulation on longer-run outcomes. While inflating a score increases the probability of a student graduating from high school by about 22 percentage points, the probability of taking advanced coursework declines by roughly 15 percentage points. There is also suggestive evidence that having a score manipulated decreases the probability of enrolling in college. We argue that these results are consistent with test score manipulation helping less advanced students on the margin of dropping out but hurting more advanced students that are not pushed to gain a solid foundation in the introductory material.

*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, 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 United States and across the globe, educational quality is increasingly measured using standardized test scores. These standardized test results can carry extremely high stakes for both students and educators, often influencing 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 among both researchers and policymakers about the fidelity of standardized test results (e.g., National Research Council 2011, Neal 2013). A particular concern is that the consequences associated with these tests can sometimes lead to outright cheating as evidenced by incidents such as the 2009 cheating scandal in Atlanta.[1]

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 or the long-run effect of such manipulation for students. 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 United States, Lavy (2009) finds that a teacher incentive program in Israel increased 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. A related literature finds that student characteristics often influence teacher grading of exams, with girls and students with higher social status often receiving better marks (Lavy 2008, Hinnerich, Hoglin and Johannesson 2011, Hanna and Linden 2012, Burgess and Greaves 2013). Most recently, Lavy and Sand (2015) and Terrier (2016) find that teachers' grading biases can have important impacts on subsequent achievement and enrollment.

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.

We begin our empirical analysis by documenting sharp discontinuities in the distribution of student scores at the proficiency cutoffs, demonstrating that teachers purposefully manipulated Regents scores in order to move marginal students over the performance thresholds. Formal estimates suggest that teachers inflated more than 40 percent of scores that would have been just below the cutoffs on core academic subjects between the years 2004 and 2010, or approximately 6 percent of all tests taken during this time period. However, test score manipulation was reduced by approximately 80 percent in 2011 when the New York State Board of Regents ordered schools

[1] See http://www.nytimes.com/2013/03/30/us/former-school-chief-in-atlanta-indicted-in-cheating-scandal.html. 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 a performance threshold (Neal and Schanzenbach 2010). There is also evidence some schools sought to manipulate the test-taking population advantageously following the introduction of test-based accountability (Figlio and Getzler 2002, Cullen and Reback 2006, Jacob 2005).

to stop 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 around performance cutoffs.

We find that manipulation was present in all New York schools prior to the reforms, but that the extent of manipulation varied considerably across students and schools. We find higher rates of manipulation for Black and Hispanic students, students with lower baseline scores, and students with worse behavioral records. Importantly, however, this is entirely due to the fact that these students are more likely to score close to the proficiency threshold – these gaps largely disappear *conditional* on a student scoring near a proficiency cutoff.

There is also notable across-school variation in rates of manipulation, ranging from 25 percent of "marginal" scores at the 10th percentile school to almost 60 percent of such scores at the 90th percentile school. This across-school variation in test score manipulation is not well explained by school-level demographics or characteristics, and there are several pieces of evidence suggesting that institutional incentives (e.g., school accountability systems, teacher performance pay, and high school graduate rules) cannot explain either the across-school variation in manipulation or the system-wide manipulation. However, we do find evidence that the extent of manipulation within a school depended on the set of teachers within a school grading a particular exam. We argue that, taken together, these results suggest that "altruism" among teachers is an important motivation for teachers' manipulation of test scores (i.e., helping students avoid sanctions involved with failing an exam).

In the second part of the paper, we estimate the impact of test score manipulation on subsequent student outcomes such as high school graduation and advanced course taking. Our empirical strategy exploits the arguably exogenous timing of the decision to prohibit the re-scoring of exams and, then later, to centralize the initial scoring of these exams.[2] 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 27.8 percent increase from the sample mean. The effects on high school graduation are larger for students with higher baseline test scores and white and Asian students, but remain economically and statistically significant for all student subgroups. These results suggest that test score manipulation had important medium- to long-run effects on the graduation outcomes of students in New York City.

While students on the margin of dropping out are "helped" by test score manipulation, we also find evidence that some students are "hurt" by this teacher behavior. Specifically, we find that having an exam score manipulated decreases the probability of taking the requirements for a more

[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.

advanced high school diploma by 16.4 percentage points, a 46.2 percent decrease from the sample mean, with larger effects for students with higher baseline test scores. As discussed in greater detail below, we find evidence suggesting that these negative effects stem from the fact that marginal students who are pushed over the threshold by manipulation do not gain a solid foundation to the introductory material that is required for more advanced coursework. These results are consistent with the idea that test score manipulation has heterogeneous effects on human capital accumulation.

Our paper is closely related to three papers conducted in parallel to our own that examine the long-term consequences of test score manipulation. Two of these papers find results consistent with the positive impact we find that manipulation has on educational attainment. Diamond and Persson (2016) document significant manipulation of test scores around discrete grade cutoffs in Sweden. Using a cross-sectional approach, where students scoring just outside the manipulable range serve as the control group for students inside the manipulable range, 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. Borcan, Lindahl, and Mitrut (2017) find similar results when studying an intervention to reduce test-manipulation in Romania by installing CCTV monitoring of the high-stakes high school exit exam. They find that this centralized oversight significantly reduced fraud but, in turn, led to decreased college access for poor students. A third paper by Apperson, Bueno, and Sass (2016) finds that students who attended middle schools where cheating occurred are more likely to drop out of high school. This result mirrors the negative effects of manipulation on advanced course-taking we find for some students.[3]

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 presents a statistical model to formalize our research questions and motivate the estimating equations for our empirical analysis. Section IV describes our empirical measurement of manipulation, documents the extent of manipulation system-wide, and explores variation in manipulation, and possible drivers of this variation in behavior. Section V presents our difference-in-differences approach and estimates the impact of manipulation on student outcomes. 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. In this section, we describe the features of these exams that are most

[3]In related work on the long-term impacts of high stakes testing, 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, and Dustmann, Puhani, and Schönberg (forthcoming) show that the significant, built-in flexibility of the German tracking system allows for initial tracking mistakes to be corrected over time.

relevant for our study. Additional details can be found in Appendix B.

A. Regents Examinations and High School Graduation

During the period we examine, 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. 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.

An uncommon and important feature of the Regents exams is that they were graded by teachers from students' own schools during most of our sample period. 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 exams are scored on a scale 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 score 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. 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 A1 provides additional details on the degree requirements for each cohort in our sample.[5]

B. The Design and Scoring of Regents Examinations

In addition to multiple choice items, Regents examinations contain open-response or essay questions. For example, the English exam typically asks students to respond to essay prompts after reading passages such as speeches or literary texts. Each of the foreign language exams also contains a

[4]The mathematics portion of the Regents exam has undergone a number of changes during our sample period (2004-2013). However, while there is some variation in how the material was organized, the required exam for graduation essentially always covered introductory algebra as well as a limited number of topics in other fields such as geometry and trigonometry.

[5]In addition to the important proficiency cutoffs at 55 and 65, cutoffs at 75 and 85 scale score points are used by some NY state public colleges as either a prerequisite or qualification for credit towards a degree and by some high schools as a prerequisite for non-Regents courses such as International Baccalaureate. The cutoffs at 75 and 85 are not used to determine eligibility for Advanced Regents coursework. 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 around scores of 75 and 85.

speaking component. Scoring materials provided to schools include the correct answers to multiple-choice questions and detailed instructions for evaluating each open-response and essay question.[6] 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.[7] While scores range from 0 to 100 on all Regents exams, all 101 scale scores are typically not possible on any single exam. Indeed, there are even some exams where it is not possible to score exactly 55 or 65, and, as a result, the minimum passing score is effectively just above those scale scores (e.g., 56 or 66).

During our primary sample period (2003-2004 to 2009-2010), grading guidelines for math and science Regents exams specified that exams with scale scores between 60 and 64 must be scored a second time to ensure the accuracy of the score, but with different teachers rating the open-response questions. 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. 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 and is clearly important for our study. We discuss this in greater depth in Section V, where we examine changes in the Regents re-scoring policies that occurred in 2011.

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 several measures, but all are at least partially based on the Regents Examinations and some are specifically linked to students meeting the 55 and 65 thresholds. Motivated by perceived shortcomings with NCLB, the New York City Department of Education (NYCDOE) implemented its own accountability system starting in 2006-2007. The central component of the NYCDOE accountability system is the school progress reports, which assigns schools a letter grade, ranging from A to F. For high schools, the school grades depend heavily on Regents pass rates, particularly pass rates in the core academic subjects that determine high school graduation. Details on the use of Regents in NCLB and NYCDOE accountability systems are provided in Appendix B. We examine the role of these accountability systems in motivating test score manipulation in Section IV.D.

[6] 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.

[7] Only graders have access to these conversion charts, so students are generally unable to know how their test answers will translate into the final scale score. As a result, it is virtually impossible for a student to target precisely an exact scale score (e.g., 55 or 65).

II. Data

Here we summarize the most relevant information regarding our administrative enrollment and test score data from the New York City Department of Education (NYCDOE). Further details on the cleaning and coding of variables are contained in Appendix C.

The NYCDOE data contain student-level administrative records on approximately 1.1 million students and include information on student race, gender, free and reduced-price lunch eligibility, behavior, attendance, matriculation, state math and English Language Arts test scores (for students in grades three through eight), and Regents test scores. Regents data include exam-level information on the subject, month, and year of the test, the scale score, and a school identifier. Importantly, they do not include raw scores broken out by multiple choice and open-response. We have complete NYCDOE data spanning the school years 2003-2004 to 2012-2013, with Regents test score and basic demographic data available starting in the school year 2000-2001.[8]

We also collected the charts that convert raw scores (i.e., number of multiple choice correct, number of points from essays and open response items), 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 do not correspond to possible scale scores on the indicated exam and must contain an error in either the scale score or test identifier. Second, we use the mapping of raw scores into scale scores for math and science exams to 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 IV.B 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 student enrollment files. We return to tests taken in the school years 2000-2001 and 2001-2002 in Section D 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 A1), although our results are again similar if we include special education

[8] We also have access to student-level National Student Clearinghouse (NSC) data on college enrollment for cohorts in the graduation files entering high school between 2001-2002 and 2004-2005. These data are not used in our main analysis but we present results on college-going using an alternate identification approach discussed in Section V.E.

students. These sample restrictions leave us with 1,629,910 core exams from 514,632 students in our primary window of 2003-2004 to 2009-2010. Table 1 contains summary statistics for the resulting dataset, and Appendix C includes additional information on our sample restrictions and the number of observations dropped by each.

III. Conceptual Framework

In this section, we develop a stylized model of test score manipulation and later educational attainment, abstracting from other inputs, such as teachers or peers, which are typically the focus of education production functions (Todd and Wolpin 2003, Cunha and Heckman 2010, Cunha, Heckman, and Schennach 2010, Chetty, Friedman, and Rockoff 2014). Using this model, we define a measure of test score manipulation that we can estimate using our data. Later, we show how we can estimate the impact of this test score manipulation on later educational attainment using a sharp policy reform.

A. Setup

Our model is characterized by a specification for test scores and a specification for later educational attainment outcomes such as high school graduation. Let s_{ieth} denote student i's observed test score for exam subject e taken at time t and graded by grader h. Let c denote a performance threshold such that a student passes an exam if $s_{ieth} \geq c$.

Test scores are determined by the following function:

$$s_{ieth} = s^*_{iet} + \xi_{ieth} + \phi(i, h, c) \tag{1}$$

Here s^*_{iet} represents the test score that the student would get in expectation on test submissions if reviewed by "unbiased" graders who have no information about the student (e.g., name, demographics, prior achievement) and simply apply the instructions for marking individual test questions to the test submissions. This persistent component of the test score reflects factors such as a student's subject knowledge at time t, the student's test taking ability, and so on. The term ξ_{ieth} represents idiosyncratic factors at the student-exam-time-grader level that affect the perceived quality of any given test submission but are not persistent across test submissions and are equal to zero in expectation. This noise component includes factors such as guessing on multiple choice items, arbitrary alignment of questions with the local curriculum, classical measurement error by graders, and so on. Finally, $\phi(i, h, c)$ represents potential "bias" by exam graders, who may manipulate the final test score s_{ieth} based on additional information they possess about student i, the beliefs and incentives of grader h, and the grader's knowledge of the cutoff c. For example, graders might inflate the exam scores of particularly well-liked students or in order to boost measured performance under a school accountability system.[9]

[9] Unlike Diamond and Persson (2016), we do not explicitly model graders' incentives, but one may have in mind a model where graders benefit from increasing the number of students passing exams but pay a cost for introducing

High school graduation G_i is a binary outcome determined by whether a student passes a required set of E exam subjects (which can be retaken multiple times) as well as performing other required work (e.g., accumulating course credits):

$$G_i = \mathbf{1}[\eta_i > 0] * \prod_{e=1}^{E} \mathbf{1}[\max_t(s_{ieth}) \geq c] \tag{2}$$

where η_i reflects individual heterogeneity in students' abilities to complete non-exam graduation requirements and may be correlated with the bias component $\phi(i, h, c)$. For example, it is possible that exams are graded more leniently for well-behaved students.

Later outcomes in life Y_i such as college enrollment or earnings depend on students' abilities to complete non-exam graduation requirements, students' knowledge and skills across various subject areas, and high school graduation itself:

$$Y_i = f_i(\eta_i, s_i^*, G_i) \tag{3}$$

where s_i^* is the set of student skills and knowledges across all subjects. The influence of these variables on outcomes may be heterogeneous across individuals i. For example, it is possible that the impact of high school graduation G_i will be different for high- and low-ability students.

One limitation of our simple framework is that we do not specify a role for student effort and learning over time in the determination of s_{iet}^*. If students fail an exam and are forced to re-take a course, it is likely that their knowledge s_{iet}^* will increase, resulting in a higher test score s_{ieth}. For this reason, our measures of manipulation are based on students' first test administration. Another limitation of our framework is we do not specify a relationship among test scores in different subjects. For example, if a student acquires higher skills s_{iet}^* in a subject such as Algebra, that student will likely perform better on the exam in Algebra 2/Trigonometry. This issue becomes relevant when we consider the impact of manipulation on students' enrolling in and passing advanced Regents courses. If a student fails a required regents' exam, such as Algebra, and is forced to re-take the course, the students' knowledge may increase, resulting in both higher test scores in Algebra and better preparation for advanced coursework such as Algebra 2/Trigonometry. This highlights a key tension involved in test score manipulation; raising a student's score s_{ieth} may help them graduate from high school but could impede accumulation of skills and knowledge. We return to this issue in Section V.

B. Defining Test Score Manipulation

Our first empirical challenge is to estimate the fraction of exams that are manipulated by grading bias so that they reach or exceed the passing cutoff c instead of falling just below the cutoff. We simplify the analysis by assuming that graders only consider manipulating exam scores that are

test score bias $\phi(i, c, h)$. Student or grader specific variation in the benefits or costs of introducing bias generates variation in test score manipulation across those dimensions.

below the performance threshold and are "close enough" so that a small amount of manipulation would allow the student to meet the high school graduation requirements. In the context of our framework, we impose the following restrictions on the bias term ϕ:

$$\phi(i,h,c)\begin{cases} 0 \text{ if } s^*_{iet} + \xi_{ieth} \geq c \\ 0 \text{ if } s^*_{iet} + \xi_{ieth} < M^-_{cet} \\ \{0, c - s^*_{iet} - \xi_{ieth}\} \text{ if } c > s^*_{iet} + \xi_{ieth} \geq M^-_{cet} \end{cases} \tag{4}$$

Grading bias is equal to zero for exam scores that would have already been at or above the threshold c, as well as for exam scores that would fall strictly below some score M^-_{cet} beneath the threshold c. For the range of potentially manipulable scores from M^-_{cet} to c, bias can be either zero (i.e., no manipulation) or equal to the additional points needed to meet the threshold c. Conditional on an exam score falling in this manipulable range, the grader can consider various student- and school-level factors when deciding whether to inflate a score to the threshold c.[10]

The amount of manipulation at cutoff c, β_{cet}, is defined as the fraction of exams inflated to meet the cutoff c:

$$\beta_{cet} = \frac{\sum_{i=1}^{I_{et}} \mathbf{1}[\phi(i,h,c) = c - s^*_{iet} - \xi_{ieth}]}{I_{et}} \tag{5}$$

where I_{et} is the total number of test takers for exam e at time t.

Let F_{set} denote the fraction of students with the observed test score of s on exam subject e at time t:

$$F_{set} = \frac{\sum_{i=1}^{I_{et}} \mathbf{1}[s_{ieth} = s]}{I_{et}} \tag{6}$$

Similarly, let F^*_{set} denote the expected fraction of students who would have received the test score s on exam e at time year t in absence of any grading bias:

$$F^*_{set} = \frac{\sum_{i=1}^{I_{et}} \mathbf{1}[s^*_{iet} = s]}{I_{et}} \tag{7}$$

It is straightforward to see that:

$$\beta_{cet} = E\left[\sum_{s \in [M^-_{cet}, c)} (F^*_{set} - F_{set})\right] = E\left[F_{cet} - F^*_{cet}\right] \tag{8}$$

In other words, manipulation can be measured using either the number of "missing exams" in the manipulable range from M^-_{cet} to just below the threshold score c, or the number of "extra exams"

[10] The simplification of zero bias outside of the range near the cutoff makes the exposition of the model and empirical strategy more transparent. In practice, however, our empirical measure of manipulation relies on the discontinuity in the distribution of test scores around the cutoff c. It is therefore possible to relax the above assumptions so long as any factors related to grading bias trend smoothly through c. In this scenario, our estimates identify the additional manipulation around c, rather than the total amount of manipulation across all test scores. We are not able to use our empirical strategy to separate any potential continuous sources of bias from any other continuously distributed factor that affects test scores such as student ability or knowledge.

exactly at the threshold score. Estimates of the amount of manipulation β_{cet} therefore require information on both the observed test score distribution F_{set} and the unobserved, counterfactual test score distribution F^*_{set}. In the next section we provide details on our method for estimating F^*_{set} and describe our findings on the magnitude of test score manipulation.

IV. The Manipulation of Regents Exam Scores

A. Estimating Test Score Manipulation

As noted above, the actual test score distribution F_{set} is observed, but the counterfactual test score distribution F^*_{set} must be estimated. We follow an approach similar to Chetty et al. (2011), who examine manipulation of taxable income at certain thresholds where marginal tax rates change discontinuously. Specifically, we calculate the counterfactual distribution of scores by fitting a polynomial to the frequency count of exams by test score s, excluding data near the proficiency cutoffs with a set of indicator variables, using the following regression specification (dropping exam e and time t subscripts for simplicity):

$$F_s = \sum_{q=0}^{Q} \pi_q \cdot s^q \; + \sum_{j \in [M_c^-, c]} \lambda_j \cdot \mathbf{1}[s = j] \; + \; \varepsilon_s \tag{9}$$

where q is the order of the polynomial and ε_s captures sampling error. We define an estimate of the counterfactual distribution $\{\widehat{F}_s\}$ as the predicted values from Equation (9) omitting the contribution of the indicator variables around the cutoffs: $\widehat{F}_s = \sum_{q=0}^{Q} \widehat{\pi}_q \cdot s^q$. In practice, we estimate $\{\widehat{F}_s\}$ using a sixth-degree polynomial ($Q = 6$) interacted with the exam subject e. Our results are not sensitive to changes in either the polynomial order or whether we interact the polynomial with both exam subject and exam year.

A key step in estimating Equation (9) is identifying the potentially manipulable test scores 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 are able to identify potentially manipulable or manipulated test scores ex-ante based on knowledge of the Regents grading rules. 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. On the math and science exams, we therefore define a score as manipulable to the left of each proficiency cutoff if it is between 50-54 or 60-64. This range is also highly consistent with the patterns we observe in the data. We set the upper bound of the manipulable range at exactly the cutoff c, since it is generally 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.[11]

[11] Note that there are rare cases in which the exact cutoff of 55 or 65 is not a possible scale score on a math or

Manipulating a score to be exactly 55 or 65 can be challenging (if not impossible) for the exams in English and social studies. This is because changes in essay ratings of just one point typically change the scale score by four points. For example, a student that initially scores a 63 might be moved to a 67 if a grader awards an additional point on one of the essay prompts (see, for example, Appendix Figure A1).[12] We therefore define a score as within the manipulable range $[M_c^-, c]$ for the English and social science exams if it is within 1 essay point of the proficiency threshold. Defining the manipulable range in this manner is highly consistent with the patterns we observe in the data (see, for example, Appendix Figure A2). Our estimates are also not sensitive to small changes in the manipulable score region to either the left or right side of the proficiency cutoffs.

If our ex-ante demarcation of the manipulable range is accurate, then the unadjusted counterfactual distribution from Equation (9) 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 (9) to the right of the tax rate kink to satisfy the integration constraint. Given that the integration constraint is satisfied in our context, we estimate manipulation using an average of the missing mass just to the left of the cutoff and excess mass at each cutoff:

$$\widehat{\beta}_c = \frac{1}{2}\left[\left(\sum_{s\in[M_c^-,c)} \widehat{F}_s - F_s\right) + \left(\sum_{s\in c} F_s - \widehat{F}_s\right)\right] = \frac{1}{2}\left[\left(\sum_{s\in[M_c^-,c)} -\widehat{\lambda}_s\right) + \left(\sum_{s\in c} \widehat{\lambda}_s\right)\right] \tag{10}$$

As seen in Equation (8), we could use either the "missing mass" or the "excess mass" to characterize the extent of manipulation. Since both of these measures will contain sampling error, we combine the two in order to increase the precision of our estimates, but our main results are nearly identical if we only use information from one side of the cutoff.

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 range: $\widehat{\beta}_c / \sum_{s\in[M_c^-,c]} \widehat{F}_s$. 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 specifications that pool multiple exams, 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 specification changes such as the polynomial order, the manipulable score region, or the weighting across exams.

We calculate standard errors for test score manipulation $\widehat{\beta}_c$ using a version of the parametric

science exam, and 56 or 66 is used instead as the upper bound for the manipulable range.

[12] To be more specific, the excess mass for the math and science exams is clearly concentrated on 55 and 65 (and 56 for the Math A and Living Environment exams where it is not possible to score exactly 55). In contrast, the excess mass for the English and Social Studies exams is spread more evenly among the small number of scores that are within 1 essay point of each cutoff.

bootstrap procedure developed in Chetty et al. (2011). Specifically, we draw with replacement from the distribution of estimated vector of errors $\widehat{\varepsilon}_s$ in Equation (9) at the score-exam-test administration level to generate a new set of scale score counts from which we can generate bootstrapped estimates of $\widehat{\beta}_c$. We define the standard error as the standard deviation of 200 of these bootstrapped estimates.

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

We begin by examining the distribution of core Regents exam scores near the proficiency thresholds at 55 and 65 points in Figure 1. We first focus on all core Regents exams taken between 2003-2004 and 2009-2010, as exams taken after 2009-2010 are subject to a different set of grading policies that we discuss in Section V.A.

To construct figures of test score distributions, we first collapse the data to the subject-year-month-score level (e.g., Living Environment, June 2004, 77 points). We then make two minor adjustments to account for two mechanical issues that affect the smoothness of the distribution of scale scores.[13] The 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 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. Scores immediately below the cutoffs appear less frequently than one would expect from a well-behaved empirical distribution, and the scores at or just above the cutoffs appear more frequently than one would expect. In Appendix Figures A2 and A3, we show that this pattern is still apparent if we examine test scores separately by subject or by year.[14]

Figure 1 includes the counterfactual density $\{\widehat{F}_s\}$ predicted using Equation (9), shown by the dotted line, as well as our point estimates for manipulation and standard errors. We estimate the average amount of manipulation on the Regents core exams to be 5.7 (se=0.02). That is, approximately 6 percent of all Regents core exams between 2004 and 2010 were manipulated to meet the proficiency cutoff. Within the range of potentially manipulable scores, we estimate that an average of 44.1 (se=0.20) percent of Regents core exams were manipulated. We also look

[13] 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, by dividing the scale score frequency by the number of raw scores that map into it. 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. We also adjust Integrated Algebra and Math A exams for an alternating frequency pattern at very low, even 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.

[14] Appendix Figure A3 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 A1). We therefore focus on the 65 cutoff when examining manipulation after 2010.

separately at all subjects and test administrations and find economically and statistically significant manipulation of all Regents core exams in our sample (see Appendix Table A2). Math and science exams tend to have somewhat lower levels of manipulation than English and social science exams. This is consistent with the notion that teachers view manipulation on multiple choice items – which have relatively high weight in the math and science exams – as more costly than on open-response items, but we lack sufficient variation for a formal test of this idea.[15]

To provide further evidence that Regents scores near cutoffs were being manipulated, Appendix Figure A5 shows the score distributions for math and English exams taken by New York City students in the third through eighth grades, which also involve high stakes around specific cutoffs but are graded centrally by the state. These distributions are smooth through the proficiency cutoff, 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.

A related concern is that the patterns we see in Figure 1 are the result of classical measurement error combined with a policy of re-grading exams with scores between 50-54 and 60-64. Several pieces of evidence suggest that this kind of mechanical relationship is not driving our results. First, such a practice would lead to a hollowing out within the marginal range and excess mass both above and below the re-grading thresholds, yet the test score distribution is clearly smooth just below 50 points (see Figure 1). Second, on the math and science exams, where it is generally possible to add points to open-ended questions in order to meet the 55 or 65 cutoffs exactly, we can easily see that almost all of the excess mass occurs exactly at 55 and 65, while the missing mass is spread smoothly across the 50-54 and 60-64 ranges (see Appendix Figure A2). This strongly supports the notion that manipulation is designed with the cutoffs in mind.[16] A third piece of evidence comes from the English exams, where the only way to increase a student's score (other than changing a multiple choice answer) is to add a raw point on an essay question. Each raw essay point is typically worth four scale points, so (focusing on the 65 cutoff for simplicity) any initial score from 61 to 64 requires just one essay point to cross the cutoff and land in the range from 65 to 68, while adding an essay point to an initial score of 60 brings the student to 64. Correction of measurement error in the 60-64 range would imply a smaller amount of missing mass at 64 than in the range 61-63, since exams moved from 60 to 64 will fill in for exams moved up from 64 to 68. However,

[15]The weight on multiple choice items varies almost exclusively across subjects, rather than over time within subjects, leaving little room to separate differences in weighting of multiple choice from other differences across subjects. One interesting and informative observation comes from the June 2001 Chemistry exam, which is the only test in our data that consists solely of multiple-choice questions. In Appendix Figure A4, one can see clear discontinuities in the distribution of scores at the 55 and 65 cutoffs despite the lack of open-response questions. However, the amount of manipulation is significantly less than similar elective exams from that time period, suggesting that the cost of manipulation of multiple choice items is higher than manipulation of open-response, but not so high as to eliminate manipulation entirely.

[16]One could say that teachers are "correcting measurement errors" in the range below the cutoffs but (a) only correcting negative errors while ignoring positive ones and (b) applying corrections just up to the point that students meet the cutoff. This is, in our view, just a different characterization of the "manipulation" we describe.

this pattern of results is not what we observe in Appendix Figure A2. If anything, there appears to be somewhat greater missing mass at 64 and, likewise 54. The data are far more consistent with teachers viewing initial English scores of 50 and 60 as much more costly to manipulate, as they require two separate changes to essay scores in order to meet the cutoff.[17]

Finally, it is important to note that the practice of manipulation on Regents exams was not unique to New York City. In an early version of this research (see Dee et al. 2011), we present evidence that similar manipulation occurred across the state of New York. Unfortunately, these state-wide data are only available for the June 2009 administration of the Regents exams and lack information on student characteristics, rendering them of little use for answering many important research questions.

C. Heterogeneity in Manipulation Across Schools and Students

Not all students with scores just below the cutoffs have their scores manipulated, raising the question of whether test score manipulation varies systematically across students and schools. We examine this issue in a number of ways.

First, we estimate manipulation for each high school in our sample and plot these distributions in Figure 2.[18] Notably, the practice of manipulation appears to have been quite widespread, as we see no significant subset of schools with estimated manipulation near zero. At the same time, the intensity with which manipulation was practiced varied widely across schools; the ranges from the 10th to 90th percentiles are 3.7 to 9.6 percent for total manipulation and 24.5 to 56.6 percent for in-range manipulation.[19] Thus, the probability of a marginally failing exam being manipulated

[17] It is worth noting that evidence from the U.S. History and Global History exams also supports our argument that mechanical correction of measurement error is not consistent with the data. The scoring of these exams bears similarities in scoring to both math/science – i.e., changing open answer ratings can raise a student's score by exactly one scale score point – and English – i.e., there are also essays where one raw point translates to four scale score points. Thus, in line with our explanations above, the scoring distributions for the social science tests look like hybrids of the other distributions, with noticeable peaks exactly at 55 and 65, extended but smaller ranges of excess mass through 58 and 68, and missing mass at 54 and 64 that is slightly larger than at lower in-range scores.

[18] Because some high schools are small, we estimate the counterfactual distribution for each test subject by splitting all high schools into five quintiles based on average Regents scores and generating a counterfactual for all exams in the quintile using Equation (9). We then calculate manipulation at the school-exam level and aggregate these to estimate manipulation across all exam administrations at the school. We also 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 279 schools from 2004 to 2010. Consistent with our results from Figure 1, total manipulation estimates are centered around 6 percent while in-range manipulation estimates are centered at around 40 percent. Results are qualitatively similar if we generate counterfactuals using either fewer or more quantiles, or if we restrict our sample to the subset of large high schools where we can estimate school x subject-specific counterfactual distributions.

[19] Of course, because each of these individual school estimates is measured with error, the distribution shown in Figure 2 could overstate the true variance in the population (Jacob and Rothstein 2016). To show that sampling error is not a major factor, Figure 2 also plots how the number of exams, both total and only in-range, varies with manipulation. While schools at the extreme tails of the distributions have lower sample sizes, consistent with larger measurement errors, schools near the 10th and 90th percentiles have at least 4,000 exams, around 1,000 of which are in the manipulable range. Additionally, we calculated manipulation at the school x subject level rather than the school level and tested for the significance of school effects in a random effects regression that controlled only for exam subject. School effects were highly significant, with a standard deviation of 2.1 percentage points for total manipulation and 11.3 percentage points for in-range manipulation, very much in line with 90-10 ranges mentioned

clearly depended on which school the student attended.

Regressions of school-level manipulation on school characteristics (Table 2) show that total manipulation is positively associated with Black and Hispanic enrollment, enrollment of students eligible for free or reduced-price lunch, and enrollment of students with lower baseline test scores.[20] This is not surprising given that these schools are likely to have higher proportions of exams with scores near the cutoffs. Indeed, when we examine in-range manipulation, schools whose students have higher 8th grade test scores exhibit (slightly) less in-range manipulation, while the estimated relationships between in-range manipulation and schools' fractions of racial minorities or students from poor households are small not statistically different from zero. Total manipulation is also (slightly) negatively associated with school size, but smaller high schools exhibit (slightly) less in-range manipulation. Thus, school level manipulation varied widely, but observables predict only a small amount of variation in total manipulation and little or no variation in in-range manipulation.[21]

We also estimate manipulation splitting the sample by student subgroup, regardless of the school they attended (Appendix Figure A7). Differences in total manipulation across student subgroups are as expected, with larger percentages manipulated for lower scoring groups. For in-range manipulation, we find fairly small differences when comparing students by gender or eligibility for free and reduced-price lunch. However, we estimate that lower percentages of in-range exams were manipulated for Black and Hispanic students, students with poor behavior (defined as having a behavioral violation or more than 20 absences), and, to a lesser extent, students with higher 8th grade test scores.

These gaps reflect both within- and across-school variation in manipulation so, we gauge the magnitude of the within-school component using a simple but intuitive Monte Carlo technique where we reassign characteristics randomly among students taking the same exam within each school.[22] Gaps by synthetic subgroup only reflect across-school differences in manipulation. Thus, if gaps disappear in the synthetic results then we have evidence of within-school differences in manipulation across students. This is precisely what happens when we assign high baseline test scores or

---

above.

[20] Regressions are weighted by the number of in-range exams, but weighting by total exams provides quite similar results.

[21] We find similar results if we simply split the sample by various school characteristics and re-estimate manipulation using all core exams (see Appendix Figure A6). Schools whose populations tend to have lower average achievement (i.e., Black/Hispanic, free lunch, low 8th grade test scores) are estimated to have manipulated higher fractions of exams overall. For example, total manipulation is twice as large for high schools with low 8th grade test scores (6.9 percent) than schools with high 8th grade scores (3.4 percent). When we compare in-range manipulation, there is less evidence of major systematic difference; rates are fairly similar across school groups and some gaps reverse sign. For example, schools with high enrollment of Black/Hispanic students show estimated in-range manipulation of 43.4 percent, while those with low Black/Hispanic enrollment have in-range manipulation of 45.2 percent. Additionally, while smaller schools total manipulation is slightly higher than large schools, rates of in-range manipulation are 5.2 percentage points lower. We split schools using the exam-weighted median for each characteristic, although results are qualitatively similar if we split using student- or school-level medians.

[22] We reassign characteristics keeping the fraction of students with each subgroup designation constant both within schools and across all schools. We then re-estimate manipulation for the randomly assigned subgroups, repeating this process 100 times. Note that one limitation of this approach is that reassignment of student characteristics will lead to differences among students both within and outside the manipulable range, thus altering our estimated counterfactual distributions.

good behavior randomly within schools (Appendix Table A4), suggesting significant within-school differences in how they are treated and supporting the idea that teachers use some soft information about students when deciding to manipulate a score near the cutoff.[23]

Finally, we examine variation in manipulation across subjects and across time within a school, and whether this variation can be linked to specific groups of teachers. Although Figure 2 shows substantial heterogeneity at the school level in the extent of manipulation, the reality is that only a subset of teachers in specific subject areas are responsible for scoring each Regents exam. Thus, manipulation may be driven to some degree by the particular groups of individual teachers doing the grading, rather than a general school-wide culture or administrative policy. Here we present some evidence in favor of this idea, using estimates of in-range manipulation at the school x subject (rather than school) level, calculated using the same methodology used to create Figure 2. Appendix Table A3 presents the mean and standard deviation of these estimates by subject, as well as the within-school correlations across each subject-pair. Average in-range manipulation is higher and more varied in English and social studies exams, but both the level of manipulation and variation across schools is still considerable in math and science. All of the within-school correlations are positive, indicating some consistency in the practice across groups of teachers within the school. However, all of the correlations except one are fairly low, with a range extending from below 0.1 to just under 0.3, suggesting that particular groups of teachers within a school may be more or less inclined to manipulate. Further support for this idea comes from the very high correlation (0.77) in manipulation estimates between the two history exams, which are likely to be graded by members of the same group of (social studies) teachers.[24] Thus, the culture of manipulation can vary within the school, and may be closely tied to the particular set of teachers performing grading duties.

In order to investigate further the importance of teachers driving manipulation, we examine the extent to which persistence over time in manipulation within a subject area and school is mediated by teacher turnover. We therefore estimate (1) manipulation at the school x subject level in two separate periods, 2004-2006 and 2007-2009, and (2) the fraction of teachers with the relevant license area for each exam (e.g., English license for the English exam, Mathematics license for the Math A and Algebra exams, etc.) who were employed at the school in both periods.[25] We begin by regressing manipulation in 2007-2009 on its "lagged" value from 2004-2006, as well as indicators

[23]The "synthetic gap" is still present (though about half as large) when ethnicity is assigned randomly within schools, suggesting that it is partially due to differences across the schools these students attend and partially due to within-school differences in how students are treated, conditional on having a score close to the cutoff. Of course, any within-school difference in manipulation by racial/ethnic groups may be driven by other characteristics correlated with race and ethnicity.

[24]The high correlation between the two history exams may also alleviate the concern that the lower correlations for other pairs are simply due to a large degree of measurement error in school x subject estimates.

[25]We assign teachers to a subject area based on license: English licenses for the English exam, Mathematics for the Math A and Integrated Algebra exams, Social Studies for the Global and U.S. History Exams, and Biology, Chemistry, Earth Science, Physics, and General Science for the Living Environment Exam. We calculate the fraction in both periods as the number within each school x subject employed in both three-year periods divided by the total number of teachers within each school x subject employed at any time during these six years. Note that while we could perform this analysis at the school x subject-year level, pooling across several years and, thus, exam administrations, greatly reduces noise in the manipulation estimates.

for subject area, and find a coefficient on lagged manipulation of 0.49 (se=0.08) (Table 3, Column 1). Adding school fixed effects (Column 2) decreases this measure of persistence slightly, to 0.42 (se=0.08), but clearly shows that variation in manipulation across subjects within the same school reflects real differences in culture that persist over time.

We then add controls for the fraction of teachers employed in both periods and its interaction with lagged manipulation (Column 3). If teachers are an important driver of manipulation practices, we would expect this interaction to be positive, i.e., greater persistence when the set of teachers remains the same. Consistent with this hypothesis, the interaction term is 0.77 (se=0.20) and highly significant, while the coefficient estimate for the main effect of manipulation (i.e., for a school with complete teacher turnover between the two periods) is just 0.09 (se=0.07) and not statistically different from zero. Of course, schools with greater teacher turnover may be changing culturally for other reasons (e.g., changes in school principal), but we find this result is robust to the addition of school fixed effects (Column 4), where the identifying variation is based on variation in the rates of turnover across subjects within the same school. Thus, while school-wide culture is a likely factor, both the correlations across subject areas and the influence of teacher turnover at the school x subject level support the notion that the extent of manipulation also depended greatly on the set of teachers within a school grading a particular exam.

D. Exploring Institutional Explanations for Manipulation

We have shown that test score manipulation was widespread among schools in New York, although clearly there was variation due to particular "cultures" which existed at the school or among groups of teachers. Here we briefly explore three additional potential drivers of the system-wide manipulation of Regents exams that relate to potentially important institutional incentives.

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 2006, 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 3 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 60.2 (se=0.81) percent of in-range exams manipulated before the implementation of these accountability systems, compared to the 43.3 (se=0.27) percent in the years after the implementation of these systems.[26]

[26] 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.

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 3 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 3 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 44.2 (se=0.21) percent of in-range exams, compared to 44.8 (se=0.62) percent for schools meeting AYP. Similarly, schools receiving a D or F from the NYC accountability system manipulate 43.4 (se=0.41) percent of in-range exams, compared to 42.3 (se=0.36) percent for schools receiving an A or B. Thus, we find no evidence that pressure from test-based school accountability systems was a primary driver of the manipulation documented above.

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 3 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).[27] We find that control schools manipulated 44.2 (se=0.33) percent of in-range exams taken during the experiment, which is higher than our estimate of 41.2 (se=0.24) percent 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 and possibly strengthen their college application. Appendix Figure A8 plots frequency distributions for scores 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 (se=0.04) percent of these elective Regents exams were manipulated in total, and that 37.4 (se=0.25) percent were manipulated among those

[27] 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 student attendance, behavior, or achievement. See Fryer (2013) for additional details.

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 institutional incentives created by school accountability systems, formal teacher pay-for-performance 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 barely 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 systematic characteristics of individuals whose behavior drives this practice.

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

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

On February 2, 2011, the Wall Street Journal published an exposé 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. The New York Times published a similar story and the results of its own analysis on February 18th, including a statement by a New York State Education Department official that acknowledged the existence of anomalies in the Regents score distribution had been known for some time. 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 re-score any of the open-ended questions on this exam after each question has been rated once, regardless of the final exam score."[28]

In October 2011, the Board of Regents further mandated that teachers would no longer be able to score their own students' state assessments starting in January 2013. In response, 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. Our comparisons of pilot and non-pilot schools (see Appendix Table A5) and our conversations with NYCDOE officials suggests there was no systematic selection of pilot schools and no major differences in their observable characteristics.[29]

[28] See for example: http://www.nysedregents.org/integratedalgebra/811/ia-rg811w.pdf

[29] Specifically, while students in pilot schools are more likely to be white and less likely to be Hispanic than students in non-pilot schools, there are not statistically significant differences in 8th grade test scores or performance on core Regents exams in the baseline period. NYCDOE officials indicated there was no targeting or specific formula used

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 2010, prior to the exposé, and 2013, by which time all of New York City's high schools used centralized scoring. 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 A1). In June 2010, pilot and non-pilot schools manipulated 73.2 (se=0.87) and 62.4 (se=0.54) percent of in-range exams, respectively.[30] When schools were told not to re-score exams below the cutoff in June 2011, in-range manipulation dropped to 17.2 (se=0.46) and 13.0 (se=0.24) percent in pilot and non-pilot schools, respectively. 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, in-range manipulation dropped from 17.2 (se=0.46) percent to a statistically insignificant -0.01 (se=0.02) percent in pilot schools, but remained fairly steady in non-pilot schools at 12.3 (se=0.28) percent compared to 13.0 (se=0.24) percent in the prior year. In June 2013, when both pilot and non-pilot schools had adopted centralized grading, manipulation appears to have been completely eliminated. Of course, we cannot say whether centralization by itself would have eliminated manipulation in absence of the state's statements regarding re-scoring marginal exams, since we do not observe high schools operating under these conditions.

At the time that 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 retaken 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 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.

to select schools, and that recruitment was driven through high school "networks," i.e., mandatory but self-selected affiliations of 20-25 schools who collaborate on administrative tasks. 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. For example, among the roughly one-third of high schools that co-located in buildings with four or more high schools, building location explains almost 90 percent of the variation in participation.

[30] As can be seen in Appendix Figure A3, 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.

B. Estimating the Effect of Test Score Manipulation on Later Outcomes

The goal of our analysis is to estimate the impact of having a score inflated above cutoff c on outcomes such as high school graduation G_i. Two important issues complicate direct estimation of these effects. First, we do not observe the bias component $\phi(i,h,c)$ for any particular student or school, making it impossible to distinguish students who would have passed an exam on their own from students who only passed due to test score manipulation. Second, the bias component $\phi(i,h,c)$ is likely to be correlated with unobserved determinants of high school graduation such as student ability or motivation. For example, it is plausible that teachers are more likely to inflate the test scores of high-performing students that just had a "bad day" on a particular exam administration. The correlation between grading bias $\phi(i,h,c)$ and unobserved student traits ε_{iet} could potentially bias cross-sectional estimates even if the bias term $\phi(i,h,c)$ was observed.

Rather than try to distinguish individual students whose scores were manipulated, we use a difference-in-differences approach that 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 right around the manipulable range, pre- and 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. We focus on the margin of scoring 65 points or above, the most relevant score cutoff for high school graduation in this time period (see Appendix Table A1).

Formally, we estimate the reduced form impact of the grading reforms using the following specification:

$$y_{iet} = W_{iet}\alpha_{11} + X_i\theta_{11} + \gamma_{11} \cdot \mathbf{1}[69 \geq s_{iet} \geq 60] \cdot Reform_t + \varepsilon_{iet} \tag{11}$$

where y_{iet} is the outcome of interest for student i who took exam e at time t and $Reform_t$ is an indicator for exam e being taken following the grading reforms implemented in 2011. W_{iet} represents a set of control variables including high school fixed effects, exam by year effects, exam by cohort effects, fixed effects for having a score within 10-point score bins (i.e., 30-39, 40-49, etc.), and linear trends in year interacted with the 10-point Regents score bins. We add these latter controls to account for the increasingly stringent graduation requirements during this time period (see Appendix Table A1). X_i includes individual controls for gender, ethnicity, free lunch eligibility, and 8th grade test scores.[31] We also control for any 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

[31] Importantly, our results are quite stable even if we drop all of our controls for students' pre-existing observable characteristics. For example, in unreported results, we find that the estimated two-stage least squares coefficient for the effect of manipulation on graduating high school without controls is 0.213. Controlling for both observable characteristics and school fixed effects only slightly increases our estimate to 0.219 (Column 7 of Table 4). These results suggest that, under the reasonable assumption that students' unobservable characteristics are correlated with observable characteristics (e.g., Altonji, Elder, and Taber 2005), our results are unlikely to be driven by student selection into our sample.

the exam. Results are similar if we drop all students scoring at or below 59 (see Section V.D). We stack student outcomes across Regents exams (i.e., include multiple exams for each student) and adjust our standard errors for clustering at both the student and school level.

The parameter γ_{11} can be interpreted as the differential effect of the reform on students scoring between 60-69 on an exam compared to the omitted group of students scoring between 70-100 on the same exam. As the reform eliminated manipulation, we might expect our estimates of γ_{11} to be negative for outcomes such as passing the exam and graduating from high school. However, the key identifying assumption is that changes in outcomes for students scoring between 60-69 at the time of the reforms would have been identical to changes for students between 70-100 in the absence of the Regents grading reforms (and conditional on our other controls). 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. For example, our identifying assumption would be violated if unobservably better students initially scoring a 69 had their scores manipulated to 70 prior to the reform (but not after the reform). However, we see no evidence of test score manipulation at the 69 to 70 score threshold. In Section V.D, we also present several tests in support of our approach.

We also present two-stage least squares estimates that provide the local average treatment effect of passing a Regents exam due to test score manipulation. The first stage regression takes the form:

$$Pass_{iet} = W_{iet}\alpha_{12} + X_i\theta_{12} + \gamma_{12} \cdot \mathbf{1}[69 \geq s_{iet} \geq 60] \cdot Reform_t + \varepsilon_{iet} \tag{12}$$

where γ_{12s} measures the effect of the grading reform on the probability of scoring at 65 points or above on the Regents exam. The associated second stage regression takes the form:

$$y_{iet} = W_{iet}\alpha_{13} + X_i\theta_{13} + \gamma_{13} \cdot Pass_{iet} + \varepsilon_{iet} \tag{13}$$

Data limitations prevent us from measuring impacts on later outcomes such as college graduation or labor market earnings. 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, Fort, and Weber 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 forthcoming). Conversely, an important study by Clark and Martorell (2014) finds negligible returns to passing "last chance" high school exit exams in the state of Texas. Note that Texas' last chance exam takers are an extremely negatively selected sample, i.e., students who failed high school exit exams multiple times and exit high school regardless of the outcome of their last attempt, and the results from their study may not be applicable to our setting.

C. Main Results

We begin with an examination of student outcomes over the period between 2004 and 2013 for those scoring between 60-69 on an exam (i.e., students likely to be affected by test score manipulation) and

between 70-100 on the same exam (i.e., students unlikely to be affected by test score manipulation) on Regents exams. We focus on students entering high school between 2003-2004 and 2009-2010 where we observe high school graduation. We also focus on the Science, Math, and Global History core exams which are typically taken first in 9th or 10th grade; as will be shown below, the policy reforms made it more difficult to pass Regents exams, and this may change the composition of students who "make it" to the English and U.S. History exams taken closer to graduation.[32] Using this sample, Figure 5 presents coefficient estimates from a regression of student outcomes on interactions of scoring between 60-69 and each year, which allows us to examine pre- and post-reform trends in how outcomes evolve differentially over time for students scoring between 60-69. Recall that we stack student outcomes across Regents exams (i.e., include multiple exams for each student) and adjust our standard errors for clustering at both the student and school level. Our estimates can therefore be interpreted as the impact of having <u>one</u> exam score manipulated (controlling for performance on the other exams in our sample).

Panel A of Figure 5 shows, not surprisingly, a sudden drop of almost 15 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 V.A, that the Regents grading reforms significantly decreased test score manipulation. We also see clear upwards trends in the probability of scoring 65 or above during the pre-reform period. As discussed below, this upwards trend is consistent with the greater emphasis on the 65 point cutoff during this time period and supports the inclusion of linear trends interacted with 10-point score bins in our difference-in-differences regression specifications. In Panel B of Figure 5, we show that the fall in pass rates also coincides with a sharp increase in test retaking, also just under 15 percentage points, 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 first attempt. We see a similar sudden drop, but of a smaller magnitude of roughly 6-7 percentage points, suggesting that the majority (but clearly not all) of these marginal students were able to pass the exam on a subsequent attempt.

Panel D of Figure 5 presents results for high school graduation, measured by possession of a non-GED high school diploma within four years. There is no trend in the coefficient values during the pre-reform period, but graduation rates suddenly drop in 2011 by about 3 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

[32] Consistent with this argument, we find no systematic changes in the characteristics of students taking the Living Environment, Math A/Integrated Algebra, and Global History exams following the reforms (see Appendix Table A6). In contrast to these early exams, we find that students taking the English and U.S. History exams have significantly higher 8th grade test scores following the grading reforms. These results are available on request. For completeness, we present estimates for all core exams together and each core exam separately in Appendix Table A7. The effect of manipulation on high school graduation is somewhat larger for Living Environment, the first exam taken by most students, and much smaller for the optional non-core exams. 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 A7 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.

just below the 65 cutoff on the Regents core exams. While most of these students still eventually graduated, 20-25 percent of them appear to have been unable to pass the exam on a subsequent attempt and thus could not graduate from high school.

Panels E and F show trends of taking the Regents diploma and Advanced Regents diploma requirements, both measures of the potential "quality" of high school diploma receipt. While there is little trend in the taking of the Regents diploma requirements over time, there is a sharp jump upward in the taking of the Advanced Regents diploma requirements in the post-policy years, suggesting that many of the marginal students forced to retake Regents exams (and coursework) ended up with a higher distinction in the long-run.[33]

Regression estimates of Equation (11), shown in Table 4, are consistent with the patterns observed in Figure 5. We report the coefficient on the interaction between scoring between 60-69 and whether the exam was 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 2 and 3) show that the grading reforms decrease the probability of scoring 65 or above by 14.1 (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 4 and 5) indicate that students scoring between 60-69 are roughly 3.1 percentage points (se=0.5) less likely to graduate high school following the grading reforms. Two-stage least squares estimates (Columns 6 and 7) suggest that the local average treatment effect of passing a Regents exam due to test score manipulation is an increase in the probability of graduating from high school of 21.9 percentage points (se=3.3). This is a substantial effect, given an exam-weighted mean graduation rate of 78.8 percent for students in our sample. In other words, consistent with the patterns seen in Figure 5, we estimate that almost 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 two-stage least squares estimates for mutually exclusive student subgroups.[34] Effects on high school graduation are larger for female students, white and Asian students, students eligible for free lunch, and students with higher baseline test scores. For example, we estimate that manipulation increases the probability that white and Asian students graduate high school by 28.2 (se=5.9) percentage points, 8.1 percentage points more than black and Hispanic students. Manipulation also has a 16.4 percentage point larger effect for students eligible for free lunch, and a 9.8 percentage point larger effect for students with above median 8th grade test scores.

[33] We focus on course taking rather than passing because our identification assumption – that students scoring above the 60-69 range are largely unaffected by the reform – would not hold for the passing outcome. Students in our control group, who score above the 60-69 range on the three earliest core regents exams, are in fact quite likely to score in the 60-69 range on later exams such as advanced math and science, and these exams are also subject to the reform. For completeness, the appendix shows results for passing (as opposed to taking) the advanced coursework are similar, if somewhat larger and more precisely estimated (see Appendix Table A8).

[34] Unfortunately, data on absences and disciplinary incidents is not available for the most recent cohorts, so we cannot test heterogeneity along this dimension.

In Panel B of Table 4, we examine the effects of manipulation on the potential quality of the high school degree passed on course taking, (i.e., taking the requirements for Regents and Advanced Regents diplomas). Having an exam manipulated increases the probability of taking the requirements for a Regents diploma, the lowest diploma for most students during this time period, by a statistically insignificant 3.3 (se=8.3) percentage points. However, consistent with Figure 5, the probability of taking the requirements for the Advanced Regents diploma decreases by 16.4 (se=10.3) percentage points. Panel B of Table 5 shows that the negative effects of manipulation on Advanced Regents taking are considerably larger for male students, free lunch students, and students with above median 8th grade test scores.

We shed further light on the mechanisms underlying the Advanced Regents result in four ways. First, we find a much larger negative effect of manipulation on meeting the requirements for an Advanced Regents diploma for the Math A/Integrated Algebra exam, consistent with the importance of advanced science and math requirements, and a small and statistically insignificant effect for the English and Social Science exams (Panel B of Appendix Table A7). Second, we show that having a Math A/Integrated Algebra score manipulated decreases the probability of taking the science and math requirements for an Advanced Regents diploma (Panel C of Appendix Table A7). Third, it is notable that Global History is typically taken before the advanced science and math exams, while English and U.S. History are usually taken concurrently, so that having to re-take Global History is more likely to crowd out these advanced courses. Consistent with this notion, we find that manipulation on the Global History exam has a positive impact on passing a physical science or advanced math exam (Appendix Table A7), while manipulation on English or U.S. History exams does not. Finally, we present direct evidence that a significant fraction of students who fail the exam because their scores are no longer manipulated end up re-taking the exam and do significantly better on the second administration, i.e., having a score manipulated significantly decreases the probability of scoring above 70, 75, and 80 points within the next calendar year (Panel A of Appendix Table A8).

These results support the idea that test score manipulation has somewhat nuanced effects. Students on the margin of dropping out may be "helped" by test score manipulation because they are not forced to retake and pass an exam or a course required 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 gain a solid foundation in the introductory material that the more advanced coursework requires.

D. Robustness Checks

Alternative Attainment Measures: One way that our setting is quite different than the "last chance" exams examined by Clark and Martorell (2014) is that most Regents exams are taken well before the end of high school, and failing may affect whether students progress towards graduation or drop out, i.e., potentially altering their educational investments. We therefore examine two additional measures of secondary school attainment: the highest high school grade in which the student is

enrolled in NYC public schools and the number of years the student is enrolled in NYC public high 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? Attainment based on highest grade does not count repetition as attainment, while attainment based on years enrolled counts repetition fully. Two-stage least squares estimates (Panel D of Appendix Table A8) show large effects of manipulation on both of these attainment measures: 0.41 (se=0.04) grade levels and 0.78 (se=0.05) school years. These results suggest that manipulation lengthened the extent of secondary educational investment for marginal students and did not just provide diplomas to students who were already at the very end of their high school careers.

Our main estimates focus on four-year graduation. However, it is possible that test score manipulation reduces time to graduation while having little impact on longer run educational attainment. Panel E of Appendix Table A8 shows that having a score manipulated also increases the probability of graduating from high school in five years by 23.0 percentage points (se=3.4), and the probability of graduating from high school in six years by 16.2 percentage points (se=2.9). While it is possible that the reforms also affected GED receipt, only about 1 percent of students in our sample appear to receive a GED within six years of starting high school. This suggests that either GED is an unimportant margin or, more likely, our data on GED receipt is of poor quality.

Alternative Specifications: Appendix Table A9 presents estimates using a variety of specifications and instruments to assess the robustness of our main two-stage least squares results further. Column 1 drops exams with scores between 0-59, limiting the control group to students scoring between 70-100. Column 2 also drops exams with scores between 80-100, further limiting the control group to students scoring between 70-79. Column 3 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 4 adds an interaction with an indicator for participating in the centralized grading pilot program for a total of six instrumental variables. None of the point estimates are meaningfully different from our preferred estimates in Table 4.

Placebo Estimates: To test for potential sources of bias in our main specification, we estimate a series of 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 A6. We find a statistically significant coefficient for only one out of seven student characteristics (a 1.97 (se=1.14) percentage point increase in students eligible for free or reduced-price lunch), and all of the estimates are small and economically trivial. We also examine differences in predicted outcomes (i.e., graduation, Regents, and Advanced Regents) using all of the baseline characteristics listed in Panel A of Appendix Table A6. Consistent with our identifying assumption, we find no statistically significant differences following the elimination of re-scoring.

E. Estimates using Across-School Variation

While we would like to examine how college outcomes were affected by the Regents scoring reforms, we only have college enrollment data on cohorts entering high school before 2005-2006.[35] However, we can implement a cross-sectional methodology motivated by the approach used by Diamond and Persson (2016) to examine these earlier cohorts. We present the details of this analysis in Appendix D and summarize here.

Intuitively, the cross-sectional strategy asks whether the outcomes of students with scores in the manipulable range (60-69), relative to those enrolled in the same school but with scores outside the range, are systematically better in schools where manipulation is more prevalent. The key identifying assumption is that, conditional on observables, the within-school differences in outcomes between students scoring "in-range" and those scoring "out-of-range" is uncorrelated with any unobserved factor other than test score manipulation. The identifying assumption for this cross-sectional approach could be violated in fairly plausible circumstances, for example, if teachers in schools that manipulate higher fractions of marginal exam scores are more generally concerned with improving the educational outcomes of marginal students. While this strategy relies on much stronger assumptions than our difference-in-differences approach, we find support for it using placebo tests in specifications with school fixed effects.

Appendix Table A10 presents estimates from this cross-sectional approach. We focus on the 65 cutoff to make the estimates more directly comparable to the difference-in-differences estimates presented above. For the same reason, we focus on graduating with a Regents or Advanced Regents diploma in the proceeding analysis, as the 65 cutoff is somewhat less relevant for the any high school graduation measure in the cohorts where we observe college enrollment outcomes. The results largely follow our preferred difference-in-differences estimates described above, although some of the point estimates are imprecisely estimated and not statistically distinguishable from zero. The probability of meeting the requirements for a Regents diploma, the lowest diploma type available to students in our preferred difference-in-differences specification, increases by 20.0 (se=3.5) percentage points. In contrast, the probability of meeting the requirements for an Advanced Regents diploma decreases by a statistically insignificant 6.4 (se=4.9) percentage points. These results are both reassuringly similar to our difference-in-differences estimates for the same outcomes.

Turning to college enrollment, we find that having a score manipulated decreases the probability of enrolling in any college by 4.5 (se=2.3) percentage points, with the effect driven by the 5.6 (se=2.4) percentage point decrease in the probability of enrolling in a two-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. These results suggest that, in addition to the Advanced Regents results, that test score manipulation may have harmful impacts for at least some students on the margin of passing an exam.

[35] Specifically, we restrict the sample to students entering high school between 2003-2004 and 2005-2006 who take at least one core Regents exams between 2004-2010 – the sample where we observe both high school graduation and college enrollment for all students.

F. Aggregate 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 IV.C, our long-run estimates suggest that test score manipulation also had important distributional effects. To quantify these effects, we multiply the two-stage least squares estimate from Table 5 by the subgroup-specific total manipulation estimates from Appendix Figure A7. We calculate all numbers at the student level, not the student by exam level.

These back-of-the-envelope calculations suggest that test score manipulation has 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. In other words, test score manipulation allowed about 1,000 additional students to graduate each year from the New York City school system.[36]

In contrast, our results suggest that test score manipulation only modestly affected 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 15.9 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 25.2 percentage points.

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. Our findings indicate that test score manipulation was sufficiently widespread that it had significant effects on the overall performance of students across and within New York public schools. For example, our estimates imply that, without manipulation, the graduation rate of our sample would have decreased from 76.6 percent to 75.3 percent.

Exploiting a series of exogenous grading reforms, we find that test score manipulation has a substantial impact on educational attainment for students on the margin of passing an exam. For these marginal students, having a score manipulated above a cutoff increases the probability of graduating from high school by approximately 22 percentage points, or nearly 28 percent. In other words, while about 80 percent of marginal students would have eventually passed the exam without test score manipulation, a significant number would have dropped out of school. However, we also find that having a score manipulated above a cutoff leads a subset of marginal students, who no longer have to study for and retake the exam, to opt out of more advanced coursework. These mixed results serve as an important reminder that lowering the bar for high school graduation

[36] 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.

can increase attainment for students who would otherwise struggle, but decrease attainment for students who may benefit from a "push" towards higher achievement.

Why did the practice of manipulation of Regents exams become so widespread prior to the reforms? While we are unable to answer this question in a definitive manner, we are able to exclude a number of potential causes, such as test-based school accountability systems or test-based teacher incentive programs. A remaining explanation, consistent with the evidence, is that manipulation is simply driven by teachers' common desire to help their students avoid the costs associated with failing an exam.

An important limitation of our analysis is that we are only able to estimate the effect of eliminating manipulation on educational attainment. While we find clear evidence that manipulation leads many marginal students to spend more time in school and graduate from high school, we also find that a subset of these students are less likely to take more advanced courses. There may also be important general equilibrium effects of eliminating test score manipulation, such as changing the signaling value of course grades or a high school diploma. Estimating the long-run impacts of manipulation on labor market outcomes remains an important area for future research.

References

[1] Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." Journal of Political Economy, 113(1): pp. 151-184.

[2] Apperson, Jarod, Carycruz Bueno, and Tim R. Sass. 2016. "Do the Cheated Ever Prosper? The Long-Run Effects of Test-Score Manipulation by Teachers on Student Outcomes." Unpublished Working Paper.

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

[4] Angrist, Joshua, Erich Battistin, and Daniela Vuri. 2014. "In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno." Forthcoming at American Economic Journal: Applied.

[5] 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.

[6] Borcan, Oana, Mikael Lindahl, and Andreea Mitrut. 2017. "Fighting Corruption in Education: What Works and Who Benefits?" American Economic Journal: Economic Policy, 9(1): 180-209.

[7] 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.

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

[9] Canaan, Serena, and Pierre Mouganie. "Returns to Education Quality for Low-Skilled Students: Evidence from a Discontinuity." Forthcoming at Journal of Labor Economics.

[10] 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.

[11] Chetty, Raj, John Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." American Economic Review, 104(9): 2593-2632.

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

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

[14] Cunha, Flavio and James J. Heckman. 2010. "Investing in Our Young People." NBER Working Paper No. 16201.

[15] Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." Econometrica, 78(3): 883-931.

[16] Cullen, Julie, and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." NBER Working Paper No. 12286.

[17] Dee, Thomas, Brian A. 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.

[18] Diamond, Rebecca, and Petra Persson. 2016. "The Long-term Consequences of Teacher Discretion in Grading of High-stakes Tests." NBER Working Paper No. 22207.

[19] Dustmann, Christian, Patrick Puhani, and Uta Schönberg. "The Long-Term Effects of Early Track Choice." Forthcoming in Economic Journal.

[20] 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 Economics.

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

[22] Fryer, Roland. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." Journal of Labor Economics, 31(2): 373-407.

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

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

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

[26] 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.

[27] Jacob, Brian A., and Jesse Rothstein. 2016. "The Measurement of Student Ability in Modern Assessment Systems." Journal of Economic Perspectives, 30(3): 85-107.

[28] 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, 78(4): 733-740.

[29] Lavy, Victor. 2008. "Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment." Journal of Public Economics, 92(10-11): 2083-2105.

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

[31] 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.

[32] National Research Council. 2011. Incentives and Test-Based Accountability in Education. Committee on Incentives and Test-Based Accountability in Public Education, M. Hout and S.W. Elliott, Editors. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

[33] 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.

[34] Neal, Derek. 2013. "The Consequences of Using One Assessment System to Pursue Two Objectives." NBER Working Paper No. 19214.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

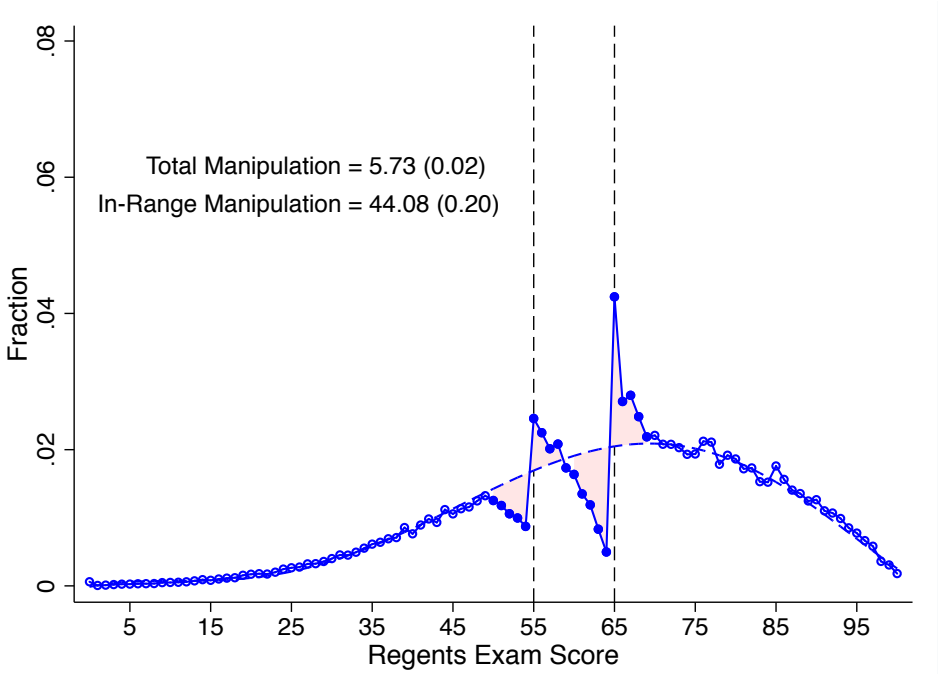
[40] Papay, John P., John B. Willett, and Richard J. Murnane. 2011. "Extending the Regression-Discontinuity Approach to Multiple Assignment Variables." Journal of Econometrics, 161(2): 203-207.

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

[42] Terrier, Camille. 2016. "Boys Lag Behind: How Teachers' Gender Biases Affect Student Achievement." IZA Working Paper 10343.

[43] Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." The Economic Journal, 113(485): F3-F33.

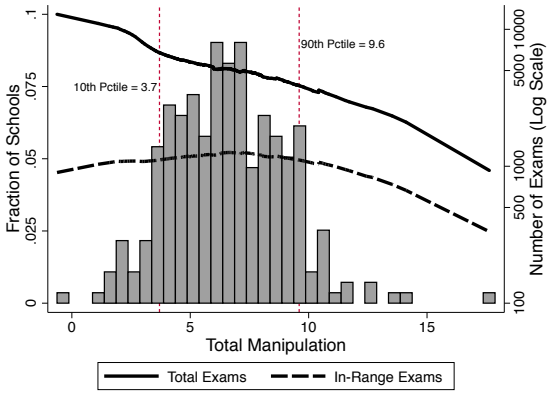
Figure 1: Test Score Distributions for Core Regents Exams, 2004-2010



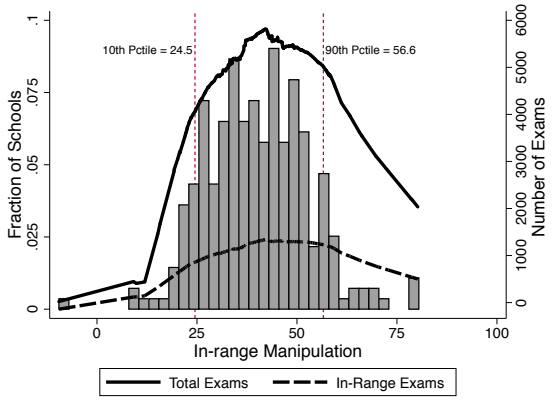
Note: This figure shows the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Core exams include English Language Arts, Global History, U.S. History, Math A/Integrated Algebra, and Living Environment. We include the first test in each 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. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III. 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: Distribution of School Manipulation Estimates, 2004-2010

(a) Total Manipulation



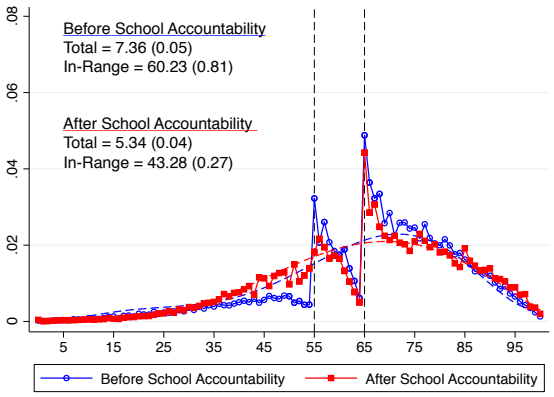
(b) In-Range Manipulation



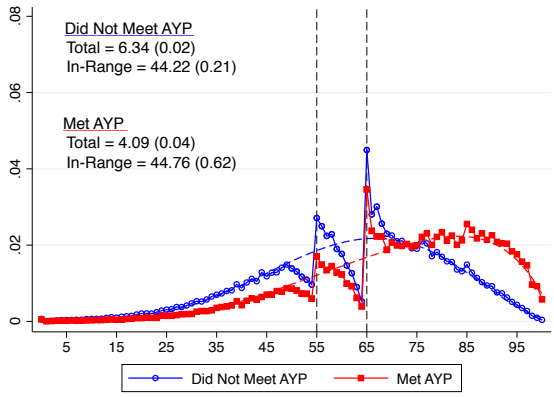
Note: These figures show the distribution of school manipulation estimates for core Regents exams around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Panel (a) is total manipulation estimates aggregated across both cutoffs. Panel (b) is in-range manipulation estimates averaged across both cutoffs. The smooth lines show the relationship between the number of both total and in-range exams and manipulation at the school level. See the text for additional details on the sample and empirical specification.

Figure 3: Results by School Accountability Pressure, 2001-2010

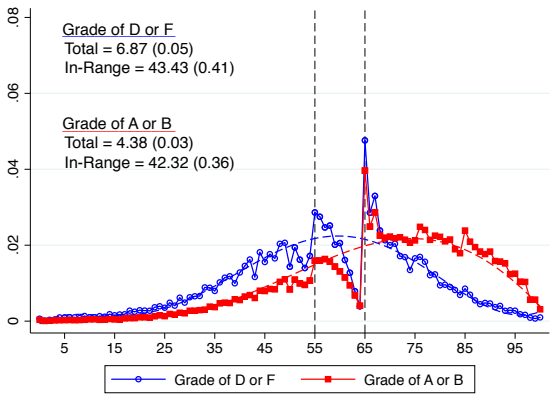
(a) Before and After School Accountability



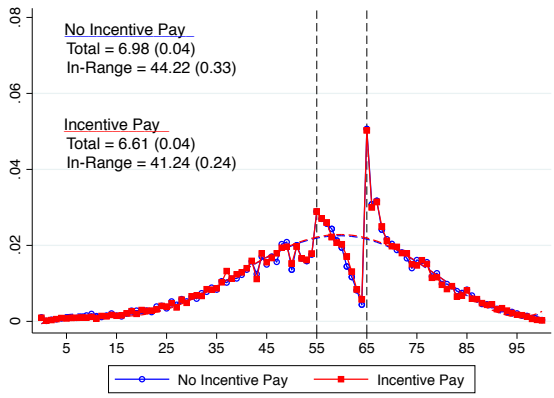
(b) NCLB Adequate Yearly Progress



(c) NYC Accountability Grade



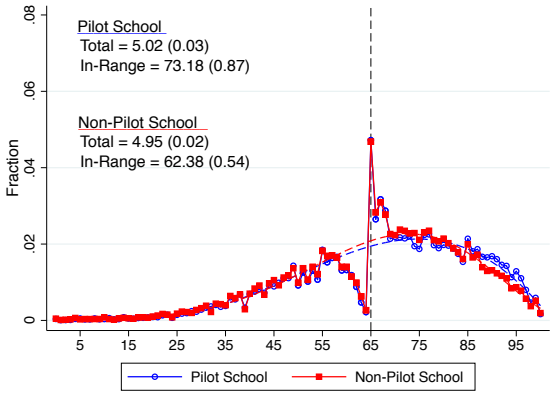
(d) Teacher Incentive Pay Experiment



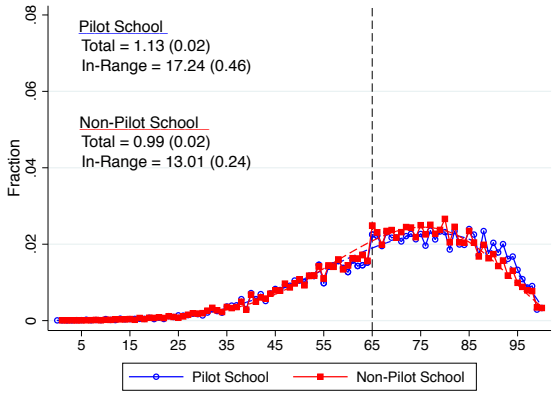
Note: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers. Panel (a) plots non-math core exams taken in 2000-2001 before the implementation of NCLB and the NYC Accountability System and in 2008-2010 after the implementation of both 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.

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

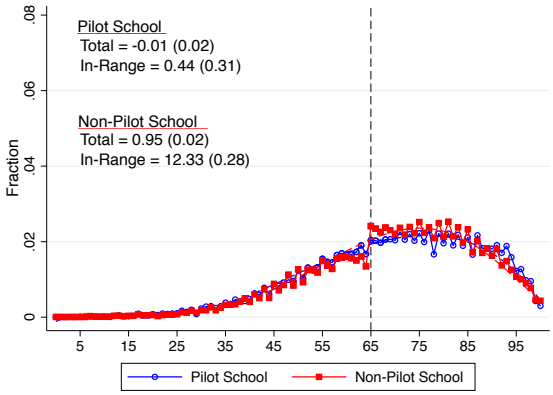
(a) 2010: Re-Scoring and Decentralized Grading



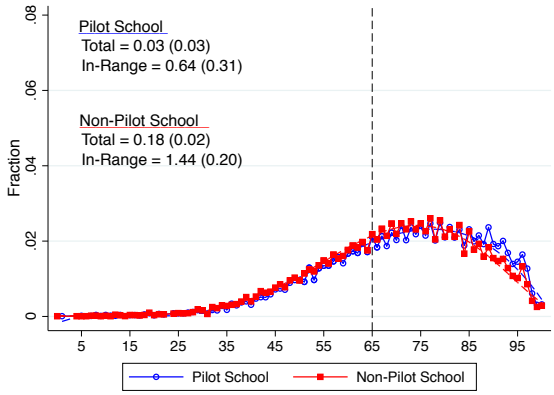
(b) 2011: No Re-Scoring and Decentralized Grading



(c) 2012: No Re-Scoring and Pilot of Centralized Grading

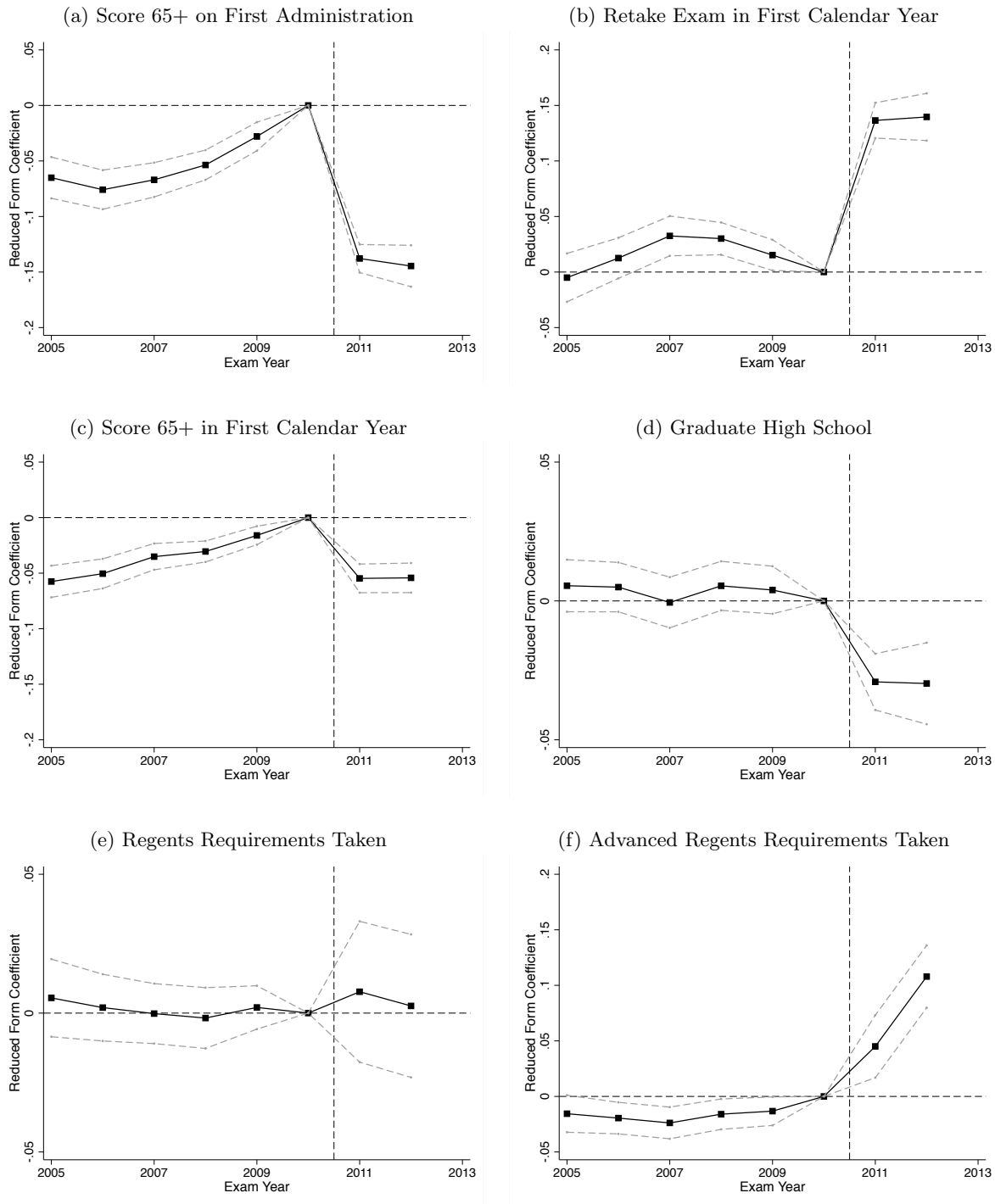


(d) 2013: No Re-Scoring and Centralized Grading



Note: 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



Note: 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. See the Table 4 notes for additional details.

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	0.602	0.589	0.483
Above Median 8th Test Scores	0.516	0.053	0.331	0.886
Core Regents Performance:				
Living Environment	69.622	38.835	63.214	82.725
Math A	69.835	40.459	65.040	84.522
Int. Algebra	66.052	40.830	61.947	79.990
Global History	67.814	32.559	60.165	86.376
Comprehensive English	69.422	29.914	63.111	85.255
U.S. History	72.499	33.010	65.201	88.994
High School Graduation:				
High School Graduate	0.730	0.129	0.672	0.926
Local Diploma	0.041	0.035	0.070	0.007
Regents Diploma	0.503	0.178	0.599	0.430
Advanced Regents Diploma	0.232	0.001	0.042	0.507
College Enrollment:				
Any College	0.524	0.132	0.457	0.716
Any Two-Year College	0.193	0.095	0.243	0.123
Any Four-Year College	0.388	0.047	0.272	0.656
Students	514,632	36,677	295,260	182,695

Note: This table reports summary statistics for students in New York City taking a core Regents exam between 2004-2010. High school graduation records are only available for cohorts entering high school between 2001-2010 (N = 457,587). High school diploma records are only available for cohorts entering high school between 2007-2009 (N = 143,222). College records are only available for students entering high school before 2001-2005 (N = 256,177). 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)	(9)	(10)
Percent Black/Hispanic	4.335***				2.573***	0.871				3.734
	(0.548)				(0.836)	(2.942)				(4.880)
Percent Free Lunch		2.555***			−1.432		−2.944			−3.476
		(0.845)			(0.883)		(4.159)			(5.154)
Standardized 8th Score			−1.332***		−1.128***			−1.330*		−1.800**
			(0.117)		(0.150)			(0.684)		(0.877)
Enrollment (in 1000s)				−0.333***	0.199				1.093**	1.822**
				(0.111)	(0.122)				(0.544)	(0.713)
Constant	3.142***	4.970***	6.639***	7.150***	5.131***	40.609***	43.116***	41.381***	39.329***	37.351***
	(0.448)	(0.540)	(0.111)	(0.240)	(0.848)	(2.404)	(2.656)	(0.650)	(1.172)	(4.954)
R^2	0.185	0.032	0.319	0.032	0.343	0.000	0.002	0.014	0.014	0.046
Dep. Var. Mean	6.551	6.551	6.551	6.551	6.551	41.293	41.293	41.293	41.293	41.293
Observations	279	279	279	279	279	279	279	279	279	279

Note: This table reports estimates from a regression of school manipulation on average school characteristics. School-exam-administration-cutoff-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. School characteristics are measured using the average for all enrolled students between 2004-2010, including non-exam takers. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Table 3: School-Subject Manipulation and Teacher Turnover

	In-Range Manipulation			
	(1)	(2)	(3)	(4)
Lagged Manipulation	0.490***	0.420***	0.090	−0.010
	(0.080)	(0.081)	(0.069)	(0.087)
Fraction of Teachers Present in Both Periods			13.122**	2.953
			(5.747)	(8.235)
Lagged Manipulation x Fraction Present			0.769***	0.829***
			(0.200)	(0.216)
Constant	41.762***	42.233***	34.343***	40.253***
	(2.374)	(1.881)	(3.397)	(4.967)
R^2	0.456	0.683	0.487	0.700
Dep. Var. Mean	46.888	46.888	46.888	46.888
Observations	984	984	984	984
School Fixed Effects	No	Yes	No	Yes

Note: This table reports estimates from a regression of school x subject in-range manipulation between 2007-2009 on school x subject lagged in-range manipulation between 2004-2006 and subject effects. All specifications above use the number of exams in-range of manipulation as weights. The fraction of teachers in each subject who were employed during both periods is calculated by dividing teachers based on license area: English licenses for the English exam, Mathematics for the Math A and Integrated Algebra exams, Social Studies for the Global and US History Exams, and Biology, Chemistry, Earth Science, Physics, and General Science for the Living Environment Exam. We drop teachers who provide instruction only to special education or bilingual populations. Standard errors are clustered by school. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Table 4: Effect of Test Score Manipulation on High School Graduation

	Dep. Var. Mean	First Stage		Reduced Form		2SLS	
Panel A: High School Graduation	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Graduate High School	0.788	−0.141***	−0.141***	−0.030***	−0.031***	0.211***	0.219***
	(0.409)	(0.007)	(0.007)	(0.005)	(0.005)	(0.035)	(0.033)
Panel B: Diploma Requirements							
Regents Requirements Taken	0.875	−0.141***	−0.141***	−0.005	−0.005	0.033	0.033
	(0.331)	(0.007)	(0.007)	(0.012)	(0.012)	(0.088)	(0.083)
Adv. Regents Requirements Taken	0.355	−0.141***	−0.141***	0.025*	0.023*	−0.176*	−0.164*
	(0.479)	(0.007)	(0.007)	(0.015)	(0.015)	(0.108)	(0.103)
Observations	1,002,804	1,002,804	1,002,804	1,002,804	1,002,804	1,002,804	1,002,804
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

Note: 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 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, exam by year-of-test effects, and cohort by year-of-test effects. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exams and cluster standard errors at the individual and school levels. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent 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
Panel A: High School Graduation	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Graduate High School	0.188***	0.251***	0.201***	0.282***	0.241***	0.077	0.127***	0.225***
	(0.050)	(0.042)	(0.035)	(0.059)	(0.036)	(0.066)	(0.040)	(0.068)
	[0.755]	[0.817]	[0.742]	[0.883]	[0.787]	[0.790]	[0.687]	[0.930]
Panel B: Diploma Requirements								
Regents Requirements Taken	0.067	0.008	0.043	0.063	−0.021	0.103	0.040	0.001
	(0.091)	(0.084)	(0.062)	(0.111)	(0.092)	(0.080)	(0.048)	(0.116)
	[0.862]	[0.886]	[0.857]	[0.913]	[0.867]	[0.887]	[0.842]	[0.939]
Adv. Regents Requirements Taken	−0.249**	−0.085	−0.165**	−0.109	−0.215**	−0.096	−0.142*	−0.286**
	(0.100)	(0.116)	(0.075)	(0.140)	(0.109)	(0.112)	(0.075)	(0.145)
	[0.342]	[0.368]	[0.244]	[0.586]	[0.328]	[0.398]	[0.159]	[0.629]
Observations	472,712	530,092	670,145	322,935	605,301	397,430	462,417	370,267
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

Note: 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, exam by year-of-test effects, and cohort by year-of-test effects. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exams and cluster standard errors at the individual and school levels. The sample mean for each subgroup is reported in brackets. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix A: Additional Results
[NOT FOR PUBLICATION]

Appendix Figure A1: Conversion of Multiple Choice Items and Essay Ratings to Scale Scores

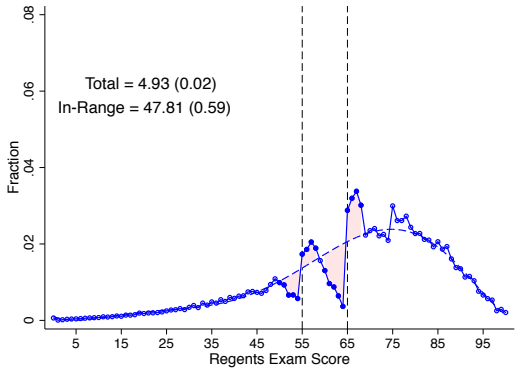
June 2009 English Exam -- Manipulable Scores Shown in Bold

Number Correct on Multiple Choice Items	Cumulative Essay Rating									
	0	1	…	15	16	17	18	19	…	24
0	0	1	…	30	34	38	41	45	…	**65**
1	1	1	…	32	36	40	43	47	…	**67**
2	1	1	…	34	38	41	45	49	…	69
3	1	2	…	36	40	43	47	**51**	…	70
4	1	2	…	38	41	45	49	**53**	…	72
5	2	2	…	40	43	47	**51**	**55**	…	74
6	2	2	…	41	45	49	**53**	**57**	…	76
7	2	3	…	43	47	**51**	**55**	59	…	77
8	2	3	…	45	49	**53**	**57**	**61**	…	79
9	3	4	…	47	**51**	**55**	59	**63**	…	80
10	3	5	…	49	**53**	**57**	**61**	**65**	…	82
11	4	6	…	**51**	**55**	59	**63**	**67**	…	84
12	5	7	…	**53**	**57**	**61**	**65**	69	…	85
13	6	8	…	**55**	59	**63**	**67**	70	…	86
14	7	9	…	**57**	**61**	**65**	69	72	…	88
15	8	10	…	59	**63**	**67**	70	74	…	89
16	9	11	…	**61**	**65**	69	72	76	…	90
17	10	13	…	**63**	**67**	70	74	77	…	92
18	11	14	…	**65**	69	72	76	79	…	93
…	…	…	…	…	…	…	…	…	…	…
25	21	24	…	77	80	84	86	89	…	99
26	23	27	…	79	82	85	88	90	…	100

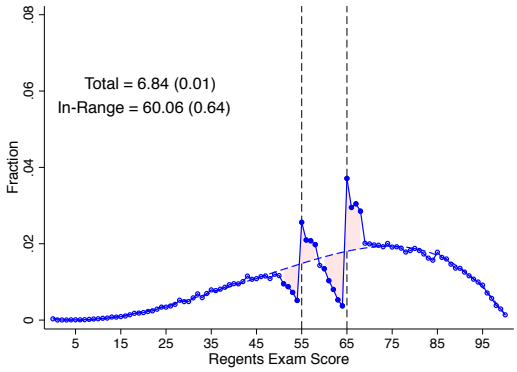
Note: This figure displays the official conversion chart for the English Language Arts Regents Exam for June 2009. For expositional purposes, the scale scores corresponding with essay points 2-14 and 20-23, and those corresponding with 19-24 multiple choice items correct, are omitted and represented by ellipsis. Cells with a white background are those scale scores for which a change in essay rating of 1 point would move the student across a cutoff at 55 or 65 scale score points.

Appendix Figure A2: Results by Subject, 2004-2010

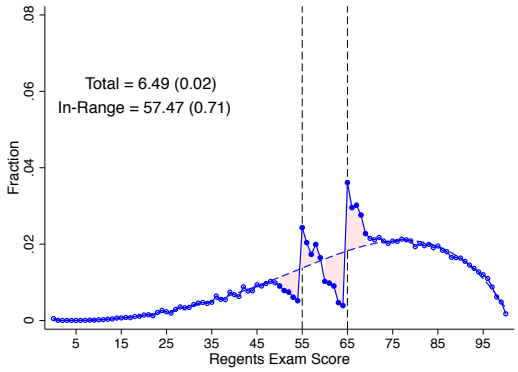
(a) English Language Arts, 2004-2010



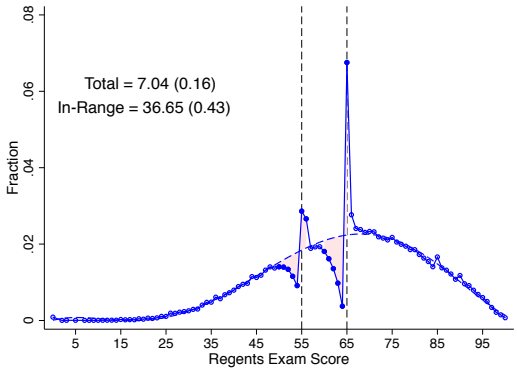
(b) Global History, 2004-2010



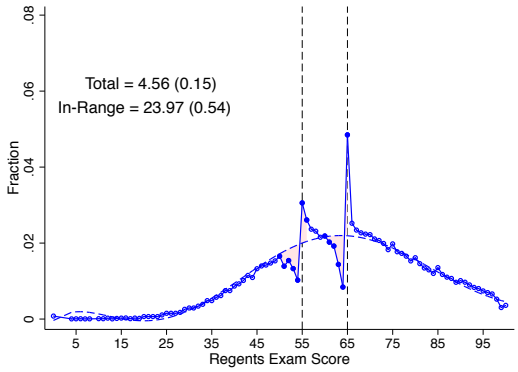
(c) U.S. History, 2004-2010



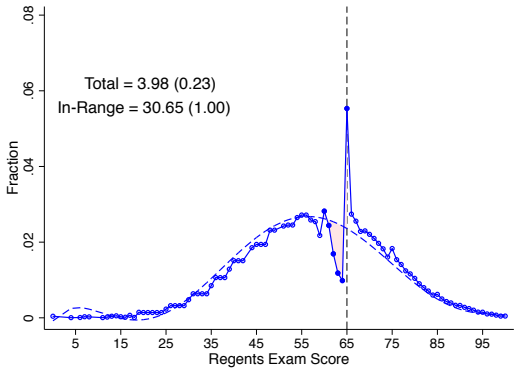
(d) Living Environment, 2004-2010



(e) Math A, 2004-2008



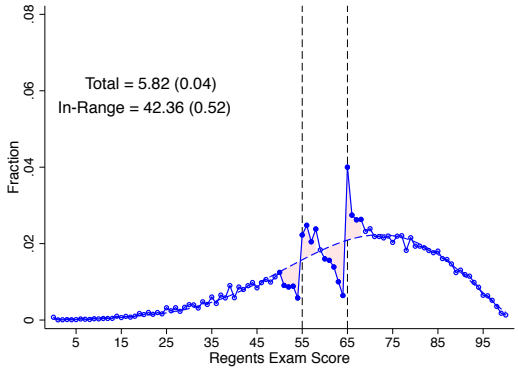
(f) Integrated Algebra, 2008-2010



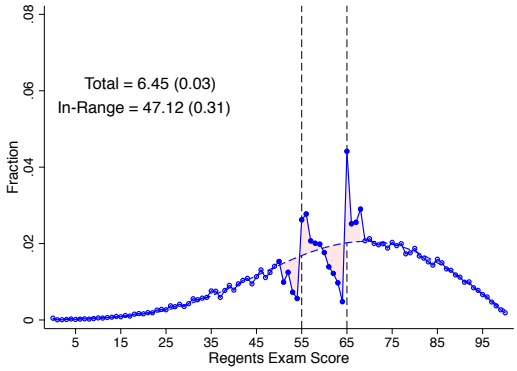
Note: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points 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. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III. 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 A3: Results by Year

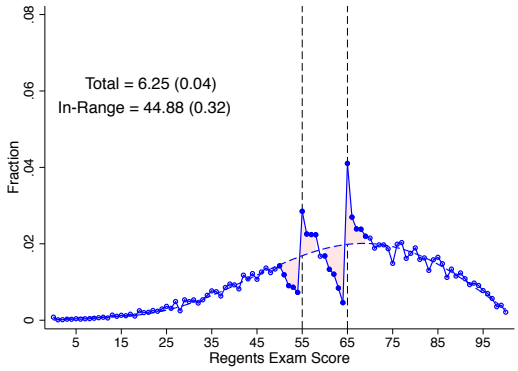
(a) 2004 Core Exams



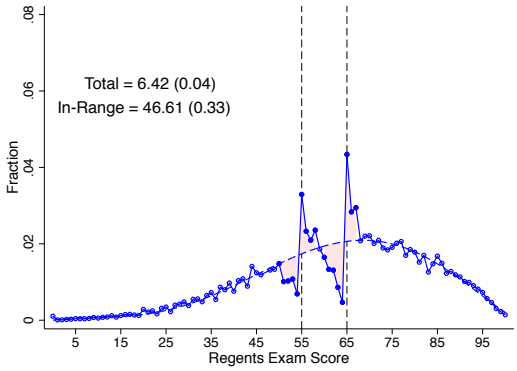
(b) 2005 Core Exams



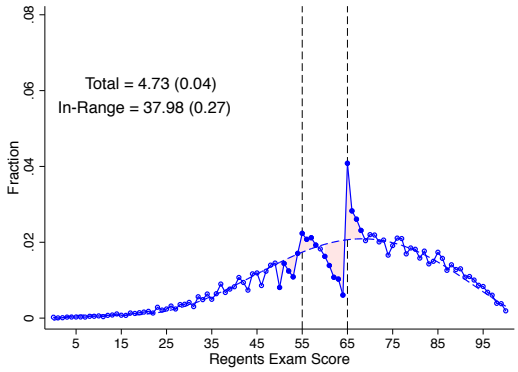
(c) 2006 Core Exams



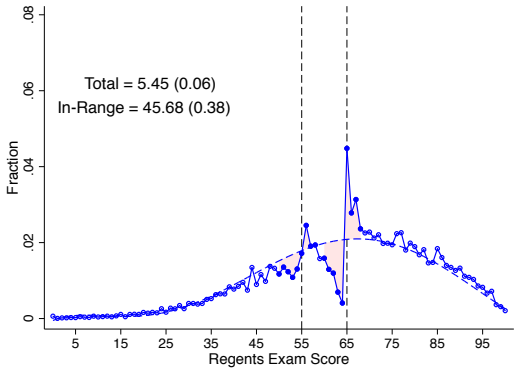
(d) 2007 Core Exams



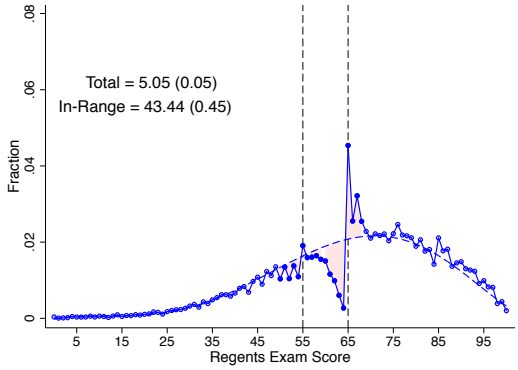
(e) 2008 Core Exams



(f) 2009 Core Exams



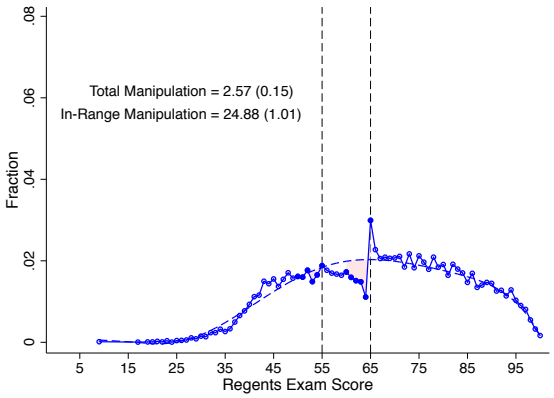
(g) 2010 Core Exams



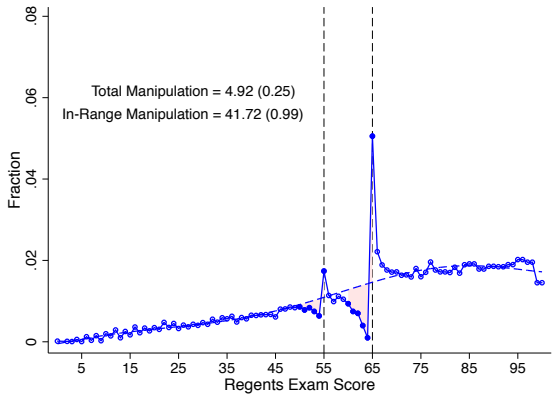
Note: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points 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. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III. 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 A4: Results for June 2001 Elective Exams With and Without Partial Credit

(a) Chemistry, No Partial Credit



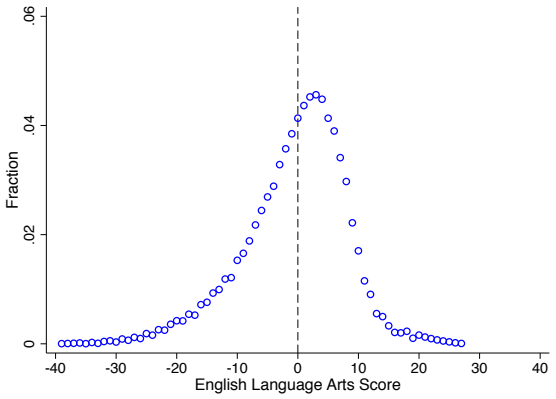
(b) Advanced Math, Partial Credit Allowed



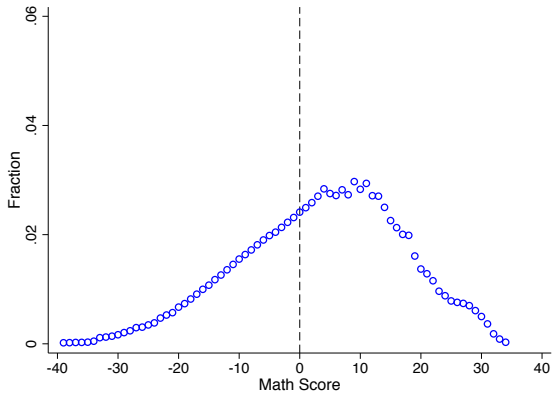
Note: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers in June 2001. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating 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. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III. 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 A5: Test Score Distributions for Centrally Graded Exams in Grades 3-8

(a) English Language Arts



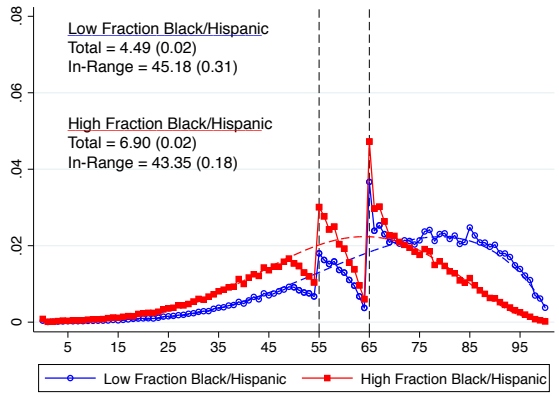
(b) Math



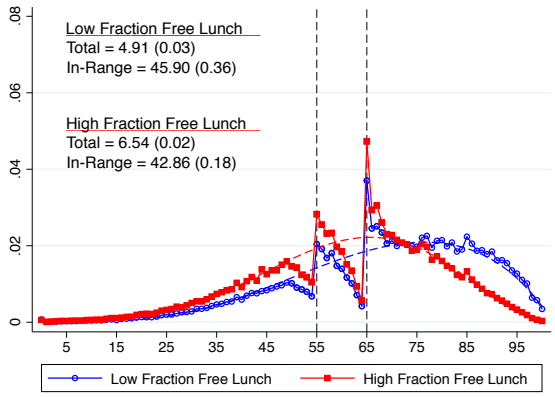
Note: These figures show the test score distribution around the proficiency score cutoff for New York City grade 3-8 test takers between 2004-2010. Each point shows the fraction of test takers in a score bin. See the data appendix for additional details on the variable definitions.

Appendix Figure A6: Results by School Characteristics, 2004-2010

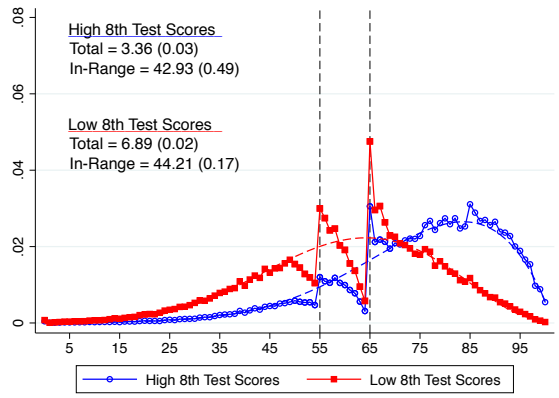
(a) High vs. Low Black/Hispanic Enrollment



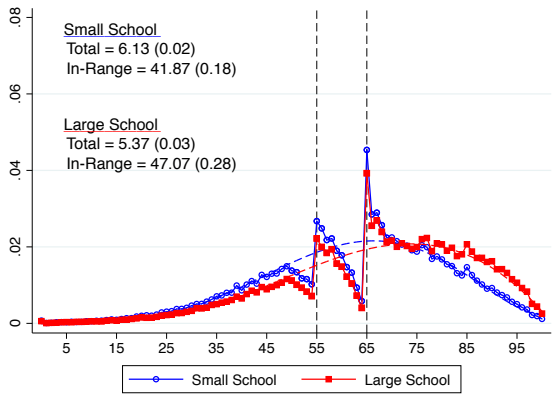
(b) High vs. Low Free Lunch Enrollment



(c) Low vs. High 8th Test Scores



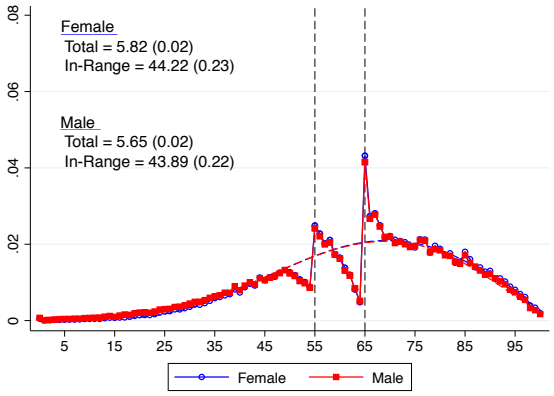
(d) Small vs. Large Schools



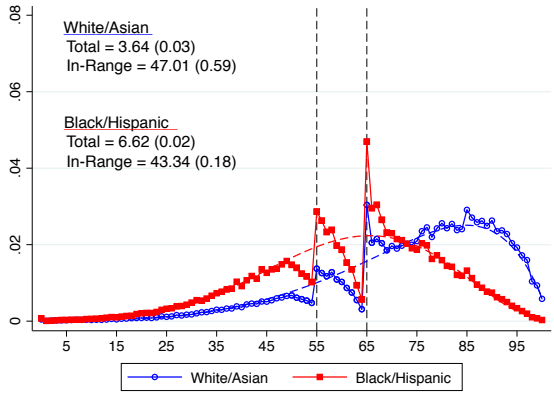
Note: These figures show the test score distribution for core Regents exams around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Panel (a) considers exams taken in schools above and below median in the fraction of black/Hispanic students. Panel (b) considers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

Appendix Figure A7: Results by Student Characteristics, 2004-2010

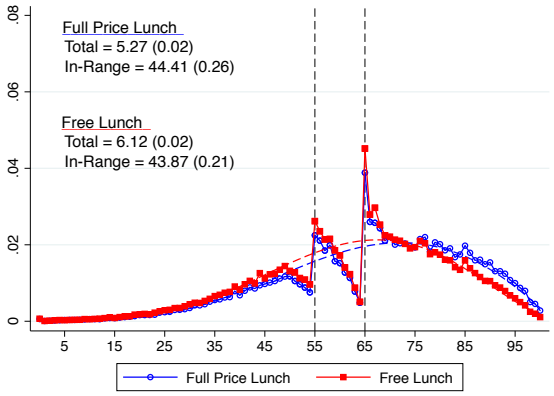
(a) Female vs. Male



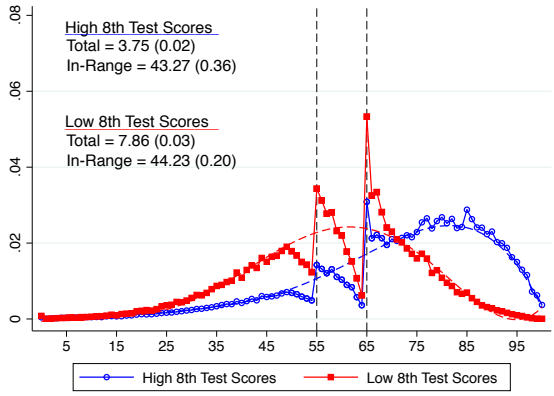
(b) White/Asian vs. Black/Hispanic



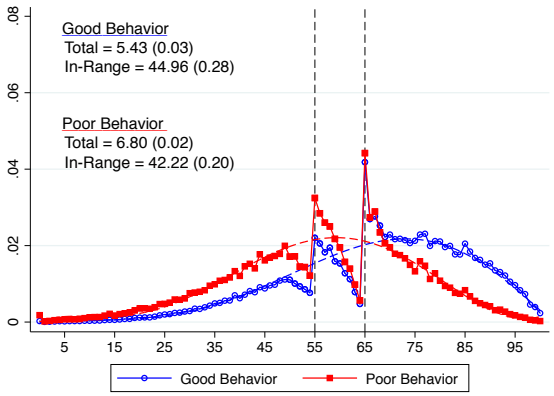
(c) Full Price vs. Free Lunch



(d) Low vs. High 8th Test Scores

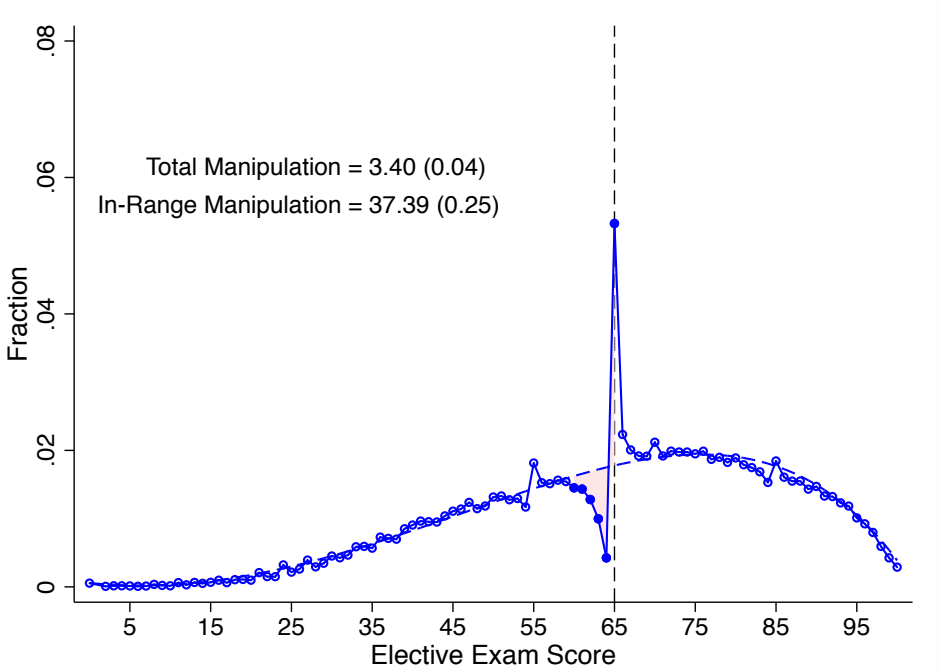


(e) Good vs. Poor Behavior



Note: 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 above and below median in 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.

Appendix Figure A8: Results for Elective Regents Exams, 2004-2010



Note: 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 each cutoff. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III. 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 Table A1: Regents Exam Requirements by Diploma Type and Cohort

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

Note: The five core Regents-Examination subjects are English, Mathematics, Science, U.S. History and Government, and 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 A2: 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	5.20		4.83			
	(0.02)		(0.08)			
In-Range Manipulation	52.46		25.25			
	(0.71)		(0.14)			
	20325		22781			
June 2004:						
Total Manipulation	6.35	7.05	5.69	5.89	6.08	
	(0.02)	(0.09)	(0.08)	(0.02)	(0.01)	
In-Range Manipulation	60.54	32.96	29.87	51.99	53.10	
	(0.78)	(0.09)	(0.09)	(0.59)	(0.54)	
	26699	41875	31158	38106	48934	
January 2005:						
Total Manipulation	5.42		4.51			
	(0.01)		(0.07)			
In-Range Manipulation	51.71		27.22			
	(0.65)		(0.10)			
	23838		24449			
June 2005:						
Total Manipulation	7.59	6.39	5.14	6.71	7.38	
	(0.01)	(0.08)	(0.08)	(0.02)	(0.01)	
In-Range Manipulation	72.42	32.59	26.95	58.93	64.29	
	(0.92)	(0.07)	(0.12)	(0.67)	(0.64)	
	24052	43572	31905	35387	47447	
January 2006:						
Total Manipulation	4.30		3.21			
	(0.02)		(0.08)			
In-Range Manipulation	42.93		16.85			
	(0.58)		(0.23)			
	27808		28171			
June 2006:						
Total Manipulation	5.94	6.94	5.30	6.77	7.71	
	(0.01)	(0.08)	(0.09)	(0.02)	(0.01)	
In-Range Manipulation	57.99	35.52	24.93	60.78	67.62	
	(0.76)	(0.04)	(0.16)	(0.69)	(0.68)	
	24483	41348	28267	36798	47147	
January 2007:						
Total Manipulation	5.76		4.16			
	(0.02)		(0.08)			
In-Range Manipulation	54.94		21.77			
	(0.71)		(0.17)			
	29929		27671			
June 2007:						
Total Manipulation	6.03	7.13	4.06	7.09	7.35	
	(0.02)	(0.08)	(0.09)	(0.02)	(0.01)	
In-Range Manipulation	57.57	36.40	19.14	63.55	64.96	
	(0.75)	(0.04)	(0.22)	(0.73)	(0.66)	
	22403	40932	27248	37687	44551	
January 2008:						
Total Manipulation	3.26		3.36			
	(0.02)		(0.07)			
In-Range Manipulation	32.58		20.47			
	(0.45)		(0.18)			
	27915		26352			

	Comp. English	Living Env.	Math A	U.S. History	Global History	Int. Algebra
June 2008:	(1)	(2)	(3)	(4)	(5)	(6)
Total Manipulation	4.02	5.62	4.27	5.60	6.33	2.81
	(0.02)	(0.06)	(0.09)	(0.02)	(0.02)	(0.10)
In-Range Manipulation	40.12	36.75	20.16	50.04	56.33	21.63
	(0.55)	(0.02)	(0.21)	(0.57)	(0.58)	(0.35)
	23618	42073	18044	38289	44951	34185
January 2009:						
Total Manipulation	3.91					3.95
	(0.01)					(0.10)
In-Range Manipulation	38.17					30.54
	(0.50)					(0.17)
	27547					10489
June 2009:						
Total Manipulation	4.15	6.63		7.52	6.52	4.42
	(0.01)	(0.08)		(0.02)	(0.02)	(0.10)
In-Range Manipulation	40.48	33.80		65.37	57.09	33.93
	(0.53)	(0.06)		(0.73)	(0.59)	(0.12)
	23697	41261		39470	43283	39513
January 2010:						
Total Manipulation	3.69					3.79
	(0.02)					(0.07)
In-Range Manipulation	35.22					38.83
	(0.46)					(0.12)
	27099					13956
June 2010:						
Total Manipulation	3.73	7.21		5.81	6.48	4.09
	(0.02)	(0.08)		(0.02)	(0.01)	(0.11)
In-Range Manipulation	35.56	36.90		51.56	56.86	31.19
	(0.46)	(0.03)		(0.59)	(0.57)	(0.17)
	22771	41477		37435	42707	34132

Note: This table reports manipulation around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. See the Figure 1 notes for details on the empirical specification and the data appendix for additional details on the sample and variable definitions.

Appendix Table A3: Summary Statistics for School x Subject In-Range Manipulation

	In-Range Manipulation			Within-School Correlation				
	Obs.	Mean	S.D.	U.S. History	Global History	English	Math	Living Env.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
U.S. History	259	54.2	33.9	1.00				
Global History	263	58.7	34.5	0.77	1.00			
English	258	46.9	23.1	0.29	0.26	1.00		
Math	271	26.1	15.6	0.22	0.22	0.21	1.00	
Living Environment	273	36.0	16.9	0.04	0.08	0.14	0.25	1.00

Note: This table presents summary statistics for our estimates of in-range manipulation at the school x subject level. Columns 1-3 present the number of estimates and means and standard deviations by subject area. Columns 4-8 present pairwise correlations weighted by the number of in-range exams for each subject area pair, where math includes both Math A and Integrated Algebra exams. See the data appendix for additional details on the sample construction and variable definitions and the text for additional details on the calculation of the school x subject in-range manipulation estimates.

Appendix Table A4: Student Subsample Results

	Total Manipulation		In-Range Manipulation	
	True Subgroup	Synthetic Subgroup	True Subgroup	Synthetic Subgroup
Panel A: Gender	(1)	(2)	(3)	(4)
Female	5.82	5.67	44.22	43.65
	(0.02)	(0.02)	(0.23)	(0.16)
Male	5.65	5.78	43.89	44.53
	(0.02)	(0.02)	(0.22)	(0.18)
Difference	0.17	−0.11	0.33	−0.87
	(0.03)	(0.04)	(0.34)	(0.33)
Panel B: Ethnicity				
White/Asian	3.64	4.24	47.01	45.21
	(0.03)	(0.02)	(0.59)	(0.28)
Black/Hispanic	6.62	6.37	43.34	43.63
	(0.02)	(0.01)	(0.18)	(0.08)
Difference	−2.98	−2.14	3.67	1.59
	(0.04)	(0.03)	(0.60)	(0.35)
Panel C: Free Lunch Eligibility				
Full Price Lunch	5.27	5.36	44.41	44.89
	(0.02)	(0.02)	(0.26)	(0.18)
Free Lunch	6.12	6.02	43.87	43.48
	(0.02)	(0.02)	(0.21)	(0.13)
Difference	−0.85	−0.65	0.54	1.40
	(0.04)	(0.04)	(0.33)	(0.30)
Panel D: 8th Test Scores				
Above Median 8th Scores	3.75	4.94	43.27	43.95
	(0.02)	(0.02)	(0.36)	(0.18)
Below Median 8th Scores	7.86	6.63	44.23	44.03
	(0.03)	(0.02)	(0.20)	(0.16)
Difference	−4.10	−1.69	−0.96	−0.08
	(0.04)	(0.04)	(0.44)	(0.34)
Panel E: Behavior and Attendance				
Good Attendance/Behavior	5.43	5.50	44.96	44.10
	(0.03)	(0.01)	(0.28)	(0.09)
Poor Attendance/Behavior	6.80	6.50	42.22	44.12
	(0.02)	(0.04)	(0.20)	(0.27)
Difference	−1.37	−0.99	2.74	−0.01
	(0.03)	(0.05)	(0.36)	(0.36)

Note: 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 fraction of students with each characteristic constant within each school when creating synthetic subgroups. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the text for additional details.

Appendix Table A5: 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	

Note: 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. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table A6: Difference-in-Differences Placebo Estimates

	Sample Mean	Reduced Form	
Panel A: Characteristics	(1)	(2)	(3)
Male	0.471	0.0059	0.0065
	(0.499)	(0.0066)	(0.0058)
White	0.143	0.0004	−0.0031
	(0.350)	(0.0056)	(0.0035)
Asian	0.180	0.0065	0.0003
	(0.384)	(0.0068)	(0.0046)
Black	0.315	−0.0081	0.0015
	(0.465)	(0.0085)	(0.0046)
Hispanic	0.355	0.0034	0.0035
	(0.479)	(0.0081)	(0.0051)
Free Lunch	0.604	0.0186*	0.0197*
	(0.489)	(0.0126)	(0.0114)
Above Median 8th Score	0.542	0.0017	−0.0005
	(0.498)	(0.0061)	(0.0056)
Panel B: Predicted Outcomes			
Predicted Graduation	0.795	0.0011	−0.0002
	(0.137)	(0.0018)	(0.0015)
Predicted Regents Requirements	0.885	0.0001	−0.0003
	(0.054)	(0.0007)	(0.0006)
Predicted Adv. Regents Requirements	0.368	0.0022	−0.0012
	(0.248)	(0.0035)	(0.0025)
Observations	1,002,804	1,002,804	1,002,804
Student Controls	–	No	No
Year x Score Trends	–	Yes	Yes
School Fixed Effects	–	No	Yes

Note: 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, exam by year-of-test effects, and cohort by year-of-test effects. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exams and cluster standard errors at the individual and school levels. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table A7: Difference-in-Differences Results by Subject

	All Core Exams	Living Env.	Int. Algebra	Global History	English	U.S. History
Panel A: High School Graduation	(1)	(2)	(3)	(4)	(5)	(6)
Graduate High School	0.227***	0.400***	0.215***	0.128***	0.361***	0.322***
	(0.025)	(0.077)	(0.085)	(0.028)	(0.080)	(0.048)
	[0.792]	[0.823]	[0.749]	[0.799]	[0.802]	[0.840]
Panel B: Diploma Requirements						
Regents Requirements Taken	0.078**	0.134	−0.061	0.060**	0.098*	0.030
	(0.039)	(0.126)	(0.191)	(0.029)	(0.051)	(0.027)
	[0.883]	[0.897]	[0.837]	[0.896]	[0.871]	[0.939]
Adv. Regents Requirements Taken	−0.049	−0.175	−0.793***	0.102*	−0.189*	−0.063
	(0.067)	(0.156)	(0.218)	(0.059)	(0.103)	(0.061)
	[0.342]	[0.398]	[0.307]	[0.370]	[0.346]	[0.381]
Panel C: Advanced Science and Math Exams						
Take Physical Science	0.015	0.051	−0.170	0.037	0.036	0.085*
	(0.033)	(0.095)	(0.150)	(0.039)	(0.081)	(0.051)
	[0.716]	[0.774]	[0.654]	[0.721]	[0.687]	[0.741]
Take Advanced Math	−0.002	−0.039	−0.557***	0.163***	−0.161	−0.070
	(0.060)	(0.120)	(0.194)	(0.056)	(0.099)	(0.060)
	[0.366]	[0.415]	[0.332]	[0.397]	[0.375]	[0.408]
Observations	2,046,649	301,881	367,517	333,406	357,374	295,584
Student Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Note: 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, cohort by year-of-test effects, and school effects. Standard errors are clustered at both the student and school level. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table A8: Difference-in-Differences Results for Additional Outcomes

	Sample Mean	First Stage		Reduced Form		2SLS	
Panel A: Future Exam Scores	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Score 70+ in Next Year	0.581	−0.141***	−0.141***	0.021***	0.020***	−0.149***	−0.142***
	(0.493)	(0.007)	(0.007)	(0.004)	(0.004)	(0.026)	(0.026)
Score 75+ in Next Year	0.443	−0.141***	−0.141***	0.009***	0.008***	−0.063***	−0.054***
	(0.497)	(0.007)	(0.007)	(0.003)	(0.003)	(0.020)	(0.020)
Score 80+ in Next Year	0.311	−0.141***	−0.141***	0.005***	0.005***	−0.034***	−0.034***
	(0.463)	(0.007)	(0.007)	(0.001)	(0.001)	(0.008)	(0.009)
Score 85+ in Next Year	0.198	−0.141***	−0.141***	−0.000	−0.001	0.003	0.004
	(0.398)	(0.007)	(0.007)	(0.002)	(0.002)	(0.016)	(0.016)
Panel B: Diploma Requirements							
Regents Requirements Passed	0.655	−0.141***	−0.141***	−0.143***	−0.145***	1.014***	1.029***
	(0.475)	(0.007)	(0.007)	(0.015)	(0.014)	(0.117)	(0.112)
Adv. Regents Req. Passed	0.219	−0.141***	−0.141***	0.030**	0.025**	−0.214**	−0.180**
	(0.414)	(0.007)	(0.007)	(0.013)	(0.012)	(0.091)	(0.086)
Panel C: Advanced Science and Math							
Take Physical Science Exam	0.713	−0.141***	−0.141***	0.007	0.006	−0.050	−0.043
	(0.453)	(0.007)	(0.007)	(0.009)	(0.007)	(0.063)	(0.051)
Take Advanced Math Seq.	0.379	−0.141***	−0.141***	0.011	0.008	−0.077	−0.055
	(0.485)	(0.007)	(0.007)	(0.012)	(0.012)	(0.088)	(0.083)
Panel D: Other Attainment Measures							
Years Enrolled in High School	4.113	−0.141***	−0.141***	−0.112***	−0.110***	0.793***	0.776***
	(0.552)	(0.007)	(0.007)	(0.006)	(0.006)	(0.057)	(0.054)
Highest Enrolled Grade	11.834	−0.141***	−0.141***	−0.056***	−0.057***	0.400***	0.405***
	(0.531)	(0.007)	(0.007)	(0.006)	(0.005)	(0.042)	(0.040)
Panel E: High School Graduation							
Graduate in 5 Years	0.845	−0.161***	−0.162***	−0.036***	−0.037***	0.222***	0.230***
	(0.362)	(0.009)	(0.008)	(0.005)	(0.005)	(0.035)	(0.034)
Graduate in 6 Years	0.872	−0.166***	−0.166***	−0.026***	−0.027***	0.156***	0.162***
	(0.334)	(0.009)	(0.008)	(0.005)	(0.005)	(0.030)	(0.029)

	Sample Mean	First Stage		Reduced Form		2SLS	
Panel F: GED Receipt	(1)	(2)	(3)	(4)	(5)	(6)	(7)
GED Diploma	0.005	−0.166***	−0.166***	0.001*	0.001*	−0.008	−0.008
	(0.073)	(0.009)	(0.008)	(0.001)	(0.001)	(0.005)	(3.016)
Observations	1,002,811	1,002,811	1,002,811	1,002,811	1,002,811	1,002,811	1,002,811
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

Note: 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. 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, exam by year-of-test effects, and cohort by year-of-test effects. Standard errors are clustered at both the student and school level. See the Table 5 notes for details.

Appendix Table A9: Robustness of Difference-in-Differences Results

Panel A: High School Graduation	(1)	(2)	(3)	(4)
Graduate High School	0.219***	0.178***	0.236***	0.224***
	(0.033)	(0.034)	(0.031)	(0.030)
Panel B: Diploma Requirements				
Regents Requirements Taken	0.040	−0.002	0.043	0.043
	(0.080)	(0.063)	(0.072)	(0.075)
Adv. Regents Requirements Taken	−0.139	−0.145*	−0.191**	−0.192**
	(0.099)	(0.074)	(0.088)	(0.088)
Observations	746,637	467,433	1,002,804	1,002,804
Student Controls	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes
Drop 0-59 Scores	Yes	Yes	No	No
Drop 81-100 Scores	No	Yes	No	No
IV Year-Specific Interaction	No	No	Yes	Yes
IV Pilot School Interaction	No	No	No	Yes

Note: 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 drops scores between 0-59. Column 2 drops scores between 0-59 and 81-100. Column 3 uses the interactions of scoring between 60-69 and year-specific indicators for taking the test between 2011-2013 as instruments. Column 4 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, cohort by year-of-test effects, and school effects. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exams and cluster standard errors at the individual and school levels. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table A10: Across-School Estimates

	Sample Mean	First Stage	Reduced Form	2SLS
Panel A: High School Graduation	(1)	(2)	(3)	(4)
Graduate High School	0.750	0.374***	−0.011	−0.028
	(0.433)	(0.017)	(0.009)	(0.024)
Panel B: Diploma Requirements				
Regents Requirements Met	0.550	0.374***	0.075***	0.200***
	(0.497)	(0.017)	(0.014)	(0.035)
Adv. Regents Requirements Met	0.178	0.374***	−0.024	−0.064
	(0.382)	(0.017)	(0.021)	(0.049)
Panel C: College Enrollment				
Any College	0.518	0.374***	−0.017*	−0.045*
	(0.500)	(0.017)	(0.009)	(0.023)
Any Two-Year College	0.194	0.374***	−0.021**	−0.056**
	(0.396)	(0.017)	(0.009)	(0.024)
Any Four-Year College	0.359	0.374***	0.006	0.016
	(0.480)	(0.017)	(0.008)	(0.022)
Observations	346,481	346,481	346,481	346,481
Student Controls	–	Yes	Yes	Yes
Year x Score Trends	–	Yes	Yes	Yes
School Fixed Effects	–	Yes	Yes	Yes

Note: 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, 10-point scale score effects interacted with year-of-test, exam by year-of-test effects, exam by cohort effects, and school effects. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exams and cluster standard errors at the individual and school levels. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table A11: Across-School Placebo Estimates

	Sample Mean	Reduced Form
Panel A: Characteristics	(1)	(2)
Male	0.465	0.0008
	(0.499)	(0.0095)
White	0.137	0.0061
	(0.344)	(0.0061)
Asian	0.147	0.0274***
	(0.354)	(0.0066)
Black	0.345	−0.0109
	(0.475)	(0.0090)
Hispanic	0.363	−0.0229***
	(0.481)	(0.0079)
Free Lunch	0.527	−0.0001
	(0.499)	(0.0088)
8th Grade Test Scores	-0.041	−0.0124
	(0.779)	(8.8749)
Panel B: Predicted Outcomes		
High School Graduation	0.755	−0.0009
	(0.154)	(0.0030)
Regents Requirements	0.563	−0.0006
	(0.264)	(0.0052)
Adv. Regents Requirements	0.183	0.0030
	(0.205)	(0.0039)
Any College	0.530	−0.0015
	(0.143)	(0.0027)
Any Two-Year College	0.196	−0.0008
	(0.040)	(0.0007)
Any Four-Year College	0.370	−0.0007
	(0.185)	(0.0034)
Observations	346,481	346,481
Student Controls	–	Yes
Year x Score Trends	–	Yes
School Fixed Effects	–	Yes

Note: 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. Column 2 reports 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, 10-point scale score effects interacted with year-of-test, exam by year-of-test effects, exam by cohort effects, and school effects. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exams and cluster standard errors at the individual and school levels. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix B: Additional Details for the New York Regents Examinations [NOT FOR PUBLICATION]

This appendix contains additional details on the history and structure of the New York Regents Examinations and their recent use in state and city school accountability policies.

A. Historical and Structural Details

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, and in 1878 the Regents began offering advanced academic exams in various subjects to be used in college admissions. A fuller accounting of the Regents first 100 years can be found in NYSED (1965). Among the most important changes in recent years was the introduction of 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 we examine in this paper (Chudowsky et al. 2002).

While requirements in the English, science, and social studies exams remained fairly constant during our sample period, there were important changes to Regents' math requirements. Until 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. An additional exam (Math B) was also available during this period for more advanced students. From 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 the first Algebra exam 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.

Scoring of regents exams followed very explicit policies. 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-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.

During our primary sample period (2003-2004 to 2009-2010), 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 these grading guidelines that we are aware of are the Chemistry exam 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.

B. Use in School Accountability

In order to meet requirements for Adequate Yearly Progress (AYP) under the 2002 No Child Left Behind Act, high schools in New York must meet several criteria related to Regents examination participations and performance. 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, where subgroups are based on race/ethnicity, poverty status, and program receipt. 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 Examination exceed 55, with students whose scores exceed 65 having twice the impact on this index. Specifically, the performance index equals 100*[(count of cohort with scale scores $\geq$ 55 + count of cohort with scale scores $\geq$ 65) / cohort size] (NYSED 2010). Thus, the performance index ranges from 0 (i.e., all students have scale scores below 55) to 200 (i.e., all students have scale scores of 65 or higher). 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 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.

New York City's separate accountability system awarded grades (A to F) to high schools starting in 2007. 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.

C. Grading Appeals

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 United States 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.

Appendix C: Data Appendix
[NOT FOR PUBLICATION]

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 2001-2002 and 2009-2010, the graduation data include information on the receipt a regular high school diploma (i.e. a local, Regents, or advanced Regents diploma) and the receipt of a GED. The data include information on four-, five-, and six-year graduation outcomes. Information on diploma type is only available for cohorts entering high school between 2007-2008 and 2009-2010.

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 2001-2002 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, U.S. 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 were 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 2,470,187 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 2,453,437 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 1,977,221 remaining exams.

4. Fourth, we drop students who are enrolled in non-high schools, special education schools, and schools with extremely low enrollments. This sample restriction leaves us with 1,820,899

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 1,728,043 remaining exams.

6. Finally, we drop special education students who are held to different accountability standards during our sample period (see Appendix Table A1). This sample restriction leaves us with 1,629,910 core exams from 514,632 students in our primary sample.

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.

Appendix D: Across-School Estimates
[NOT FOR PUBLICATION]

In this appendix, we provide additional details on the cross-sectional methodology we use to examine college enrollment outcomes.

Intuitively, our cross-sectional strategy asks whether the outcomes of students with scores in the manipulable range, relative to those enrolled in the same school but with scores outside the range, are systematically better in schools where manipulation is more prevalent. Throughout the analysis, we focus on the 65 cutoff to make the estimates more directly comparable to the difference-in-differences estimates presented above. For the same reason, we focus on graduating with a Regents or Advanced Regents diploma in the analysis, as the 65 cutoff is somewhat less relevant for the any high school graduation measure in the cohorts where we observe college enrollment outcomes.

Formally, we estimate the reduced form impact of attending a high manipulation school using the following specification:

$$y_{isemth} = \alpha_{14h} + \alpha_{14et} + \alpha_{14s} + \alpha_{14s} \cdot \text{Year}_{ie} + X_i\beta_{14} + \gamma_{14} \cdot \mathbf{1}[69 \geq \text{Score}_{iemt} \geq 60] \cdot Manipulation_{eh} + \varepsilon_{isemth} \quad (14)$$

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_{14h} are school effects, α_{14et} are exam by year effects, α_{14s} are 10-point Regent score effects, $\alpha_{14s} \cdot \text{Year}_{ie}$ 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 A1), and X_i includes controls for gender, ethnicity, free lunch eligibility, 8th grade test scores. $Manipulation_{eh}$ is an estimate of in-range manipulation for exam e and high school h as described as Section IV.B. We estimate $Manipulation_{eh}$ only using exams around the 65 cutoff in the cohorts. Results are similar but less precise if we use a measure of manipulation $Manipulation_{eh}$ that uses information across both cutoffs. When estimating Equation (14), we stack student outcomes across all core Regents exams and adjust our standard errors for clustering at both the student and school level. The estimating regressions for the first and second stage specifications follow naturally from the above.

The parameter γ_{14} 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, conditional on observables, the within-school differences in outcomes between students scoring between 60-69 ("in-range") and those scoring either below 60 or above 70 ("outside the range") is uncorrelated with any unobserved factor other than test score manipulation. Diamond and Persson (2016) estimate the extent of manipulation at the county level, not school level, but the identification assumptions are essentially the same in their analysis.

This identifying 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 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 A11).