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

By Thomas S. Dee, Will Dobbie, Brian A. Jacob, and Jonah Rockoff*

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 academic outcomes. While inflating a score increases the probability of a student graduating from high school by about 17 percentage points, the probability of taking advanced coursework declines by roughly 10 percentage points. 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.

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]

* Dee: Stanford University and NBER, Graduate School of Education, Stanford University, 520 Galvez Mall, CERAS Building, 5th Floor, Stanford, CA 94305, tdee@stanford.edu. Dobbie: Princeton University and NBER, Industrial Relations Section, Louis A. Simpson International Building, Princeton University, Princeton, NJ, 08544, wdobbie@princeton.edu. Jacob: University of Michigan and NBER, Gerald R. Ford School of Public Policy, University of Michigan, 735 South State Street, Ann Arbor, MI, 48109, bajacob@umich.edu. Rockoff: Columbia University and NBER, Columbia Business School, Columbia University, 3022 Broadway, Uris Hall 603, New York, NY, 10027, jonah.rockoff@columbia.edu. 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, James Reeves, Sean Tom, and Yining Zhu provided excellent research assistance.

[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

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 did not affect test score manipulation, and Angrist, Battistin, and Vuri (2017) 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, Höglin, 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 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 stu-

accountability (Figlio and Getzler 2006, Cullen and Reback 2006, Jacob 2005).

dents, 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 24 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 graduation 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 16.7 percentage points, a 21.1 percent increase from the pre-reform mean. The effects on high school graduation are economically and statistically significant for all student subgroups. These results suggest that test score manipulation had important 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 advanced high school diploma by 9.8 percentage points, a 26.6 percent decrease from the pre-reform mean, with larger effects for students with lower 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

[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.

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 effects of manipulation on educational attainment we find for some students. 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 negative effects of test manipulation on students' later academic outcomes. Specifically, students who attended middle schools where cheating occurred are more likely to drop out of high school. Combined with the results of our study, these recent papers support the idea that test manipulation can have either positive or negative effects on different students and in different contexts.[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 relevant

[3] In related work on the long-term impacts of high stakes testing, Ebenstein, Lavy, and Roth (2016) 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 (2017) show that the significant, built-in flexibility of the German tracking system allows for initial tracking mistakes to be corrected over time.

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–although not necessarily each student's actual teacher–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]

[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.

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 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 the 2002-2003 academic year, 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

[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).

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.IV.D.

II. Data

Here we summarize the most relevant information regarding our administrative enrollment and test score data from the 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, nor do they include an identifier for the teacher(s) who graded the exams. 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.

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.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 IV.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 these 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.

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.[8]

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

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

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 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:

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

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.[9]

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

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

[9]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.

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" 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 and time t.[10]

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. The widespread knowledge of these norms for math and science may well have influenced grading norms on in other subject areas. We therefore define the lower bound of the manipulable range on all Regents exams to be a score of 50 for the cutoff at 55 and a score of 60 for the cutoff at 65. The only exceptions are a few cases in which the exact score of 50 or 60 is not a possible scale score, and 51 or 61 is used instead as the lower bound for the manipulable range around that cutoff.

For the upper bound of the manipulable range we follow the framework laid out above and assume that teachers manipulate in order to push a student's score above a passing threshold. We also assume that teachers would, all else equal, prefer to manipulate a score through changing their subjective ratings of essays and open-response items, as opposed to changing or filling in multiple choice answers. These assumptions lead us to define the upper bound of the manipulable range as the highest score a student could receive if the student was initially within the 50-54 (or 60-64) range and a teacher awarded one additional raw point on an essay or open-response item. In other words, the top of the manipulable range is the best score a student could get if they initially were failing but a teacher awarded them the minimum credit needed to pass.

These assumptions, coupled with the scoring rules for different subjects, lead to specific definitions of the upper bound of the manipulable range for each exam. On math and science exams, the upper bound is typically the exact 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. The only exceptions on math and science

[10] Given the empirical distribution of exam scores, it is obvious that a fairly high-order polynomial is needed, but it is unclear whether a sixth-order polynomial is sufficiently flexible. Appendix Table A2 shows the Akaike Information Criterion (AIC) for fitting the data with polynomials of order 1 through 7; we see the criterion fall monotonically until the sixth-order polynomial (indicating a better fit) and then increase when we add a seventh-order polynomial. Our results are not sensitive to small changes in the polynomial order.

exams are a few cases in which the exact cutoff of 55 is not a possible scale score, and 56 is used instead as the upper bound for the manipulable range around that cutoff. For exams in English and social studies, the scoring rules generally result in an upper bound beyond the exact cutoff c. This is because manipulating a score to be exactly 55 or 65 can be challenging if not impossible for any given student. Changes in essay ratings of just one raw point typically change the scale score by four points. For example, a student that initially scores a 63 would be moved to a 67 if a grader awards an additional point on one of the essay prompts.[11] This means that the upper bound for English and social studies is usually 67, 68, or (in a few cases) 69.

Defining the manipulable range in this manner is highly consistent with the patterns we observe in the data (see, for example, Appendix Figure A2). We also provide more formal specification checks and tests for the robustness of our results to changes in the manipulable score region. For purposes of transparency, Appendix Table A3 shows exactly which scores are included in the manipulable range above the proficiency cutoffs for each of the June exams in each of the core subjects.

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.[12] 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:

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

[11] The conversion chart for the June 2009 English Exam shown in Appendix Figure A1 illustrates this point. A student that has 17 correct multiple choice items and originally receives 15 raw points on the essay questions would end up with a scale score of 63. If a teacher awards that student one additional raw point on one of the essay responses (for a total of 16 raw points), the scale score will jump to 67. The same situation arises for a student who had 6 multiple choice items correct and an initial essay total of 18 raw points, which translates to a scale score 53. Moving that student's essay total to 19 points will shift his or her final scale score to 57.

[12] We regress the estimated excess mass to the right of the cutoff on the estimated missing mass to the left, weighting by the number of tests used to generate the estimates, and find a coefficient of -0.99 with an R-squared of 0.53. Adding subject fixed effects raises the R-squared to 0.63 and the coefficient remains stable at -1.02. This supports our view that both of these estimates of manipulation are measured with error but capture the same underlying behavior, and that taking the average is likely to yield more precise estimates.

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 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.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

[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,

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 43.9 (se=0.13) percent of Regents core exams were manipulated. We also look 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 A4). 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 explore the robustness of our method for selecting the manipulable range, we also estimate manipulation allowing scale scores immediately above and below our chosen manipulable regions to be potentially manipulable. In other words, we add additional indicators $\mathbf{1}[s = j]$ to the regression shown in Equation (9) for scores j just outside of our manipulable range $[M_c^-, c]$. This effectively removes those scores from contributing to the estimated counterfactual density. Appendix Figure A5 shows that the amount of excess (or missing) mass estimated to occur at

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.

[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.

these points is small and statistically insignificant; moreover, the point estimates at scores just below and just above our chosen range are of the incorrect sign. This provides us with confidence that our selection of upper and lower bounds for the manipulable range accurately capture the scope of manipulation by exam graders.

To provide further evidence that Regents scores near cutoffs were being manipulated, Appendix Figure A6 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. As mentioned above, 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, 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

[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.

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 a similar manipulation rate (i.e., 3 to 5 percent) occurred across the state of New York.

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 24.4 to 55.6 percent for in-range manipulation.[19] Thus, the probability of a marginally failing exam being manipulated clearly depended on which school the student attended.

Regressions of school-level manipulation on school characteristics are shown in Table 2, where we examine the correlation of manipulation with serving disadvantaged student populations as well as with school size, since teachers in larger

[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 above.

schools may be less likely to know students personally. We find 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 (Panel A, Columns 1-3).[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 and not statistically different from zero (Panel B, Columns 1-3). We also find that total manipulation is negatively associated with school size (Panel A, Column 4), but the coefficient becomes positive when we control for student characteristics (Panel A, Column 5) and when we examine in-range manipulation (Panel B, Column 4). 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 A8). 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

[20] Regressions are weighted by the number of in-range exams, but weighting by total exams provides quite similar results. For reference, the (weighted) standard deviations of the independent variables are 24.5 percent (for percent Black/Hispanic), 16.3 percent (for percent free lunch), 28.8 (for test score percentile), and 1,237 students (for enrollment).

[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 A7). 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 43.3 percent, while those with low Black/Hispanic enrollment have in-range manipulation of 45.0 percent. Additionally, while smaller schools' total manipulation is slightly higher than large schools', rates of in-range manipulation are 5.0 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

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 good behavior randomly within schools (Table 3), 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 A5 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 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 (0.78) 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

reassignment of student characteristics will lead to differences among students both within and outside the manipulable range, thus altering our estimated counterfactual distributions.

[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.

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 for subject area, and find a coefficient on lagged manipulation of 0.50 (se=0.09) (Table 4, Column 1). Adding school fixed effects (Column 2) decreases this measure of persistence slightly, to 0.42 (se=0.09), 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 0.84 (se=0.25) 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 0.08 (se=0.08) 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.[26] 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.

[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.

[26] Greater teacher turnover over this period could also be associated with decreases in teacher experience, which may in turn be linked to changes in manipulation. Indeed, when we calculate the change in average teacher experience within each school-subject cell, we find a positive and significant correlation of about 0.2 with the fraction of teachers present in both periods. However, including the change in experience and its interaction with lagged manipulation does not change the results shown in Table 4, and the coefficients on the variables related to experience are not statistically significant.

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 2006, 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 60.0 (se=0.68) percent of in-range exams manipulated before the implementation of these accountability systems, compared to the 42.4 (se=0.27) percent in the years after the implementation of these systems.[27]

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 44.1 (se=0.13) percent of in-range exams, compared to 43.8 (se=0.35) percent for schools meeting AYP. Similarly, schools receiving a D or F from the NYC accountability system manipulate 43.4 (se=0.35) percent of in-range exams, compared to 42.1 (se=0.28) 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).[28] We find that control schools

[27]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.

[28]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

manipulated 44.2 (se=0.27) percent of in-range exams taken during the experiment, which is higher than our estimate of 41.3 (se=0.23) 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 A9 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=0.03) percent of these elective Regents exams were manipulated in total, and that 37.3 (se=0.19) percent were manipulated among those with scores within the range just below the cutoff. The latter is only a few percentage points less than the amount of in-range manipulation for core Regents exams.

In sum, these estimates suggest that manipulation was unrelated to the 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.

graduation rates. Fryer (2013) finds no effect of the teacher incentive program on teacher absences or student attendance, behavior, or achievement.

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."[29]

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 A6) and our conversations with NYCDOE officials suggests there was no systematic selection of pilot schools and no major differences in their observable characteristics.[30]

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

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

[30]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 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.

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=0.99) and 62.4 (se=0.52) percent of in-range exams, respectively.[31] 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=0.26) 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 0.44 (se=0.31) percent in pilot schools, but remained fairly steady in non-pilot schools at 12.3 (se=0.27) percent compared to 13.0 (se=0.26) 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.

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

[31] As can be seen in Figure 4, in-range manipulation in 2010 across both the 55 and 65 cutoffs was above 60 percent, although manipulation had greatly decreased at the 55 cutoff (which was no longer relevant for almost all students taking exams in 2010) and manipulation at 65 was substantial. Appendix Figure A10 shows manipulation for pilot and non-pilot schools for each of the pre-reform years 2004 to 2009. Manipulation is fairly stable over this time period, with the decreasing (increasing) importance of the 55 (65) cutoff becoming apparent starting in 2008.

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 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 in 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. Recall that by 2008, New York State had finished phasing in new graduation rules requiring scores of 65 on all core exams (see Appendix Table A1). The exams in our analysis are typically taken in 9th and 10th grade, so when the policy changed in 2011 the only cutoff that mattered for the vast majority of students would have been 65.

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

$$y_{iet} = X_i\theta_{11} + W_{iet}\alpha_{11} + \gamma_{11} \cdot \mathbf{1}[M^-_{ce} \leq s_{iet} \leq c_e] \cdot Reform_t + \varepsilon_{iet} \tag{11}$$

where y_{iet} is the outcome of interest for student i who took exam e at time t. 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. X_i includes student gender, ethnicity, free lunch eligibility, and 8th grade test scores, although dropping these controls has little bearing on our estimates.[32] W_{iet} represents a set of fixed effects (high school and year by subject) and additional controls to help ensure comparability of students over time. Specifically, we place each exam into one of ten score "bins": $[0,9]$, $[10,19]$ ..., $[50,59]$, $[M^-_{ce}, c_e]$, $[c_e+1, 79]$, $[80-89]$, and $[90-100]$. We interact these score bin indicators with subject fixed effects and subject-specific linear trends. In this way, we remove differences in both levels and trends of outcomes for students at different parts of the score distribution, including students with scores in and around our manipulable range $[M^-_{ce}, c_e]$. $\mathbf{1}[M^-_{ce} \leq s_{iet} \leq c_e]$ is an indicator for a

[32]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 0.158. Controlling for both observable characteristics and school fixed effects only slightly increases our estimate to 0.167 (Column 5 of Table 6). 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.

score in the manipulable range, and $Reform_t$ is an indicator for exam e being taken following the grading reforms implemented in 2011.

A key assumption is that students with scores in the manipulable (or indeed any) score range are comparable over time, conditional on our controls. For math and science, test scaling is invariant over time and $[M_{ce}^{-}, c_e]$ always extends from 60 to 65. For Global History, differences in test scaling would cause the range $[M_{ce}^{-}, c_e]$ to vary between 60-67 and 60-68 depending on the year of test administration (see Appendix Table A3). To align with the comparability assumption stated above, we use a consistent range of 60-68 to define $[M_{ce}^{-}, c_e]$ in Global History in our difference-in-differences analysis.[33] However, allowing the top of this range to vary between 67 and 68 has very little effect on our estimates. We also address the possibility that students scoring 0-59 could be affected by the reform, as many will retake exams under the reform conditions, by including an interaction of $Reform_t$ and an indicator for scores 0-59. In Section V.V.D, we also show that results are similar if we drop all students scoring at or below 59.

The parameter γ_{11} can be interpreted as the differential effect of the reform on students scoring in the range $[M_{ce}^{-}, c_e]$ on an exam compared to students scoring in the range $[c_e + 1, 100]$ 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 in the manipulable range at the time of the reforms would have been identical to changes for students scoring above the manipulable range 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 65 (which is always in the manipulable range) had their scores manipulated to 70 (which is always above the manipulable range) prior to but not after the reform. However, the evidence discussed above and presented in Appendix Figure A5 shows no evidence of test score manipulation outside our defined manipulable range. In Section V.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:

$$Pass_{iet} = X_i\theta_{12} + W_{iet}\alpha_{12} + \gamma_{12} \cdot \mathbf{1}[M_{ce}^{-} \leq s_{iet} \leq c_e] \cdot Reform_t + \varepsilon_{iet} \quad (12)$$

where γ_{12} 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

[33] Following the same logic, for English and U.S. History, which are not used in our main results but are included in some appendices, we use the range 60-69.

takes the form:

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

Data limitations prevent us from measuring impacts on later outcomes such as college graduation or labor market earnings.[34] 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 2010) 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 2018). 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

Before turning to our difference-in-differences estimates, we begin with a descriptive examination of student test outcomes over the period between 2004 and 2013 for students with scores in the range $[M_{ce}^{-}, c_e]$ (i.e., students likely to be affected by test score manipulation). We focus on the Science, Math, and Global History core exams which are typically taken first in 9th or 10th grade.[35] Using this sample, we run a regression of student outcomes on interactions of exam subject and year indicators. This allows us to examine differential pre-trends in the outcomes of students with exam scores that may have been subject to manipulation, and to assess visually whether there is a post-reform change in outcomes consistent with the change in manipulation documented above. 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.

[34] We lack college attendance information on recent cohorts affected by the grading reforms that drive our identification strategy. A previous draft of this study (Dee et al. 2016) includes an examination of college attendance for earlier cohorts using cross-sectional regressions that rely on much stronger assumptions and produced somewhat imprecise results.

[35] As will be shown below, the policy reforms made it more difficult to pass Regents exams, and this could in turn change the composition of students who "make it" to the English and U.S. History exams taken closer to graduation. 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 A7), but we find a shift towards higher 8th grade test scores for students taking the English and U.S. History exams. These results are available on request. For completeness, we also present difference-in-difference estimates based on all core exams and based on each core exam separately in Appendix Table A8.

Coefficient estimates from these regressions for student test outcomes are presented in Figure 5. In Panel A we see results for whether a student passed with a score of 65 or above the first time they took the exam, the first stage dependent variable in Equation (12). The figure shows a smooth upward trend in the probability of passing during the pre-reform period and then a sudden drop of more than 25 percentage points following the implementation of the grading reforms in 2011. The pre-trend is consistent with the greater emphasis on the 65 point cutoff during the pre-reform period and supports our inclusion of controls for differential linear trends for students with marginal scores in our difference-in-differences specification. The post-2011 drop strongly supports the idea that the Regents grading reforms significantly decreased test score manipulation. Panel A also shows results for whether a student was able to pass with a score of 65 or more within one year and within two years of taking the exam for the first time. These plots provide a visual assessment of whether students who did not pass due to the reforms were eventually able to do so. The post-2011 drop in passing rates within one and two years are around 20 percentage points and 10 percentage points, respectively, suggesting that a majority of the students who failed because the reforms caused teachers not to manipulate their scores were able to retake and pass the exam within two years.

Panel B of Figure 5 explores the issue of re-taking further by plotting rates of test retaking within one and two years of a student's first exam attempt.[36] There is a clear spike upward in re-taking behavior after the reforms and re-taking within two years rises by even more, with magnitudes suggesting that the majority of marginal students who failed due to the lack of manipulation in the post-reform period tried to re-take the exam after further study. Panels C and D of Figure 5 show results for the probability of scoring well in excess of the passing cutoff (70+ and 75+) within two years of the first-time taking the exam. This plot shows visual evidence of the extent to which students who (due to the reforms) failed and re-took exams eventually demonstrated a significantly higher level of subject matter knowledge. These plots show increases in the fraction of high eventual scores among students who had initially scored in the manipulable range around the cutoff, with larger increases at lower thresholds like 70+ but still a noticeable increase even at scores of 75+.

Altogether, the descriptive evidence in Figure 5 suggests that the policy reforms had three short-run effects: (1) a significant fraction of students with scores near the 65 cutoff failed because their scores were not manipulated; (2) the majority, but not all, of these marginal failing students re-took the exams and eventually passed; (3) the marginal failing students re-taking exams, on average, increased their knowledge of the tested subject matter and some achieved reasonably high test scores. We provide more evidence of these effects in difference-in-differences regressions; these include all test takers and additional controls discussed above

[36] We do not look beyond two years as two of the three exams in our analysis sample are typically taken in 10th grade.

for Equation (11), including linear trends for each score bin. The "reduced form" regression results are quite consistent with our graphical analysis (Table 5). Initial pass rates are estimated to have fallen by 27.2 (se=1.0) percentage points, while pass rates within one and two years fell by 20.2 (se=0.9) and 11.0 (se=0.8) percentage points, respectively. Retaking within one year and two years rose by an estimated 15.4 (se=1.1) and 18.7 (se=1.1) percentage points. Finally, the probability of achieving higher scores (70+, 75+, 80+, and 85+) are all estimated to have increased after the reforms, with estimated effects of 9.6 (se=0.5) percent for 70+, 3.8 (se=0.4) percent for 75+, and between 1.3 (se=0.2) and 0.6 (se=0.2) percentage points for even higher scores. We also provide two-stage least squares estimates which scale up the reduced form coefficients by the first-stage coefficient of -0.27 (se=0.01) and can be interpreted as the impact of test score manipulation on students whose scores were manipulated. For example, we estimate that manipulation reduced two-year re-taking rates by 70.3 (se=2.6) percentage points, increased passing rates within two years by 41.5 (se=2.3) percentage points, and *decreased* the probability of scoring 75 or higher within two years by 14.4 (se=1.5) percentage points.

We now turn to examine the impact of the reforms on high school graduation and coursework completion. We focus on students entering high school between 2003-2004 and 2010-2011 for whom we can observe these outcomes.[37] Again, before turning to the full difference-in-difference regressions, we present plots that help assess our identifying assumption of parallel trends and visually illustrate changes in outcomes coincident with the grading reforms. Figure 6 shows year-specific coefficients for having a test score in the manipulable range in the range $[M_{ce}^{-}, c_e]$, taken from regressions that include test takers with scores in the range $[c_e + 1, 100]$, score bin fixed effects, and year effects. Our omitted year is 2010, the year prior to the reform, so these coefficients should thus be interpreted as the year-specific differences in outcomes between students with marginal scores and those safely above the manipulable range in each year, relative to 2010. In Panel A, we see that graduation rate differences between these groups of test-takers were fairly stable between 2004 and 2010 but then fall by almost 4 percentage points between 2010 and 2011 and remain significantly lower thereafter. In Panel B, we see no trends and no break after the reform period for students completing the course requirements for a Regents diploma, suggesting no overall effect on this margin of course-taking for the marginal failing students. In Panel C, however, there is evidence of a slight upward trend prior to 2011 for students completing the Advanced Regents diploma course requirements, with a clear rise in Advanced Regents course taking in the post-reform years.

Table 6 shows results from our difference-in-difference regressions, which include all test-takers, score-bin fixed effects and linear trends, and controls for student

[37]For completeness, Appendix Figure A11 displays graphical evidence of the impacts on test-taking outcomes for the same sample that we use to estimate effects on graduation and advanced course taking.

characteristics.[38] The "reduced form" coefficient indicates that students scoring in the manipulable range are 4.6 (se=0.6) percentage points less likely to graduate high school following the grading reforms. Two-stage least squares estimates 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 16.7 (se=2.1) percentage points. In other words, we estimate that one out of every six "marginal" Regents passers would not have graduated from high school if their scores had not been manipulated. In line with Figure 6, having an exam manipulated does not have a statistically significant effect on the probability of taking the requirements for a Regents diploma, the lowest diploma for most students during this time period, but *reduces* the probability of taking the requirements for the Advanced Regents diploma by 9.8 (se=5.1) percentage points.

Together these results provide clear evidence that the grading reforms' elimination of manipulation led to many additional students failing Regents exams. Broadly speaking, only a minority of these "marginal failures" saw changes in ultimate outcomes such as graduation and course completion. Many of these students – roughly 50 percent according to our estimates – re-took and passed the exams within two years and went on to graduate from high school with a normal Regents diploma. Many others – about 23 percent by our estimation – were never able to pass the exam but would not have graduated from high school regardless (e.g., they would have failed another required exam or some other graduation requirement). For those remaining students whose outcomes were altered by the reforms, we find evidence of two very different causal effects. We estimate that about 17 percent of these students were never able to pass and did not graduate from high school, but would have if their scores had been manipulated.[39] Thus the reforms, by eliminating manipulation, had a clear negative effect on academic attainment for this subset of test-takers. However, our estimates also suggest that about 10 percent of students re-took and passed the exams at a later date, some likely scoring considerably higher than before due to additional study, and went on to take the coursework required for an Advanced Regents diploma.[40] For this subset of students, the elimination of manipulation seems to have had a positive

[38] We also present results with and without school fixed effects, but these controls have very little impact on our estimates.

[39] The magnitude of our estimate is broadly consistent with related results from other states. Clark and Martorell (2014) find an increase of 40 to 50 percentage points in Texas but these students are taking their "last chance" exam. In contrast, Papay et al. (2010) find an increase of about 8 percentage points in Massachusetts for students who can re-take the exam up to four times.

[40] It would be ideal if we could link directly those students who went on to re-take and get higher scores to advanced coursework but this is of course not possible given our estimation method. However, we can ask the question of whether the additional higher scores we estimate were caused by re-taking among marginal failures might reasonably explain higher rates of Advanced course-taking. To do so, we take the estimated percentage of marginal failures scoring [70, 75), [75,80), [80,85), and [85-90) – i.e., we assume those above 85 are lower than 90 – and multiply by the average Advanced regents course-taking for all students with scores each of these bins. This calculation would imply an increase of 14.1 percentage points in Advanced Regents coursework, well within the confidence range for our difference-in-difference estimate of 9.8 percentage points.

effect on academic attainment, although it did require additional effort to pass the required exams.

We attempt to shed further light on the mechanisms underlying our results by examining additional outcomes and focusing on individual exams (see Appendix Table A8). First, the impacts on high school graduation were not driven by a particular exam among the three used in our main results (Living Environment, Math A/Integrated Algebra, and Global History). Second, the estimated effects on Advanced Regents course taking appear to be driven by enrollment in advanced math (rather than science) courses, and by students "marginally failing" the Math A/Integrated Algebra exam due to the grading reforms (Appendix Table A8). We also note that manipulation on the Global History exam is estimated to have a positive impact on taking advanced math; this could be driven by failing students having to re-take Global History when they otherwise would enroll in advanced math.

Overall, our results support the idea that test score manipulation has somewhat nuanced effects. Some students are "helped" by test score manipulation because they are not forced to retake and pass an exam or a course required to leave high school. Others are "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. It is difficult to ascertain which students are likely to be affected by manipulation in these ways, but we run several tests to see if the effects of manipulation were heterogeneous across different types of students. We explore this in Table 7, which reports two-stage least squares estimates for mutually exclusive student subgroups.[41] Estimated effects on high school graduation are not statistically different by gender or race; they are somewhat larger for students with higher baseline test scores, but these are marginally significant and we have not adjusted p-values for multiple comparisons. Unfortunately, data on absences and disciplinary incidents is not available for the most recent cohorts, so we cannot test heterogeneity along this dimension.

D. Robustness Checks

Alternative Attainment Measures: Most Regents exams are taken well before the end of high school, and failing an exam may affect whether students progress towards graduation or drop out. This is one of the ways in which our setting is different than the "last chance" exams examined by Clark and Martorell (2014). 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.[42] We

[41] For completeness, Appendix Figure A12 shows our year-specific coefficients for having a test score in the manipulable range $\left[M_{ce}^{-}, c_e\right]$, taken from regressions that include test takers with scores in the range $[c_e + 1, 100]$, score bin fixed effects, and year effects by subgroup.

[42] Since we only observe the most recent cohort for four years since high school entry, we calculate highest grade attained within four years of entry and cap the number of years enrolled in high school at

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 C of Appendix Table A9) show large effects of manipulation on both of these attainment measures: 0.29 (se=0.03) grade levels and 0.39 (se=0.03) 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. These positive effects are notable given that manipulation allows students to avoid re-taking a course and might have been expected to shorten the number of years spent in school for some students.

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 D of Appendix Table A9 shows that having a score manipulated also increases the probability of graduating from high school in five years by 20.3 (se=2.2) percentage points, and the probability of graduating from high school in six years by 15.6 (se=1.9) percentage points. 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 A10 presents estimates using a variety of specifications and instruments to assess the robustness of our main two-stage least squares results further. Column 1 replicates our main results. Column 2 limits the control group to students scoring from just above the manipulable range to 80. Column 3 limits the control group to students scoring between 81-100. Column 4 uses the interactions of scoring in the manipulable range and year-specific indicators for taking the test between 2011-2013 for a total of three instrumental variables. Column 5 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 6.

Alternative Manipulable Range: Appendix Table A11 presents estimates using an expanded manipulable range. Our analysis generally assumes that manipulation does not extend beyond our specified ranges, e.g. a teacher manipulating an Algebra exam would only award points up to the 65 cutoff, not to 66 or 67. We check that our specific choice of range is not driving the results by estimating specifications that expand the range upward by 1, 2, or 3 points, e.g., we presume

four.

Algebra manipulation extends to 66, 67, or 68 points. Column 1 replicates our main results, while Columns 2 to 4 show the estimates with expanded manipulable ranges. The first stage coefficients on passing the exam are attenuated when we add observations that, by construction, are above 65. The two-stage least squares coefficient estimates grow somewhat larger, implying that the reduced form effects shrink to a lesser extent than the first stage, but essentially confirm our main findings.

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 A7. We find a statistically significant coefficient for only one out of seven student characteristics (a 2.45 (se=0.76) 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), where predictions are based on pre-reform cross-sectional regressions using all of the baseline characteristics listed in Panel A of Appendix Table A7. Consistent with our identifying assumption, we find no statistically significant differences following the elimination of re-scoring.

E. Aggregate Implications

Our estimates from this section suggest that test score manipulation had economically important effects on student outcomes. In light of the differential benefits of manipulation documented in Section IV.IV.C, our 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 7 by the subgroup-specific total manipulation estimates from Appendix Figure A8. 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 75.6 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.[43]

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 16.1 percentage points in the absence of test score manipulation,

[43]The high school graduation rate is higher in our sample compared to the district as a whole (65.2 percent) because we drop students in special education, students in non-traditional high schools, and students without at least one core Regents score.

while the graduation gap between high- and low-achieving students would have increased from 25.0 percentage points to 25.1 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 widespread and that it had significant effects on the overall performance of students across and within New York public schools.

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 17 percentage points, or more than 21 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 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.

A clear advantage of studying manipulation on a standardized exam with clear rules is that we can be precise in our measurement of the magnitude of manipulation and how this magnitude varies across students and settings. While New York was unusual in allowing high-stakes exams to be locally graded by teachers, it is important to keep in mind that the primary measures of student achievement in most educational settings are course grades, which are almost exclusively locally graded and are often a requirement for advanced high school course eligibility, high school graduation, and college admissions. Our results suggest that teachers are also likely to manipulate to some degree on course exams and grades in order to help students avoid failure. Our evidence on the patterns of manipulation (e.g., more likely for marginal students who are white and Asian than for those who

are Black or Hispanic) may be relevant in other dimensions of teacher behavior, such as discipline, classroom engagement, and grading.

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

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): 151-184.

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." CALDER Working Paper 155.

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

Angrist, Joshua, Erich Battistin, and Daniela Vuri. 2017. "In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno." *American Economic Journal: Applied Economics*, 9(4): 216-249.

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.

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.

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

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

Canaan, Serena, and Pierre Mouganie. 2018. "Returns to Education Quality for Low-Skilled Students: Evidence from a Discontinuity." *Journal of Labor Economics*, 36(2) 395-436.

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." *Quarterly Journal of Economics*, 126(2): 749-804.

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.

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

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

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

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

Cullen, Julie, and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." In *Advances in Applied Microeconomics*, vol 14., edited by Timothy J. Gronberg & Dennis W. Jansen, 1-34.

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." Columbia Business School Research Paper.

Dee, Thomas, Will Dobbie, Brian Jacob, and Jonah Rockoff. 2016. "The Causes and Consequences of Test Score Manipulation: Evidence from the New York Regents Examinations." NBER Working Paper No. 22165.

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

Dobbie, Will, and Roland Fryer. 2014. "The Impact of Attending a School with High-Achieving Peers: Evidence from the New York City Exam Schools." *American Economic Journal: Applied Economics*, 6(3): 58-75.

Dustmann, Christian, Patrick Puhani, and Uta Schönberg. 2017. "The Long-Term Effects of Early Track Choice." *Economic Journal*, 127(603): 1348-1380.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth. 2016. "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution." *American Economic Journal: Applied Economics*, 8(4): 35-65.

Figlio, David, and Lawrence Getzler. 2006. "Accountability, Ability and Disability: Gaming the System." In *Advances in Applied Microeconomics*, vol 14., edited by Timothy J. Gronberg & Dennis W. Jansen, 35-49.

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

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

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

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

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.

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

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.

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.

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

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.

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.

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.

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

New York State Education Department. 2008. History of Elementary, Middle, Secondary & Continuing Education. http://www.regents.nysed.gov/about/history-emsc.html. (accessed January 29, 2011).

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

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.

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.

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.

Papay, John P., Richard J. Murnane, and John B. Willett. 2010. "The Consequences of High School Exit Examinations for Low-Performing Urban Students: Evidence from Massachusetts." *Educational Evaluation and Policy Analysis*, 32(1): 5-23.

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

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

Tinkelman, Sherman. 1965. "Regents Examinations in New York After 100 Years." Albany, NY.

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

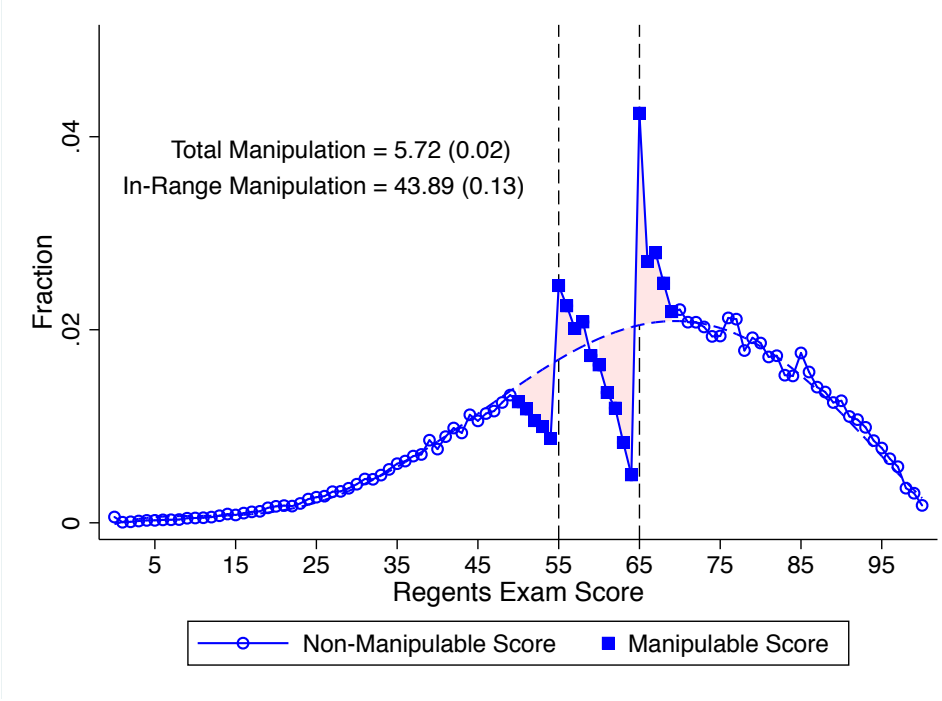

FIGURE 1. TEST SCORE DISTRIBUTIONS FOR CORE REGENTS EXAMS, 2004-2010

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-by-year 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 and detailed in Appendix Table A3. 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.

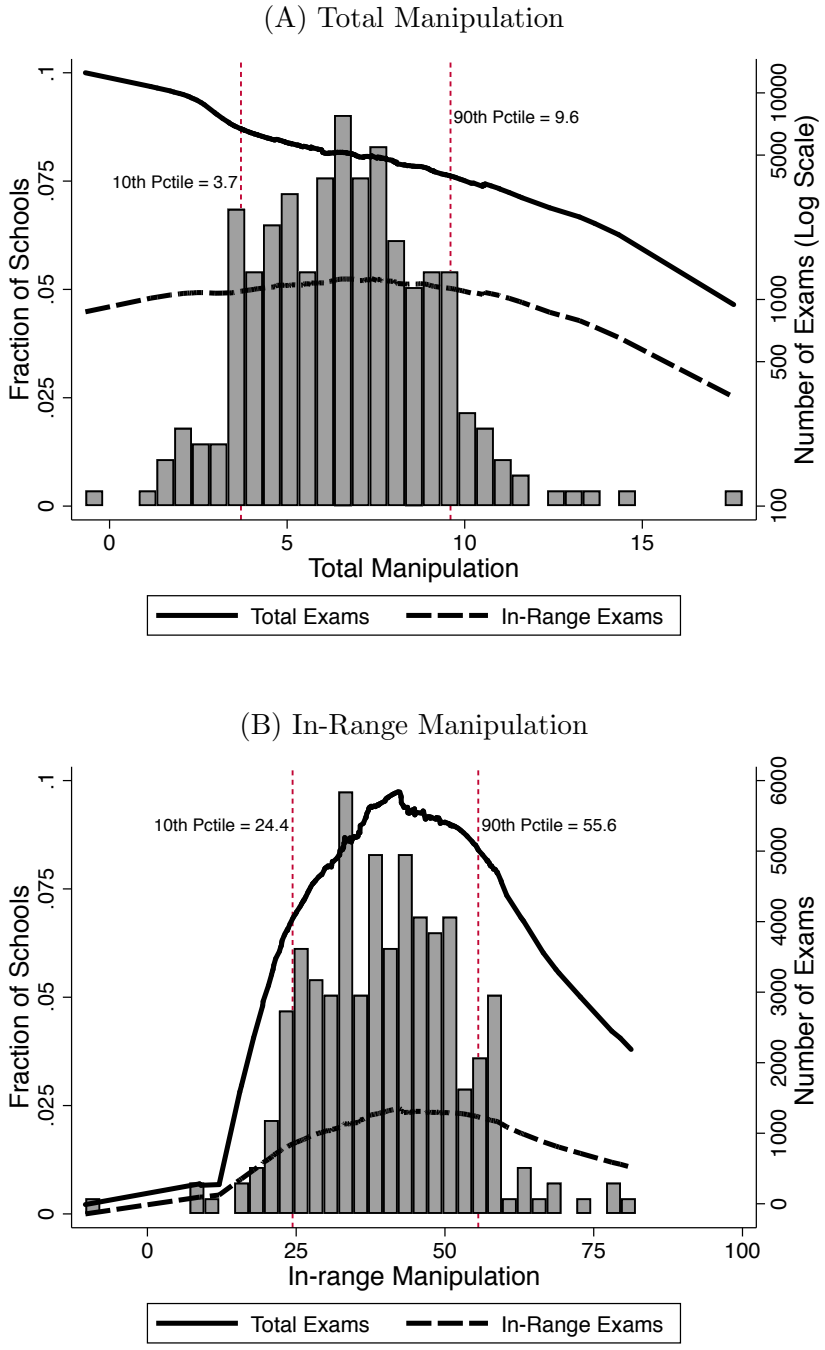

FIGURE 2. DISTRIBUTION OF SCHOOL MANIPULATION ESTIMATES, 2004-2010

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.

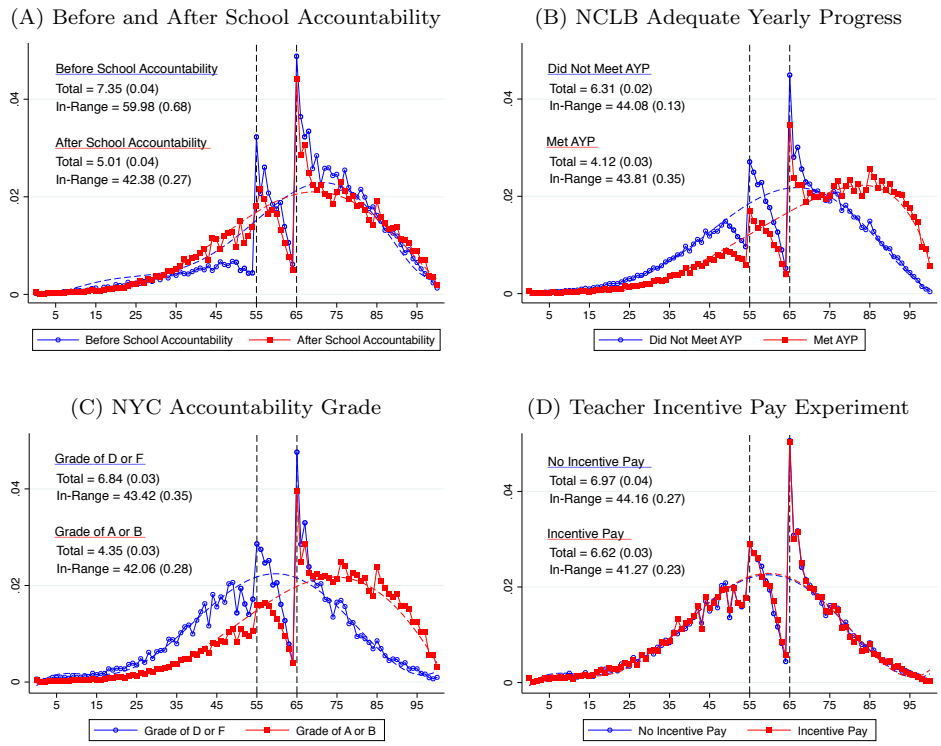

FIGURE 3. RESULTS BY SCHOOL ACCOUNTABILITY PRESSURE, 2001-2010

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.

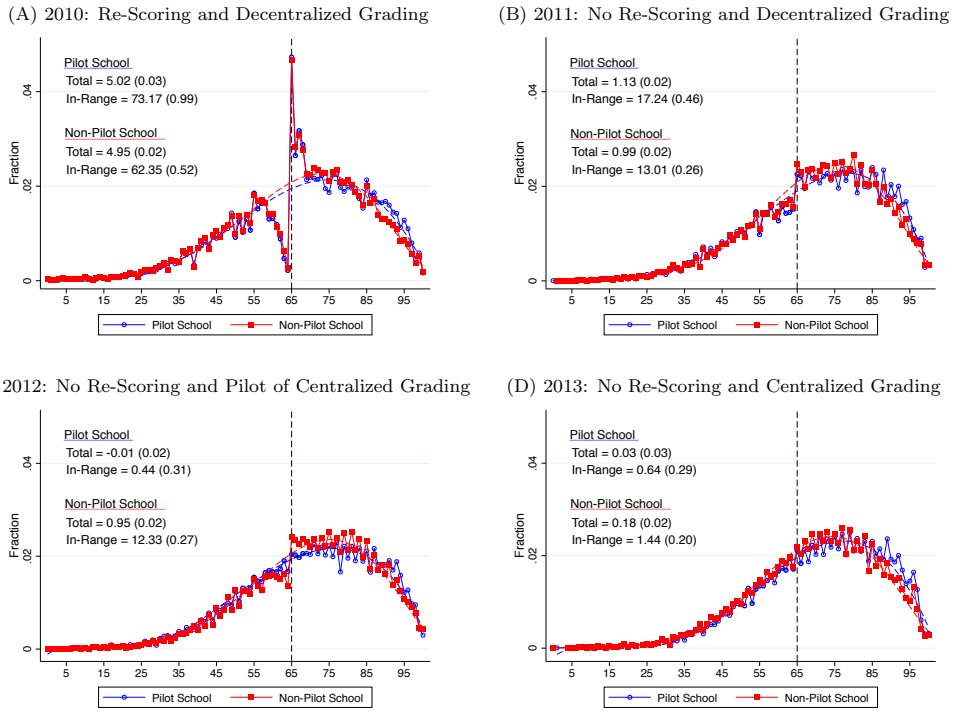

FIGURE 4. TEST SCORE DISTRIBUTIONS BEFORE AND AFTER GRADING REFORMS, 2010-2013

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.

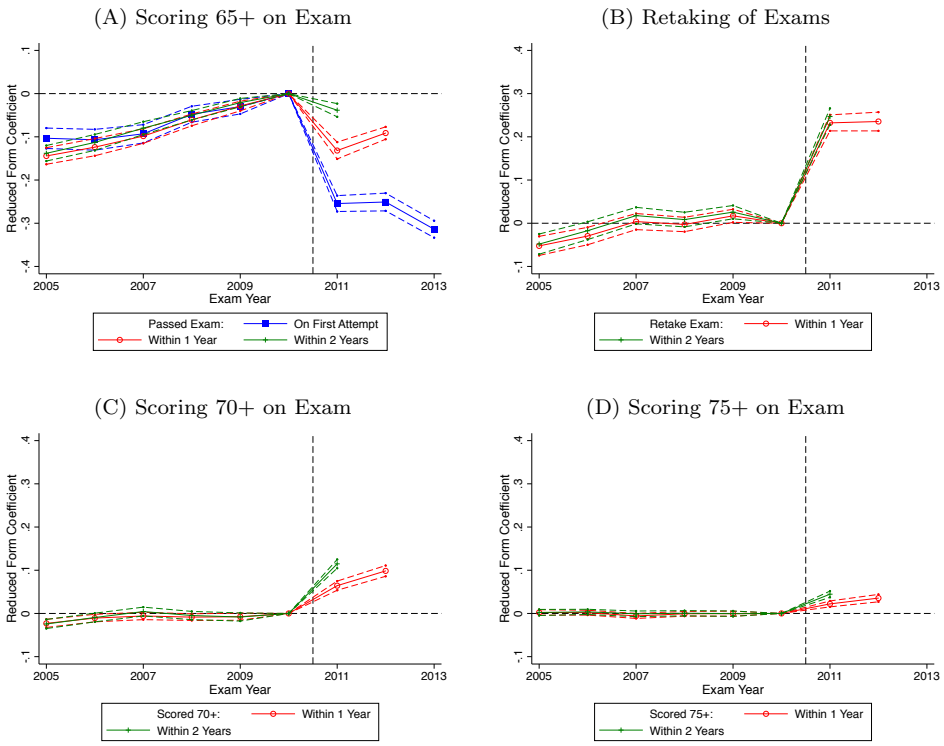

FIGURE 5. REGENTS GRADING REFORMS AND REGENTS OUTCOMES

Note: These figures plot the reduced form impact of the Regents grading reforms on Regents outcomes. The sample includes students taking core Regents exams between 2004-2013. We report reduced form results using the interaction of taking the test in the indicated year and score in the manipulable range around the 65 cutoff. We control for an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects, and exam by year-of-test effects. We stack student outcomes across the Living Environment, Math A/Algebra, and Global History exams and cluster standard errors at the individual and school levels. See the text for additional details.

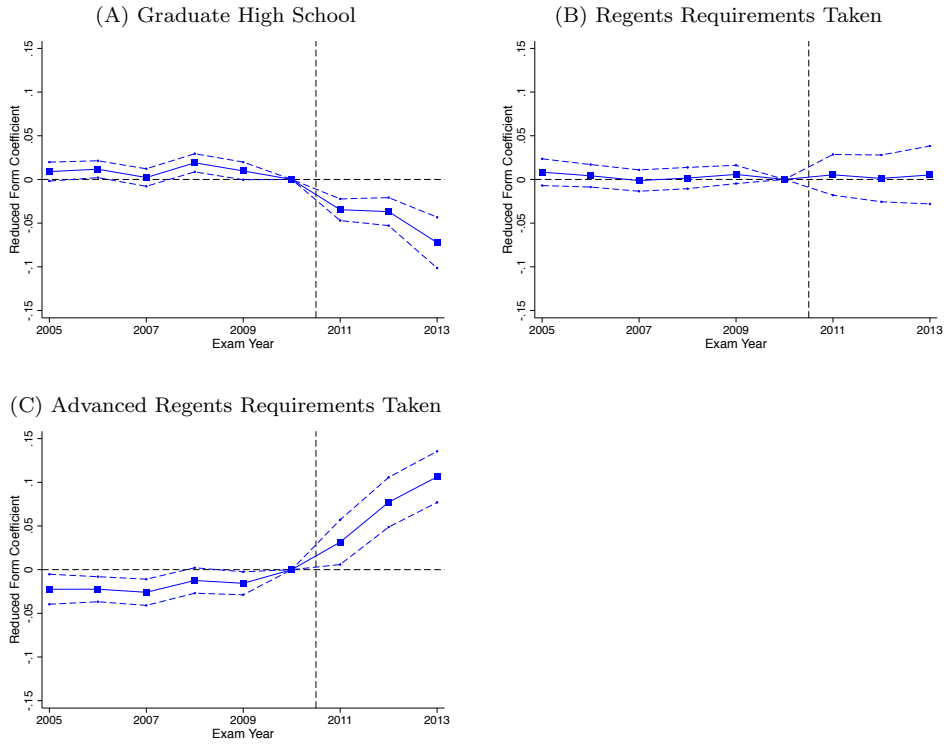

FIGURE 6. REGENTS GRADING REFORMS AND HIGH SCHOOL GRADUATION

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 score in the manipulable range around the 65 cutoff. We control for an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects, and exam by year-of-test effects. We stack student outcomes across the Living Environment, Math A/Algebra, and Global History exams and cluster standard errors at the individual and school levels. See the text 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
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). Enrollment, test score, and high school graduation information comes from Department of Education records. Column 1 reports mean values for the full estimation sample. Column 2 reports mean values for students with all Regents scores 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

	Manipulation				
Panel A: Total Manipulation	(1)	(2)	(3)	(4)	(5)
Percent Black/Hispanic	4.896				2.808
	(0.511)				(0.827)
Percent Free Lunch		4.248			-0.436
		(0.850)			(0.934)
8th Test Score Percentile			-0.051		-0.040
			(0.004)		(0.005)
Enrollment (in 1,000s)				-0.489	0.220
				(0.113)	(0.122)
Constant	2.472	3.533	9.197	7.035	6.306
	(0.396)	(0.526)	(0.272)	(0.263)	(0.937)
R^2	0.250	0.083	0.369	0.064	0.398
Dep. Var. Mean	6.071	6.071	6.071	6.071	6.071
Observations	277	277	277	277	277
Panel B: In-Range Manipulation					
Percent Black/Hispanic	2.152				3.520
	(2.669)				(4.724)
Percent Free Lunch		0.191			-1.052
		(4.018)			(5.337)
8th Test Score Percentile			-0.051		-0.064
			(0.023)		(0.029)
Enrollment (in 1,000s)				0.737	1.720
				(0.528)	(0.699)
Constant	39.080	40.547	43.837	39.207	39.259
	(2.067)	(2.488)	(1.535)	(1.229)	(5.353)
R^2	0.002	0.000	0.019	0.007	0.043
Dep. Var. Mean	40.661	40.661	40.661	40.661	40.661
Observations	277	277	277	277	277

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. See the data appendix for additional details on the sample construction and variable definitions.

Table 3—Student Subsample Results

	Total Manipulation		In-Range Manipulation	
	True Subgroup	Synthetic Subgroup	True Subgroup	Synthetic Subgroup
Panel A: Gender	(1)	(2)	(3)	(4)
Female	5.81	5.67	43.99	43.52
	(0.02)	(0.02)	(0.16)	(0.15)
Male	5.64	5.78	43.74	44.37
	(0.02)	(0.02)	(0.15)	(0.17)
Difference	0.17	-0.11	0.25	-0.85
	(0.02)	(0.04)	(0.22)	(0.33)
Panel B: Ethnicity				
White/Asian	3.64	4.24	46.66	44.91
	(0.02)	(0.03)	(0.46)	(0.30)
Black/Hispanic	6.61	6.37	43.23	43.55
	(0.02)	(0.01)	(0.12)	(0.08)
Difference	-2.97	-2.14	3.43	1.36
	(0.03)	(0.04)	(0.47)	(0.38)
Panel C: 8th Test Scores				
Above Median 8th Scores	3.75	4.93	43.02	43.73
	(0.02)	(0.02)	(0.29)	(0.17)
Below Median 8th Scores	7.86	6.63	44.22	44.02
	(0.02)	(0.02)	(0.12)	(0.15)
Difference	-4.11	-1.70	-1.20	-0.29
	(0.03)	(0.04)	(0.35)	(0.32)

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. See the text for additional details.

Table 4—School-Subject Manipulation and Teacher Turnover

	In-Range Manipulation			
	(1)	(2)	(3)	(4)
Lagged Manipulation	0.504	0.416	0.077	-0.029
	(0.091)	(0.093)	(0.079)	(0.100)
Fraction of Teachers Present in Both Periods			14.060	2.664
			(5.751)	(8.626)
Lagged Manipulation x Fraction Present			0.836	0.870
			(0.254)	(0.276)
Constant	41.525	42.114	33.539	40.244
	(2.445)	(2.086)	(3.515)	(5.095)
R^2	0.439	0.661	0.471	0.679
Dep. Var. Mean	46.420	46.420	46.420	46.420
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. See the data appendix for additional details on the sample construction and variable definitions.

Table 5—Effect of Test Score Manipulation on Regents Outcomes

	Pre-Reform Mean	Reduced Form		2SLS	
Panel A: Scoring 65+	(1)	(2)	(3)	(4)	(5)
Score 65+ in First Administration	0.682	-0.272	-0.272	1.000	1.000
	(0.466)	(0.010)	(0.010)	(0.000)	(0.000)
Score 65+ in First Year	0.767	-0.201	-0.202	0.746	0.752
	(0.423)	(0.010)	(0.009)	(0.027)	(0.027)
Score 65+ in First Two Years	0.805	-0.109	-0.110	0.408	0.415
	(0.396)	(0.008)	(0.008)	(0.024)	(0.023)
Panel B: Retaking Same Exam					
Retake in First Year	0.225	0.155	0.154	-0.575	-0.572
	(0.418)	(0.011)	(0.011)	(0.032)	(0.032)
Retake in First Two Years	0.260	0.188	0.187	-0.707	-0.703
	(0.439)	(0.012)	(0.011)	(0.027)	(0.026)
Panel C: Scoring Above Higher Thresholds					
Score 70+ in First Two Years	0.586	0.098	0.096	-0.367	-0.361
	(0.493)	(0.005)	(0.005)	(0.018)	(0.018)
Score 75+ in First Two Years	0.441	0.040	0.038	-0.149	-0.144
	(0.497)	(0.004)	(0.004)	(0.015)	(0.015)
Score 80+ in First Two Years	0.306	0.013	0.013	-0.050	-0.051
	(0.461)	(0.002)	(0.002)	(0.008)	(0.008)
Score 85+ in First Two Years	0.195	0.006	0.006	-0.022	-0.021
	(0.396)	(0.002)	(0.002)	(0.008)	(0.008)
Observations	1,002,804	1,002,804	1,002,804	1,002,804	1,002,804
Student Controls	–	Yes	Yes	Yes	Yes
Year x Score Trends	–	Yes	Yes	Yes	Yes
School Fixed Effects	–	No	Yes	No	Yes

Note: This table reports estimates of test score manipulation on Regents outcomes. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. Column 1 reports the sample mean for the pre-reform period between 2004-2010. Columns 2-3 report reduced form results using the interaction of taking the test between 2011-2013 and scoring in the manipulable range. Columns 4-5 report two-stage least squares results using the interaction of taking the test between 2011-2013 and scoring in the manipulable range around the 65 cutoff as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring below the manipulable range in 2011-2013, 10-point scale score x subject effects, year x subject effects, and 10-point scale score x subject linear trends. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exam subjects and cluster standard errors at the individual and school levels. See the data appendix for additional details on the sample construction and variable definitions.

Table 6—Effect of Test Score Manipulation on High School Graduation

	Pre-Reform Mean	Reduced Form		2SLS	
Panel A: High School Graduation	(1)	(2)	(3)	(4)	(5)
Graduate High School	0.791	-0.044	-0.046	0.162	0.167
	(0.407)	(0.006)	(0.006)	(0.022)	(0.021)
Panel B: Diploma Requirements					
Regents Requirements Taken	0.890	-0.005	-0.004	0.019	0.016
	(0.313)	(0.012)	(0.012)	(0.044)	(0.043)
Adv. Regents Requirements Taken	0.369	0.031	0.027	-0.114	-0.098
	(0.482)	(0.015)	(0.014)	(0.054)	(0.051)
Observations	1,002,804	1,002,804	1,002,804	1,002,804	1,002,804
Student Controls	–	Yes	Yes	Yes	Yes
Year x Score Trends	–	Yes	Yes	Yes	Yes
School Fixed Effects	–	No	Yes	No	Yes

Note: This table reports estimates of test score manipulation 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. Column 1 reports the sample mean for the pre-reform period between 2004-2010. Columns 2-3 report reduced form results using the interaction of taking the test between 2011-2013 and scoring in the manipulable range. Columns 4-5 report two-stage least squares results using the interaction of taking the test between 2011-2013 and scoring in the manipulable range around the 65 cutoff as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring below the manipulable range in 2011-2013, 10-point scale score x subject effects, year x subject effects, and 10-point scale score x subject linear trends. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exam subjects and cluster standard errors at the individual and school levels. See the data appendix for additional details on the sample construction and variable definitions.

Table 7—High School Graduation Effects by Student Subgroup

	Male	Female	Black/ Hispanic	White/ Asian	Low 8th Score	High 8th Score
Panel A: High School Graduation	(1)	(2)	(3)	(4)	(5)	(6)
Graduate High School	0.153	0.183	0.159	0.192	0.121	0.201
	(0.030)	(0.027)	(0.022)	(0.037)	(0.023)	(0.042)
Pre-Reform Mean	0.757	0.821	0.745	0.887	0.688	0.927
p-value on difference	0.458		0.445		0.094	
Panel B: Diploma Requirements						
Regents Requirements Taken	0.024	0.012	0.023	0.028	0.019	0.065
	(0.047)	(0.044)	(0.032)	(0.055)	(0.027)	(0.061)
Pre-Reform Mean	0.878	0.900	0.871	0.930	0.855	0.950
p-value on difference	0.853		0.942		0.489	
Adv. Regents Requirements Taken	-0.140	-0.060	-0.101	-0.037	-0.085	0.034
	(0.049)	(0.059)	(0.038)	(0.072)	(0.035)	(0.076)
Pre-Reform Mean	0.356	0.380	0.254	0.610	0.163	0.642
p-value on difference	0.299		0.432		0.158	
Observations	472,712	530,092	670,145	322,935	462,417	370,267
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 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 in the manipulable range around the 65 cutoff as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring below the manipulable range in 2011-2013, 10-point scale score x subject effects, year x subject effects, and 10-point scale score x subject linear trends. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exam subjects and cluster standard errors at the individual and school levels. See the data appendix for additional details on the sample construction and variable definitions.