Rules and Discretion in the Evaluation of Students and Schools: The Case of the New York Regents Examinations[*]

Thomas S. Dee
University of Virginia and NBER
dee@virginia.edu

Brian A. Jacob
University of Michigan and NBER
bajacob@umich.edu

Justin McCrary
University of California at Berkeley and NBER
jmccrary@econ.berkeley.edu

Jonah Rockoff
Columbia University and NBER
jonah.rockoff@columbia.edu

February 21, 2011
PRELIMINARY DRAFT, PLEASE DO NOT CITE WITHOUT PERMISSION

Abstract

The challenge of designing effective performance measurement and incentives is a general one in economic settings where behavior and outcomes are not easily observable. These issues are particularly prominent in education where, over the last two decades, test-based accountability systems for schools and students have proliferated. In this study, we present evidence that the design and decentralized, school-based scoring of New York's high-stakes Regents Examinations have led to pervasive manipulation of student test scores that are just below performance thresholds. Specifically, we document statistically significant discontinuities in the distributions of subject-specific Regent scores that align with the cut scores used to determine both student eligibility to graduate and school accountability. Our results suggest that roughly 3 to 5 percent of the exam scores that qualified for a high school diploma actually had performance below the state requirements. Using multiple sources of data, we present evidence that score manipulation is driven by local teachers' desire to help their students avoid sanctions associated with failure to meet exam standards, not the recent creation of school accountability systems. We also provide some evidence that variation in the extent of manipulation across schools tends to favor traditionally disadvantaged student groups.

[*] We would like to thank Tom McGinty and Barbara Martinez of the Wall Street Journal for bringing this issue to our attention and providing us with some of the data used in this analysis. We would also like to thank Don Boyd, Jim Wyckoff, personnel at the New York City Department of Education and New York State Education Department, and seminar participants at Northwestern University's Institute for Policy Research and the Massachusetts Institute of Technology for helpful discussions and comments. Sean Tom provided outstanding research assistance. All errors are our own.

1. Introduction

A fundamental challenge across diverse economic settings where behavior and outcomes cannot be observed easily involves the measurement of performance and the corresponding design of effective incentives linked to those measures. In particular, a key concern in these contexts is that procedures that create high-stakes incentives linked to a particular outcome measure are likely to induce behavioral distortions along other dimensions as agents seek to "game" the rules (see, for instance, Holmstrom and Milgrom 1991, Baker 1992). In recent years, these issues arguably have been nowhere more prominent than in education where student and school accountability policies linked to test scores have expanded dramatically.

The proliferation of test-based accountability in education has generated a variety of concerns about unintended consequences. These concerns have been underscored by evidence that teachers narrow their instruction to the tested content (i.e., "teaching to the test", see Jacob 2005), that instructional effort is targeted to students who are near performance thresholds (Neal and Schanzenbach 2010) and that schools seek to shape the test-taking population advantageously (Jacob 2005, Figlio and Getzler 2002, Cullen and Reback 2002). Jacob and Levitt (2003) also document instances of test-score manipulation on the part of teachers.

In this study, we examine the extent and potential causes of manipulation in the scoring of New York State's signature high school assessment, the Regents Examinations. The Regents are the oldest high school examinations in the U.S., dating back to the mid 19th century, and, more importantly, they are locally evaluated, with teachers scoring the students within their schools. They also carry important stakes. Students' eligibility for graduation and consequences for schools under New York's accountability system are based largely on meeting strict score cutoffs of 55, 65, and 85 on a 100 point scale. Using state-wide data from 2009 and historical data from New York City, we document sharp discontinuities in the distribution of student scores at these three cutoffs, strongly suggesting that teachers purposefully manipulate scores in order to move marginal students over the performance thresholds.

Several pieces of evidence help us to shed light on the mechanisms and motivations for this behavior. First, we find no evidence of any manipulation on the math and English exams given annually to students in grades 3 to 8; these exams are scored centrally, suggesting that local evaluation is a key factor in teachers' willingness to manipulate scores. Second, variation in the methods of scoring across subjects and variation in the distribution of scores around the cutoff points suggest that teachers manipulate students' scores by altering their subjective evaluations of student essays or answers to open-response questions, rather than changing students' answers to multiple choice items (as was found by Jacob and Levitt, 2003). Third, we find that the extent of manipulation was just as pervasive in the earliest years of our data, the school year 2000-2001, before the passage of No Child Left Behind and the creation of New York's accountability systems. This suggests teachers who manipulate students' scores are not driven by a desire to influence these recently created institutional measures of school performance. Fourth, the manipulation of scores is widespread; it occurs across a wide variety of subjects and is present both for low cutoffs and core exams, which determine students' eligibility for a basic high school diploma, as well as for high cutoffs and elective exams, which provide students with additional accolades that may be used as prerequisites for advanced coursework, admission to college, and the granting of college credits.

While many students at the margin of missing an important cutoff may benefit from having a teacher manipulate his/her score upwards, those who are left below face negative consequences. Thus, the use of discretion versus rules in deciding students' final scores raises issues of equitable treatment. We present evidence that teachers are more likely to manipulate scores when their student populations have greater shares of traditionally disadvantaged students.

Our study is organized as follows. In Section 2, we describe the Regents Examinations and their use in student and school evaluations. We also provide details on state-mandated scoring procedures and prior evidence on the quality of these practices from governmental audits. In Section 3, we describe the data and methodologies used in our analysis and present results on the presence and extent of manipulation in Section 4. In Section 5 we discuss our results related to equity. In Section 6, we summarize our conclusions and briefly characterize their implications for policy and practice.

2. New York's Regents Examinations

In 1866, the Regents of the University of the State of New York implemented what was effectively the first statewide system of standardized, high-stakes examinations in the United States (Beadie 1999, NYSED 2008). The first versions of these exams were entrance exams, which were taken prior to attending secondary schools and influenced the allocation of state funds to support those institutions. Beginning in 1878, a new set of Regents examinations functioned instead as exit exams, assessing student performance in the secondary-school curricula and forming the basis for awarding differentiated graduation credentials to students, a practice that has continued in New York to the present.

2.1 Regents Examinations and High School Graduation

In more recent years, public high school students in New York must meet certain performance thresholds on Regents examinations in designated subjects to graduate from high school.[1] Regents exams are administered within schools in January, June, and August of each calendar year and are given in a wide variety of subjects, but scores range from 0 to 100 for every Regents exam. Students typically take exams at the end of the corresponding course, so that most students take the exams in June. Unlike most other standardized exams, teachers grade the Regents exams for students in their own school. 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).

Regents exam requirements have changed somewhat during the years we examine (2001 to 2010). While Appendix Table 1 provides these requirements in greater detail, there are few generalities we can describe here. To graduate, students generally must score at least 55 on each of five "core" Regents examinations: English, Mathematics, Science, U.S. History and Government, and Global History and Geography. In order to receive a more prestigious Regents Diploma, students must receive a score of at least 65 in each of these five core subjects. To earn an "Advanced" Regents Diploma, students must also score at least a 65 on elective exams in math, science, and foreign language.

Currently, the option of receiving a local diploma is being eliminated entirely. Beginning with those who entered the 9th grade in the fall of 2008 (NYSED 2010), students are required to meet the Regents Diploma requirements (i.e., a score of 65 or higher in each of the five core subjects) in order to graduate from high school in New York State.[2] The shift from local diploma to Regents diploma requirements was done gradually, with students entering 9th grade in

[1] In the late 1970's, New York introduced a minimum competency test, the Regents Competency Test, which students were required to pass in order to graduate from high school. However, in the late 1990s, the state began phasing out these tests and replacing them with graduation requirements tied to the more demanding, end-of-course Regents Examinations (Chudhowsky et al. 2002).

[2] New York has recently approved alternatives to the Regents Examinations, based exclusively on comparable performance thresholds in Advanced Placement, International Baccalaureate and SAT II exams (NYSUT 2010). We do not examine data on these exams, none of which are scored locally.

fall 2005 having to score 65 in at least two core subjects, and each subsequent cohort facing stricter requirements. However, the option of a local diploma with scores of 55 or higher in all five core subjects remains available to students with disabilities.

In addition to the importance of cutoffs at 55 and 65, a score of 85 is labeled as achieving "mastery" of the subject matter. While scoring 85 or higher is not relevant for high school graduation, meeting this cutoff is often used by high schools as a prerequisite for courses (e.g., Advanced Placement) and by New York State colleges as either a prerequisite or qualification for credit towards a degree.[3] Beginning with students who entered 9th grade in the fall of 2009, an additional accolade of "Annotation of Mastery" in science and/or math became available for students who score above 85 on three Regents exams in science and/or math.

2.2 The Design and Scoring of Regents Examinations

Regents examinations contain both multiple-choice and open-response (or essay) questions. For example, the English examination includes both multiple-choice questions as well as the opportunity to write essays in response to prompts such as a speech, an informative text with tables or figures, or a literary text. Similarly, the two social-science examinations (i.e., U.S. History and Government and Global History and Geography) include a thematic essay in addition to multiple-choice and open-ended questions. The foreign language exams also contain a speaking component. The sole exception to this design during the time period we examine (2001 to 2010) are the Chemistry exams administered in 2001, which were completely based on multiple choice questions. Scoring materials provided to schools include the correct answers to

[3] Google search using the terms: 'credit "state university of New York" mastery regents 85' returned numerous examples.

multiple-choice questions and detailed, subject-specific instructions and procedures for evaluating open-ended and essay questions.[4]

To help ensure consistency of scoring, essays are given a numeric rating (e.g., on a scale of one to four) by two teachers working independently. If the ratings do not agree but are contiguous, the ratings are averaged. If non-contiguous, a third teacher rates the essay, the modal rating is taken if any two of the three ratings are the same, and the median rating is taken if each of the three ratings is unique. The number of multiple-choice items answered correctly, points awarded on open-ended questions, and cumulative ratings across essay questions are recorded and converted into a final "scale score", using a conversion chart that is specific to each exam. While scale scores range from 0 to 100, typically not all 100 integers are possible on any single exam (see Appendix Table 2 for an example).

During our sample period, Regents exams in math and science with scale scores ranging from 60 to 64 were required to be re-scored, with different teachers rating the open-ended responses. (Two exceptions are the Chemistry examination in June 2001, which was only based on multiple choice questions, and the Living Environment exam in June 2001, where exams with scale scores from 62 to 68 were to be re-scored.) Principals also have the discretion to mandate that math and science exams with initial scale scores from 50 to 54 be re-scored. This policy is clearly importantly for our study. Although we find evidence of manipulation in every Regents

[4] For the English and two social-studies exams, principals are required to designate a scoring coordinator who is responsible for managing the logistics of scoring, assigning exams to teachers, and providing teachers with necessary training. For essay questions, the materials available to support this training include scoring rubrics and pre-scored "anchor papers" that provide detailed commentary on why the example essays merited different scores. For open-ended questions, the materials include a rubric to guide scoring. A single qualified teacher grades the open-ended questions on the social-science exams. In the math and science, the school must establish a committee of three mathematics (or two science) teachers to grade the examinations, and no teacher should rate more than a third (a half) of the open-ended questions in mathematics (science). That is, each member of the committee is supposed to be the first grader of an equal portion of the open-ended response items.

exam for which we have data, the policy of re-scoring in math and science may influence how principals and teachers approach scoring Regents exams more generally.

2.3 Regents Examinations and School Accountability

Since 2002-2003, high schools across the 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 critically on five measures, all of which are at least partially based on the Regents Examinations, particularly in mathematics and English. First, 95 percent of a school's 12th graders (both overall and for sub-groups with 40 more students) must have taken the Regents Examinations in mathematics and English or an approved alternative (NYSED 2010). Second, for both its overall student population and among accountability sub-groups with at least 30 members, performance indices based on the Regents examinations in math and English meet statewide objectives. The subject-specific performance indices are increasing in the share of students whose scale scores on the Regents Examination exceed 55. However, students whose scores exceed 65 have twice the impact on this index.[5] These state-mandated performance objectives increase annually in order to meet NCLB's mandated proficiency goals for the school year 2013-2014.

The fifth measure relevant to whether a high school makes AYP under New York's accountability system is whether its graduation rate meets the state standard, which is currently set at 80 percent. Like the other criteria, this standard is also closely related to the Regents

[5] Specifically, the performance index equals 100 x [(count of cohort with scale scores ≥ 55 + count of cohort with scale scores ≥ 65) ÷ cohort size] (NYSED 2010). So, this index equals 200 when all students have scale scores of 65 or higher and 0 when all students have scale scores below 55.

Examinations, since eligibility for graduation is determined in part by meeting either the 55 or 65 scale score thresholds in the five core Regents Examinations.

2.4. Prior Evidence on Scoring Inaccuracies & Potential Manipulation

In 2009, the New York State Comptroller released the results of an audit of local scoring practices, concluding that the oversight by the NY State Education Department (SED) was not adequate to assure the accuracy of Regents scores, and identifying a number of specific shortcomings in scoring procedures (DiNapoli 2009). The report also makes clear that the Education Department had known about widespread scoring inaccuracies, based on periodic statewide reviews in which trained experts rescore randomly selected exams from a sample of schools throughout the state. A review of June 2005 exams found that, on rescoring, 80 percent of the randomly selected exams received a lower score than the original (i.e., official) score. In 34 percent of cases, the difference in scoring was substantial–as much as 10 scale score points. The audit notes that an earlier analysis covering the school year 2003-2004 found similar patterns.

While the 2009 audit report and earlier departmental reviews clearly suggest the presence of scoring inaccuracies, they provide little sense of whether inaccuracies are the result of human error or purposeful manipulation of scores.[6] They are also limited to an extremely small number of students, schools and subjects, and do not speak to whether the likelihood and/or magnitude of scoring inaccuracies vary across the state or has changed over time.

3. Data and Methodology

We base our analysis on two sets of data. The first is the complete sample of exams administered from August 2008 to June 2009in the five "core" Regents Examinations that

[6] While scores were often re-scored lower, this may have been due to misapplication of the scoring standards by local teachers, rather than intentional score inflation.

determine student eligibility for high school graduation. The data identify the particular test, the student's scale score, and a school and district identifier. Summary statistics on these data are shown in Table 1.

The second set of data covers Regents exams taken by students in New York City from January 2001 to June 2010—including both the core five subjects and other elective exams—and math and English exams taken by students in grades 3 to 8. Like the Regents exams, the tests in grades 3 to 8 contain both multiple choice and open-response items, and they are high stakes. Students who do not meet strict score cutoffs for "proficiency" may be sent to summer school or retained in the same grade level, and the calculation of AYP for elementary and middle schools under No Child Left Behind is based on the fraction of students meeting proficiency cutoffs.

The Regents exam data provide information on the subject, month, and year of the test, the scale score, and a school identifier, but the data on tests in grades 3 to 8 contain information on student demographics, free lunch status, and indicators for limited English proficiency and receipt of special education services. Because both datasets contain a scrambled student identifier, we can link variables for a subset of students present in both datasets. Summary statistics on the New York City Regents examination data, including information on variables linked from the data on students in earlier grades, are provided in Table 2.

We utilize publicly available data at the school level on student demographics, staff characteristics and performance on the state accountability system. We also collected the conversion charts used to determine scale scores for the Regents exams during this time period. We use these charts in three ways. First, they help us identify a handful of observations in the New York City data that contain errors in either the scale score or test identifier, i.e., they do not correspond to possible scale scores on their respective exam. Second, for math and science exams, we use them to map raw scores (which we cannot observe) into scale scores, which can cause predicable spikes in the frequency of scale scores when this mapping is not 1 to 1. Third, on some exams, the scale scores are affected differently by changes in multiple choice and essay questions, and these charts help us to interpret the data and understand which type of questions are most likely the source of score manipulation.

Throughout the paper, we focus on the examinations administered in June, which is when most students are tested. However, the patterns we describe also appear in the January and August administrations of the exam.

4. Results

In Figure 1, we plot the frequency distribution of scale scores in New York State for each of five core Regents exams from June 2009. In the absence of any test score manipulation or odd features of the test metric, one would expect the distributions shown in Figure 1 to be relatively smooth. However, there appear to be large "jumps" in the frequency of student scores as the distribution passes through the 55 and 65 scale score cutoffs. The scores immediately below these cutoffs appear less frequent than one would expect from a well-behaved statistical distribution, and the scores at or just above the cutoffs appear more frequent than one would expect. This pattern is apparent in all five of the core subjects.

The patterns around the passing thresholds in Figure 1 are strongly suggestive of manipulation. To be certain that these patterns could not have occurred simply by chance, we conducted a series of statistical tests, shown in Table 3. We compare the number of students scoring just below and exactly at the passing cutoffs of 55 and 65.[7] Table 3 shows the number of

[7] These comparisons are typically made at 64-65 and 54-55, but occasionally one of these scores is impossible to obtain (e.g., the English Language Arts exam allows scores of 53 and 63, but not 54 or 64.). There is some evidence from Figure 1 of manipulation in the score cutoff at 85, particularly for the science examination. We plan to investigate manipulation at this higher cutoff in future revisions to this paper.

students scoring immediately below and exactly at the cutoff for each exam, as well as the difference in frequencies between these points. For example, in U.S. History and Government, 6,412 students scored at 65 while only 395 students received a score of 64. We conduct simple t-tests and present the corresponding p-values, all of which indicate that the differences we observe in the data are indeed statistically significant – i.e., they are extraordinarily unlikely to have occurred by chance under the null hypothesis that the two scores occur with equal probability.[8]

The frequency distribution for math scores exhibits two other phenomena unrelated to the score cutoffs. First, we see a repeated pattern of alternating frequency for low scores. We believe this is driven by scoring design and the differential points for multiple choice and open ended responses.[9] Second, among relatively high scores there appears to be three somewhat parallel lines. This is due to the fact that multiple raw scores can convert to a single scale score (i.e., a student can answer more questions correctly yet not improve their scale score) at particular points in the distribution. In Figure 2, we plot frequencies that adjust for the mapping of raw to scale scores; specifically, we take the frequency of each scale score and divide this by

[8] One might be concerned that the frequency of scores is generally increasing in the range of the cutoffs, so one might expect the score of 65 (55) to be somewhat more common than the score of 64 (54). While this is true, one can test this by comparing the differences between adjacent scores at these thresholds with the analogous differences between other adjacent scores. In the U.S. History and Geography exam, for example, the number of students scoring 45 and 46 are 1,103 and 1,211, a difference of only 108 students or 10 percent. Similarly, the number of students scoring 76 and 77 was 3,957 and 3,987, a difference of 30 students or 1 percent. These are vastly smaller than the differences we observed at the 55 and 65 thresholds – i.e., 2,529 students or 307 percent at the 55 threshold and 6,017 or 1,523 percent at the 65 threshold.

[9] Correct answers to each question are all worth two raw points. However, partial credit of one raw point may be given on open ended responses. If a significant fraction of low scoring students are awarded no partial credit, the frequency of their raw scores will be higher on even numbers, resulting in this alternating pattern in the data.

the number of raw scores mapping into it. One can see clearly that this adjustment results in a smooth distribution throughout the scale score range *except* around the cutoffs at 55 and 65.[10]

We will argue that patterns around the cutoffs shown in Figures 1 and 2 are driven by manipulation, made possible by local scoring of these examinations. While we cannot know for certain what the frequency distributions of Regents scores *would* have been in the absence of local manipulation, we can compare them to the score distributions for math and English exams taken by New York City students in grades 3 to 8, which involve high stakes around specific cutoffs but are graded centrally by the state.[11] These distributions (shown in Figure 3) appear quite smooth, particularly as they pass through the proficiency cutoff, and estimates of a discontinuity in the distribution at the proficiency cutoff (see McCrary 2008) produce very small point estimates that are statistically insignificant, with t-statistics of about 0.3.

Returning to the score distributions across the five core Regents exams, we see variation in behavior around the 55 and 65 cutoffs across subjects. Specifically, in math and science we see large spikes in frequency precisely at scores of 55 and 65, but scores just slightly higher (e.g., 66 or 67) do not appear with nearly as much frequency and appear in line with a smooth distribution. In contrast, in English and social studies, unusually high frequencies are seen both for scores precisely at the cutoffs (i.e. 55, and 65) and those just above (e.g., 57 and 67 for English).

This variation is strikingly consistent with differences in how the exams are scored and allows us to shed light on the manner in which scores are manipulated. In math and science, it is always possible to award enough additional raw points through partial credit on open-response

[10] One might also be concerned that multiple raw scores map into the scale scores at 55 or 65, which could easily produce spikes in frequency. This is not the case. Indeed, multiple raw scores map into a scale score of 64 of the Living Environment exam, understating the discontinuity at 65 in the unadjusted figures.

[11] While Figures 1 and 2 are based on state-wide Regents exam data, similarly striking discontinuities are seen if we limit the sample to exams taken in New York City schools.

questions in order to move a student from just below the cutoff to exactly a score of 55 or 65. In contrast, for the exams in English and social studies, a score of exactly 55 or 65 may not always be possible if manipulation is done through changes in scores to essay questions. This is because changes in essay ratings of just one point typically change the scale score by four points.

We take English as an example and display its conversion chart in Appendix Table 2. Students with an odd (even) number of multiple choice items correct can, depending on their essay ratings, reach exactly the cutoff at 55 (65); students with an even (odd) number of multiple choice items correct can only reach scores of 53 or 57 (63 or 67), depending on their essay ratings. Thus, a teacher who wishes to alter a student's essay rating by one point in order to move them to the 55 or 65 score cutoffs will often only have the option of moving them strictly past the cutoff, to the next highest score. We therefore argue that the roughly equal jumps in score frequency between pairs of scores surrounding the cutoffs (i.e., 51-55, 53-57, 61-65, and 65-67) is evidence that essay ratings are the primary method through which scores in English were manipulated.[12]

In further support of this argument, we note that in the social sciences, only every third integer in the number of multiple choice items correct allows a student to meet the exact cutoffs at scale scores of 55 and 65, so that teachers willing to raise a student's essay rating by one point in order to meet the cutoff are often forced to move them to two different scores strictly above the cutoff.[13] Thus, the fact that we see roughly equal jumps for three scores on either side of the 55 and 65 cutoffs, with one unusually high frequency just at the cutoff and two strictly above it, provides evidence supporting our hypothesis about the use of essays in score manipulation.

[12] Scale scores of 52, 54, 56, 61, 64, and 66 are not possible on this exam.

[13] For example, to have a scale score of 65 in Global History and Geography, June 2009, the number of multiple choice items answered correctly must be a multiple of three. For other integer values, a teacher willing to increase a student's essay rating by one point in order to raise a student to meet this cutoff can raise the scale score from 62 to 66 or from 63 to 67. A scale score of 64 is not possible on this exam.

A graphical depiction of this argument is shown in Figure 4, where we plot frequency distributions for the English and social studies exams again but now mark those scores which, if the essay rating was changed by one point, could cross through one of the cutoff scores. Aside from these particular scale scores, the frequency distributions appear smooth. Figure 4 also contains frequency distributions for the math and science exams, adjusted as in Figure 2, where we mark scores that just meet the cutoff score, those that fall within the range 60 to 64 (which must be re-scored according to state guidelines), and those within two "raw" points of the 55 and 85 cutoffs. This last demarcation is done under the notion that teachers wishing to manipulate a student's score would be willing to give them partial credit on up to two open-response items.[14] Again, aside from these particular scores, the frequency distributions appear quite smooth.

Having argued that the patterns we see are due to manipulation, we proceed to gauge the magnitude of this behavior. To do so, we use an interpolation strategy to estimate what the "true" distribution should have been in the absence of any score manipulation.[15] We present these predictions in Figure 5, with our interpolation shown by the heavy dashed curve that appears to "connect" the actual data points on either side of the suspect region. By comparing the predicted frequency of scores at different points with the actual frequency of scores at the same points, we estimate that between 13 and 27 percent of exams in the interpolated range had inflated scores (Table 4). This represents between 2.6 and 5.2 percent of *all* students taking each exam, i.e., including those with very high or very low scores. We also estimate the number of student scores affected at each of the cutoffs. For example, on the Global History exam, we estimate that

[14] The choice of two instances of partial credit is admittedly arbitrary but is only used here for illustration and is not crucial to the remainder of our analysis.

[15] Our forecast matches the observed frequencies and slopes at the density for points just outside of the range where manipulation appears to occur (50 and 69 for English and social studies, 50 and 65 for math and science), and the total count of all exams within the range where we expect manipulation. The interpolation has a 4-degree polynomial which fits our five constraints and allows for a closed form solution. Further details on the interpolation are given in Appendix A.

38.4 percent of exams scored just at or above 65 should have scored below 65. The corresponding percentages for English and U.S. History tests are of similar magnitude, while those for math and science are above 50 percent. *4.1 Motivations for Manipulation*

We have established considerable evidence that local scoring of Regents examinations combined with the stakes attached to particular cutoff scores induces teachers to manipulate test scores at particular parts of the score distribution. We have also shown that manipulation is likely due to changes in the grading of essays and open-response items, rather than altering of students' multiple choice responses. However, there are a number of hypotheses for why manipulation might be occurring. One issue may be more demanding graduation requirements, which increased the stakes around meeting the cutoff at 65 and has raised concerns about possible increases in the dropout rate (Medina 2010). In addition, the passage of NCLB and New York City's accountability system (see Rockoff and Turner, 2010), both based heavily on Regents exams, may have driven school staff to manipulate student exam results.

We address both of these hypotheses using data from New York City. To see whether manipulation is driven mainly by a desire to allow students to graduate from high school, we look for manipulation on exams that are not required for graduation but optional for students seeking greater distinction on their diploma. Figure 6 plots frequency distributions for exams on two optional science exams (Earth Science and Chemistry) and an optional math exam (Geometry). 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 purely geared towards meeting basic high school graduation requirements.

To test whether manipulation is driven by the rise of school accountability, we look for manipulation on the core Regents exams administered in June 2001, before the NCLB legislation was passed and many years prior to the creation of the city's own accountability system. In Figure 7 we present distributions for the English, math, and science, but results for social studies exams are quite similar; the figure also displays June 2009 frequencies for purposes of comparison. It is quite evident that manipulation was prevalent in June 2001, well before the rise of school accountability systems. Moreover, it appears that the discontinuities at the 55 scale score cutoff are much wider in 2001 than in 2009, at least in English and math. This is consistent with the reduced importance of this cutoff in the latter period due to changing graduation requirements for students without disabilities.

We can also examine the importance of school accountability by taking advantage of the fact that a considerable fraction of students taking the core math and science Regents exams are still in 8th grade. These are typically advanced students, e.g., taking an Algebra course in grade 8 as opposed to grade 9, who wish to begin fulfilling their high school graduation requirements. While the tests are still high stakes for students, they play no role in school accountability metrics for schools that do not serve high school students (e.g., middle schools). Using our state-wide data for the math and science exams from June 2009, we plot score frequencies separately for schools serving only high schools students and for schools only serving grades 8 and below (Figure 8). As expected, exams from schools serving only grade 8 or lower have higher average scores, but there is clear evidence of manipulation in scores at the cutoffs. Indeed, manipulation in Regents scores around the cutoff of 85 appears substantially greater among schools serving grade 8.

Finally, we take advantage of our historical data to address whether manipulation would have occurred if the Regents exams were purely based on multiple choice items and offered little ambiguity as to students' correct scale scores. To do so, we present the frequency distribution for the June 2001 Chemistry exam, which is the only test during this time period which did not contain any open response or essay questions. The Chemistry exam was not required for graduation, so, for comparison purposes, we also present the frequency distribution from an optional, advanced math exam (Sequential Math 2) also from June 2001. These distributions, shown in Figure 9, present a somewhat mixed picture. Despite the lack of open-response questions, we do see a clear discontinuity in the distribution of Chemistry test scores at the cutoff of 65. However, it is much smaller in magnitude than the discontinuity in the mathematics exam. In Chemistry, 294 students scored a 64, relative to 645 students scoring 65, a ratio of roughly 1:2. In advanced math, 26 students scored a 64, relative to 1,058 students scoring 65, a ratio of roughly 1:40. This suggests that teachers view manipulation on multiple choice items as much more costly than manipulation of scores on open-response items, but not so costly as to eliminate manipulation entirely. Nevertheless, we only have one exam based purely on multiple choice questions and we draw this conclusion with considerable caution.

5. Whose Exam Scores Are Being Manipulated?

We now examine whether the extent of manipulation varies systematically across schools and students with different observable characteristics. If exam scores that initially fell just below 55 or 65 were subsequently changed to meet cutoffs in *all* schools for *all* students, then the manipulation simply reduces the effective cutoff scores by a few points. It is possible that lowering standards in this way may have negative impacts on student and teacher motivation, but that is beyond the scope of this paper. A more important issue in our view is the issue of equitable treatment, since it is evident that not all students with scores just below the cutoffs have their scores manipulated.

We first investigate this issue descriptively by taking the state-wide data and plotting the school-level average characteristics by scale score. If manipulation is concentrated within schools serving particular kinds of students, or if manipulation is targeted at certain types of students within schools, then we should see jumps in average characteristics precisely at the score cutoffs of 55 and 65. Using our state-wide data, we plot the average (school-level) percent of students who are minority (i.e., Black or Hispanic), by score, for each of the five core exams (Figure 10). Because of the relatively high concentration of Black and Hispanic students in New York City, these figures contain separate calculations for New York City and the rest of New York State.

The results suggest that the extent of manipulation around the cutoffs does vary systematically across exams in schools serving different student populations. Specifically, we see sharp increases in the percentage of minority students exactly at the 55 and 65 scale score cutoffs for several exams, and these jumps occur both in New York City and elsewhere in the state. Analogous figures for the percentage of students receiving free lunch and the percentage with limited English proficiency provide similar qualitative results.

Nevertheless, schools serving greater percentages of minority students may differ in other ways. For example, they may have a higher fraction of students scoring near these cutoffs and therefore pay greater attention to the scoring of these exams. In order to provide a clearer picture of how manipulation varies across schools and students, we move to a regression approach. We first isolate exams with scale scores close to the cutoffs at 55 and 65, using our definition of "manipulable" scores outlined above. For each school and subject, we calculate the fraction of these exams that met or exceeded the relevant cutoff score, and then regress these "pass rates" on the school-wide percentages receiving free lunch, limited English proficient (LEP), and minority ethnicity, as well as subject-cutoff fixed effects.[16]

Columns 1 to 3 of Table 5 show regressions of pass rates around the cutoffs on a single demographic variable and subject-cutoff fixed effects. As suggested by our discussion of Figure 10, the percentage of poor, LEP and minority students are all positively related to pass rates at the cutoff scores. When we include the three covariates in the same regression, the LEP and minority coefficients remain the same but the coefficient on free lunch receipt becomes negative and statistically insignificant. We then add a control for the fraction of exams near the cutoff. This coefficient is positive and highly significant, suggesting that schools with greater fraction of students near the cutoff tend to have higher pass rates. Finally, we split the sample and look separately at the cutoffs of 55 and 65. The LEP coefficient remains positive and (marginally) significant for both cutoffs, while the minority coefficient is significant and positive for the pass rate at the 65 cutoff.

Taken together, these results suggest that traditionally disadvantaged students may be helped more than their peers by manipulation of Regents exam scores. We plan to investigate this further in future work, using the New York City data to look within schools.

6. Conclusion

The state of New York has been at the forefront of the movement to establish rigorous academic standards for high school graduates and to hold students and schools accountable for meeting those standards. In particular, over the last 15 years, New York has implemented ambitious graduation requirements linked to student performance on the Regents Examinations,

[16] Regressions are weighted by the number of tests close to the cutoff and standard errors allow for clustering at the school level.

and raised the stakes on these exams for schools and teachers under the state accountability system developed in response to the No Child Left Behind Act (NCLB).

The design of the Regents Examinations may provide a relatively nuanced assessment of student performance in that these tests include open-response and essay questions in addition to multiple-choice questions. However, the scoring of such rich assessments effectively requires that a human rater evaluate each student response to each question. New York currently uses a decentralized, school-based system in which teachers grade the responses for the students in their school. In this study, we present evidence that New York's combination of high-stakes testing and decentralized scoring has led to a striking and targeted manipulation of student test scores. More specifically, we document sharp, statistically significant discontinuities in the distribution of scores on the Regents Examination at the performance thresholds relevant both for eligibility to graduate and for school accountability. Estimates based on these sharp discontinuities imply that a substantial fraction of the exam scores that just met the performance thresholds (i.e., 40 to 60 percent) should have not have been graded as passing.

The social-welfare implications of this manipulation are also not entirely clear. For example, making additional students with marginal test performance eligible for high school graduation may convey substantial economic benefits to them at a comparatively low cost to other, higher-achieving students. However, this behavior undermines the intent of the learning standards embodied in the Regents Examinations and raises substantive issues of horizontal equity. That is, if the degree of test-score manipulation is more common in some schools than others, the awarding of high school diplomas could be viewed as undesirably capricious. Our preliminary findings indicate that manipulation may have relatively larger effects on pass rates among traditionally disadvantaged students, but we plan to pursue this issue much further in revisions to this draft.

Regardless, our results do suggest that there may be attractive, sensible reforms to New York's scoring procedures for the Regents Examination. For example, the statistical procedures used in our analysis could be used to generate school-level estimates of the degree of scoring manipulation. And these school-level estimates can provide a straightforward, cost-effective way to target state audit and training resources to where they are most needed. But it should be noted there may also be common-sense redesigns of the scoring procedures for the Regents Examination that attenuate the need for more costly, labor-intensive audits and training. For example, an approach with strong historical precedent in the context of the Regents Examinations would be to move scoring responsibilities from schools to a central office.[17] Alternatively, the test-score manipulation documented here might be limited if state procedures were changed so that school-based graders only reported the raw-score components for each exam without exact knowledge of how the state would map these results into scale score performance thresholds. Additionally, having a school's tests graded by a neighboring school might also attenuate the clear tendency to rate just-failing scores as passing. Combining pilots of such alternative procedures with careful ex-post assessments may constitute a compelling strategy for ensuring and harmonizing the standards associated with New York's signature assessments.

[17] NYSED (2008) notes that "for many decades all higher-level Regents exams were rated, and diplomas issued in Albany."

References

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.

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

DiNapoli, Thomas P. (2009). “Oversight of Scoring Practices on Regents Examinations.” Office of the New York State Comptroller. Report 2008-S-151.

Medina, Jennifer. “New Diploma Standard in New York Becomes a Multiple-Choice Question,” The New York Times, June 27, 2010, page A17.

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

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

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

New York State Education Department. How No Child Left Behind (NCLB) Accountability Works in New York State: Determining 2010-11 Status Based on 2009-10 Results. Albany NY, October 2010. http://www.p12.nysed.gov/irs/accountability/, Accessed January 29, 2011.

New York State United Teachers. “NYS Education Department Approved Alternatives to Regents Examinations,” Research and Educational Services 10-02, February 2010.

Jacob, B. (2005). “Accountability, Incentives and Behavior: Evidence from School Reform in Chicago.” *Journal of Public Economics*. 89(5-6): 761-796.

Jacob, B. and Levitt, S. (2003). “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating.” *Quarterly Journal of Economics*. 118(3): 843-877.

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.

Cullen, J., Reback, R., 2002. Tinkering Toward Accolades: School Gaming under a Performance Accountability System. Working paper, University of Michigan.

Figlio, D., Getzler, L., 2002. Accountability, Ability and Disability: Gaming the System? Working paper, University of Florida.

Table 1: Summary Statistics for New York State Core Regents Exams, June 2009

	English	Global Hist.	U.S. Hist.	Algebra	Living Environ.
Number of Exams	136,587	207,479	186,240	222,930	206,270
Exams ≤1 Essay Rating from 55 Cutoff	7,963	17,905	11,335	*n/a*	*n/a*
Exams at or just Below 55 Cutoff	*n/a*	*n/a*	*n/a*	7,914	11,335
Exams ≤1 Essay Rating from 65 Cutoff	15,642	25,482	18,339	*n/a*	*n/a*
Exams at or just Below 65 Cutoff	*n/a*	*n/a*	*n/a*	10,060	18,339
Exams ≤1 Essay Rating from 85 Cutoff	20,672	29,572	27,351	*n/a*	*n/a*
Exams at or just Below 85 Cutoff	*n/a*	*n/a*	*n/a*	12,398	27,351
Number of Schools with an Exam	1,086	1,109	1,110	1,826	1,337
Percentage of Schools in NYC	34.7%	33.9%	33.8%	32.0%	33.0%
Mean School-Level Characteristics					
Ending in Grade 8	0.0%	0.0%	1.0%	36.3%	14.6%
Ending in Grade 12	96.5%	94.8%	94.8%	57.9%	78.2%
Other Grade Config.	3.5%	5.2%	4.2%	5.8%	7.2%
Black or Hispanic	39.9%	39.6%	38.8%	37.8%	39.9%
Free Lunch	36.1%	35.8%	35.4%	35.6%	36.7%
Limited English Proficient	5.0%	4.9%	4.6%	4.8%	4.9%

Notes: Scores within 1 essay rating from a cutoff are those for which the addition or subtraction of one point on any essay would move the scale score across the cutoff. Scores at or just below a cutoff refer to the lowest score that meets the cutoff and the highest score that does not meet the cutoff. Mean school characteristics are calculated using one observation per school, conditional on the school having at least one exam in the relevant subject.

Table 2: Summary Statistics for New York City Regents Examinations, 2001-2010

	English	U.S. Hist.	Global Hist.	Math (Core)	Science (Core)	Math (Elective)	Chemistry	Earth Science	Physics	Foreign Language
Regents Dip. Req.	√	√	√	√	√					
Adv. Regents Req.	√	√	√	√	√	√	√*	√*	√*	√
Total # of exams	1,075,792	908,628	1,171,748	479,670	1,024,464	281,782	336,340	432,755	126,140	322,193
# Schools w/ Exams	805	813	829	993	969	689	648	804	434	719
Mean Student Characteristics										
% Linked w/Grade 8 Data	66%	68%	73%	70%	76%	46%	69%	78%	60%	66%
% Black or Hispanic	51%	52%	57%	54%	57%	29%	40%	55%	25%	42%
% Limited English Prof.	16%	14%	14%	14%	15%	10%	10%	13%	7%	17%
% Receiving Free Lunch	63%	62%	62%	44%	59%	44%	55%	57%	49%	55%
% Special Education	5%	5%	5%	6%	5%	2%	1%	3%	0%	1%

Notes: Core science is the Living Environment exam. Elective Math consists of higher-level math: Math B, Sequential Math 2 & 3, Geometry, Algebra 2/Trigonometry pooled. Core Math consists of Math A, Sequential Math 1, and Integrated Algebra pooled. *One physical science required for Advanced Regents Diploma. Average student characteristics are taken from data on testing in grades 3 to 8, linked to students later taking Regents exams, and are calculated at the exam level.

Table 3: Tests of Equal Probability Frequency Around Cutoffs at Scale Scores of 55 and 65

	English	Global Hist.	U.S. Hist.	Algebra (Math Core)	Living Env. (Science Core)
Exams with score of 55	2,607	4,485	3,354	4,651	3,132
Exams scoring just below 55	1,187	1,612	825	3,263	1,100
P-value of difference	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
Exams with score of 65	6,547	7,687	6,412	8,962	8,216
Exams scoring just below 65	1,315	1,016	395	1,100	1,227
P-value of difference	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Notes: Data pertain to Regents exams administered in June 2009. The scale scores just below and at the cutoff for each subject exam are as follows: ELA (53,55,63,65), Global History and Geography (54,56,63,65), Algebra (54,56,64,65), Living Environment (54,55,64,65) and U.S. History and Government (54,56,64,65).

Table 4: Interpolation Estimates of Manipulation on NYS Core Regents Examinations

	English	Global History	U.S. History	Algebra	Living Environment
Number of exams	136,587	207,479	186,240	222,930	206,270
Percent of exams we predicted are inflated	4.0%	5.2%	4.6%	2.6%	3.2%
Exams "in range" of interpolation	26,047	46,539	31,589	43,850	29,161
Percent of "in range" exams we predict are inflated	20.9%	23.1%	26.9%	13.4%	22.6%
Percent of "in range" exams scored 65+ we predict below 65	41.1%	38.4%	39.7%	50.4%	62.6%
Percent of "in range" exams scored 55+ we predict below 55	53.3%	56.6%	65.6%	8.0%	23.2%

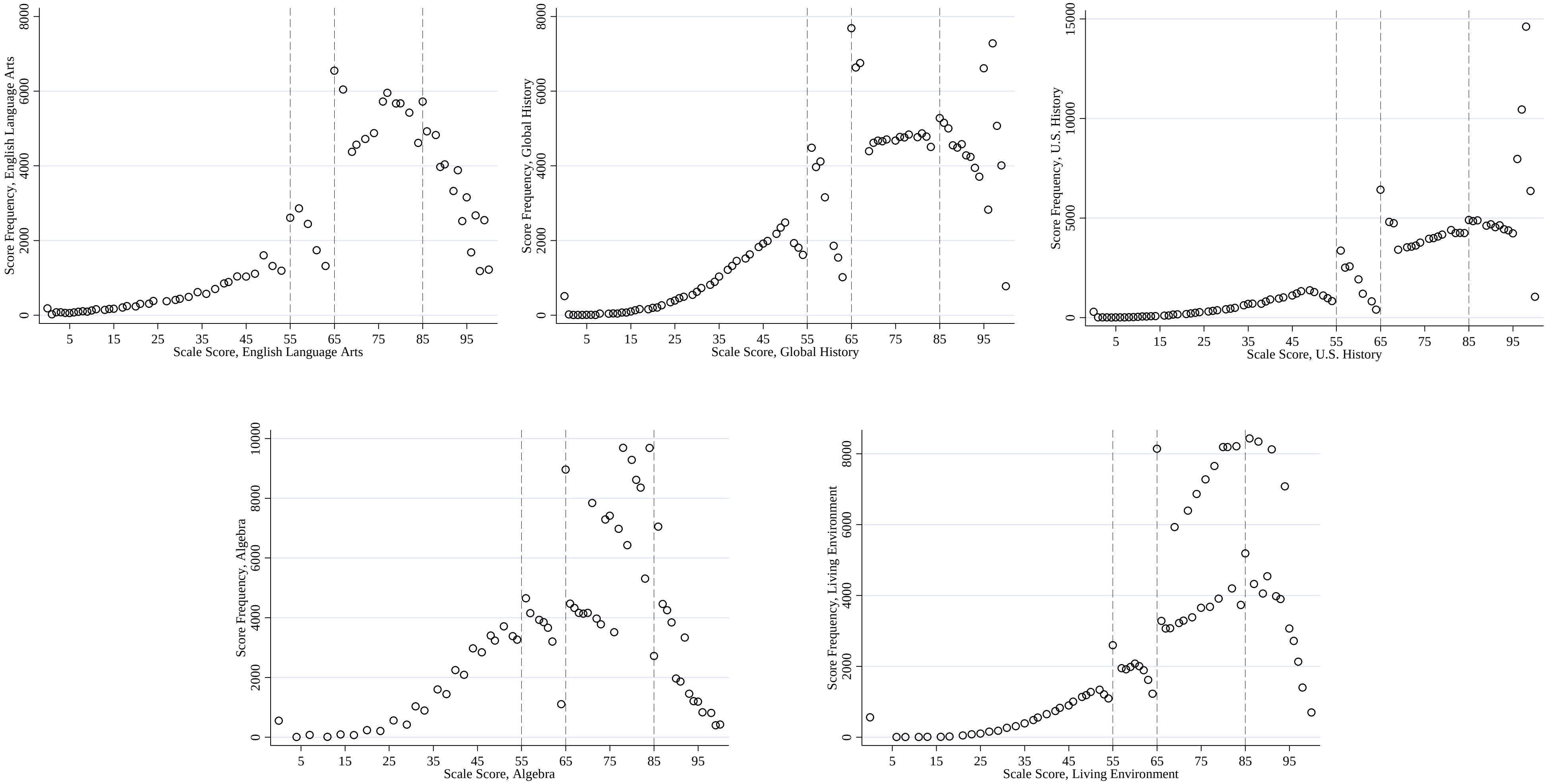
Notes: Interpolation estimates were conducted on core Regents examinations from June 2009. Exams scoring in range of interpolation are defined as follows: English (scores above 49 and below 69); Global History (scores above 50 and below 69); U.S. History (scores above 47 and below 69); Algebra (scores above 51 and below 66); Living Environment (scores above 50 and below 66).

Table 5: Regression Estimates of Manipulation Across Schools and Students

	Fraction of "Close" Exams that Pass Cutoffs						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Percent Free Lunch	0.030			-0.015	-0.026	0.005	-0.042
	(0.010)**			(0.022)	(0.023)	(0.030)	(0.025)+
Percent Limited English Proficient		0.077		0.062	0.065	0.073	0.059
		(0.028)**		(0.032)*	(0.032)*	(0.040)+	(0.036)+
Percent Minority			0.028	0.028	0.019	-0.021	0.041
			(0.008)**	(0.016)+	(0.016)	(0.020)	(0.018)*
Percentage of Exams Close to a Cutoff					0.135	0.126	0.134
					(0.051)**	(0.082)	(0.061)*
Restricted to 55 Score Cutoff						√	
Restricted to 65 Score Cutoff							√
Subject-Cutoff Fixed Effects	√	√	√	√	√	√	√
Observations	10,169	10,197	10,207	10,169	10,169	4,685	5,484
R-squared	0.51	0.51	0.51	0.52	0.52	0.33	0.59

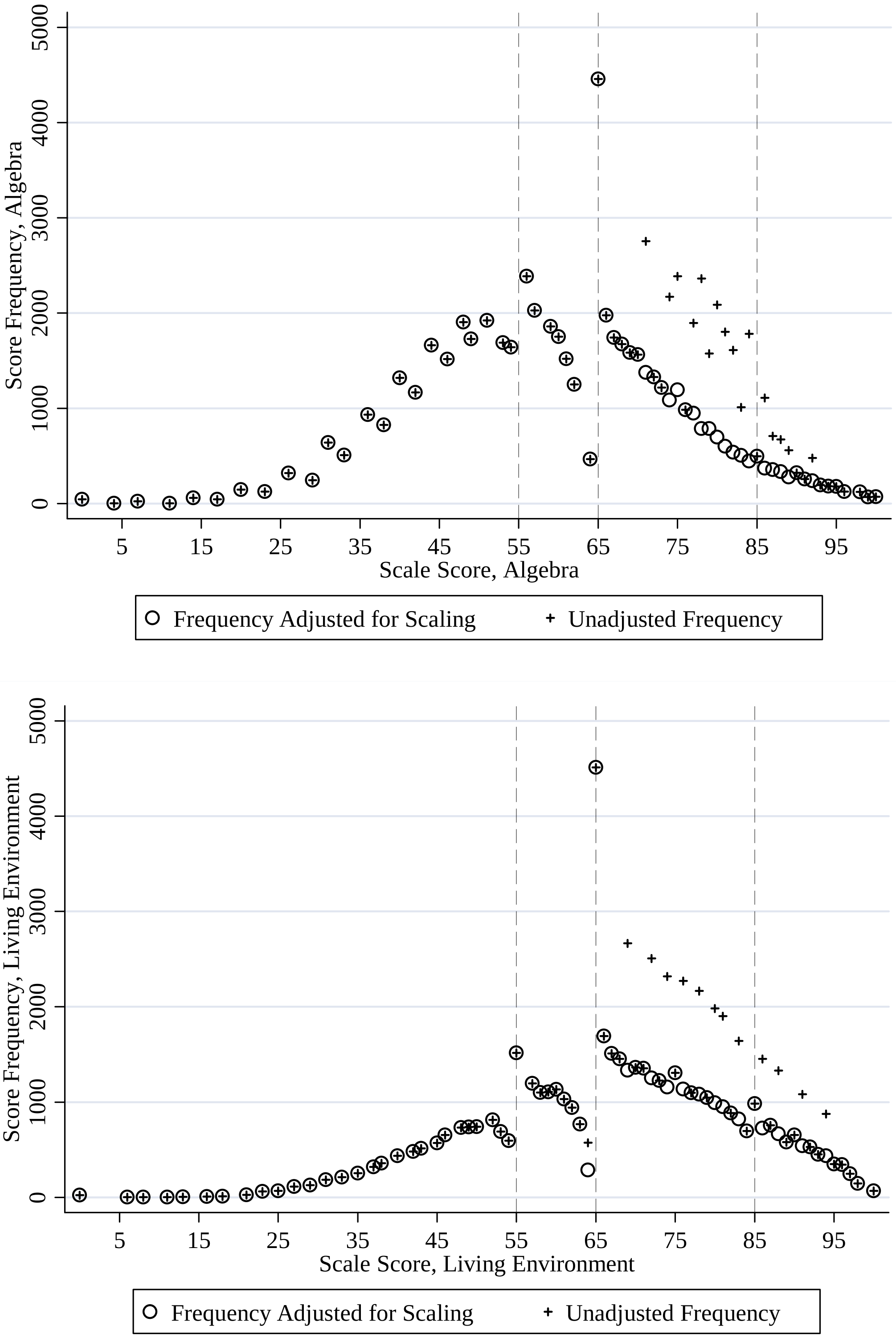
Notes: Regression estimates were conducted using data on core Regents examinations from June 2009, with observations aggregated to the school-subject-cutoff cell level. Regressions are weighted by the number of exams in each cell. Exam scores close to cutoffs are defined as follows: English (scores 51-57 and 61-67); Global History (scores 52-58 and 62-67); U.S. History (scores 52-58 and 62-68); Algebra (scores 53-56 and 60-65); Living Environment (scores 53-55 and 60-65). + significant at 10 percent; * significant at 5 percent; ** significant at 1 percent; standard errors allow for clustering at the school level.

Figure 1: Frequency of Scale Scores on Core Regents Exams



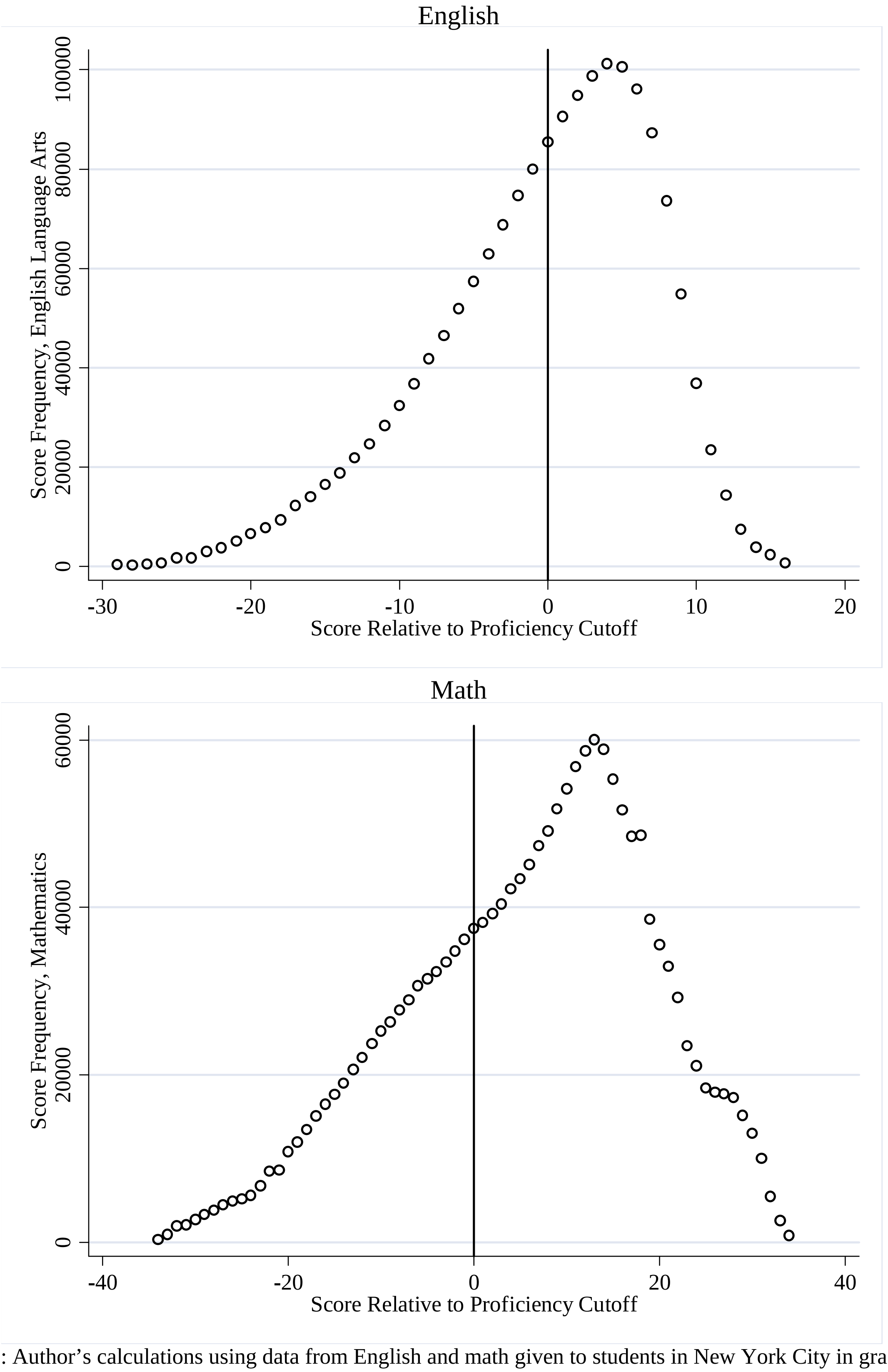
Source: Authors' calculations from data on all Regents exams taken in June, 2009. The x-axis shows students' scale scores, from which cutoffs of 55 and 65 are used for graduation requirements and a cutoff of 85 is used for other accolades. See text for details.

Figure 2: Adjusted vs. Unadjusted Frequencies of Scores in Math and Science



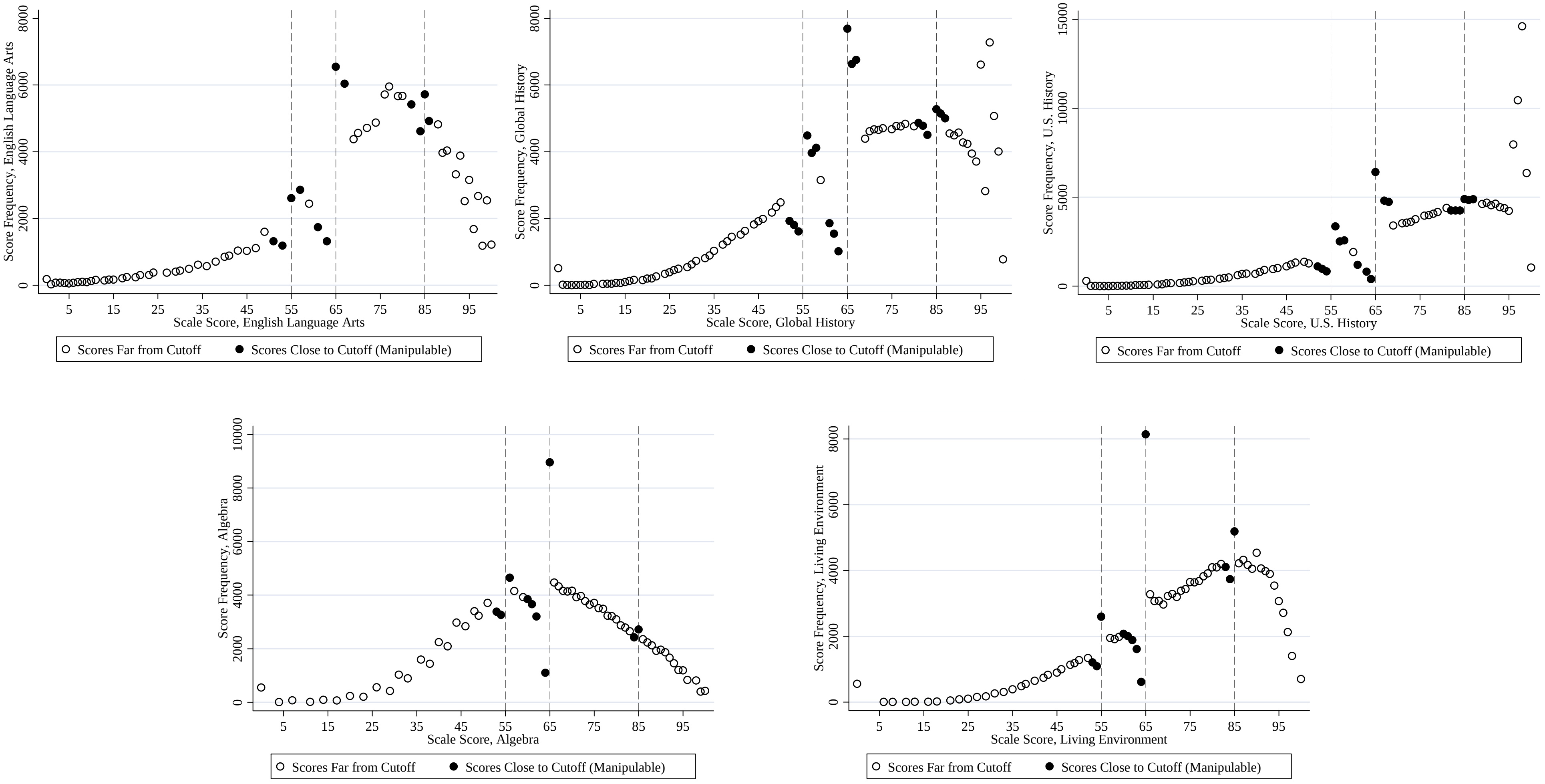
Source: Authors' calculations from data on all New York State Regents exams taken in June, 2009. The x-axis shows students' scale scores, from which cutoffs of 55 and 65 are used for graduation requirements and a cutoff of 85 is used for other accolades. Adjusted frequencies are calculated by dividing the raw frequencies by the number of raw points mapping into each scale score. See text for details.

Figure 3: Frequency Distributions for Centrally Graded Exams in Grades 3 to 8



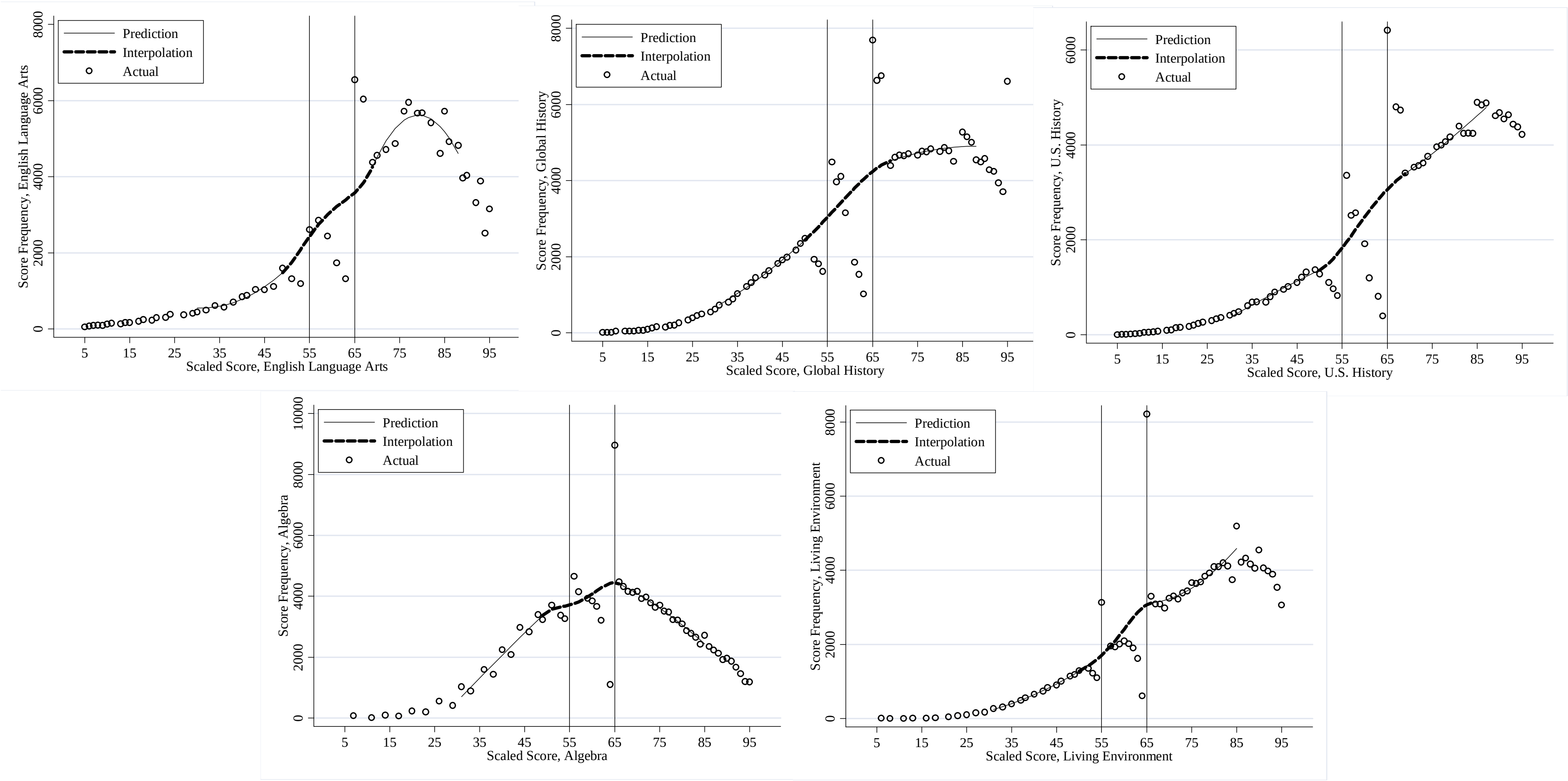
Source: Author's calculations using data from English and math given to students in New York City in grades 3-8 between 2006 and 2009, during which the scoring of these tests remained constant. To pool across grades, scale scores are assigned a rank relative to the New York State cutoff for proficiency. See text for more details.

Figure 4: Frequency of Scale Scores, Adjusted, with “Manipulable” Scores Marked



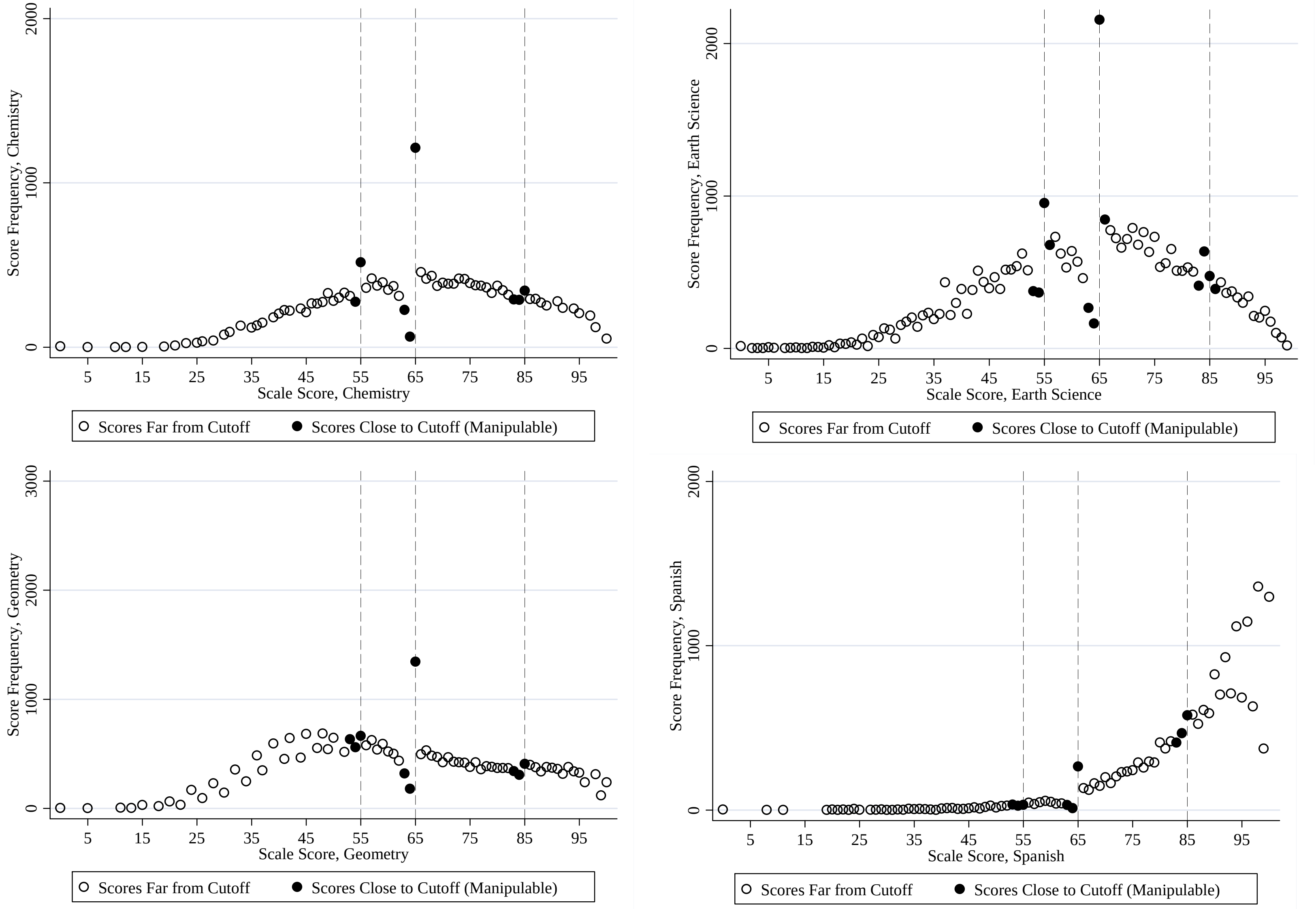
Source: Authors’ calculations from data on all New York State Regents exams taken in June, 2009. The x-axis shows students’ scale scores, from which cutoffs of 55 and 65 are used for graduation requirements and a cutoff of 85 is used for other accolades. Adjusted frequencies are calculated by dividing the raw frequencies by the number of raw points mapping into each scale score, and are only relevant for Algebra and Living Environment exams. Manipulable scores, shown by solid circles, are those which we assert to be most at risk of manipulation by teachers seeking to move students past a cutoff point. See text for details.

Figure 5: Interpolations of Counterfactual Scale Score Distributions, Statewide Core Regents Exams



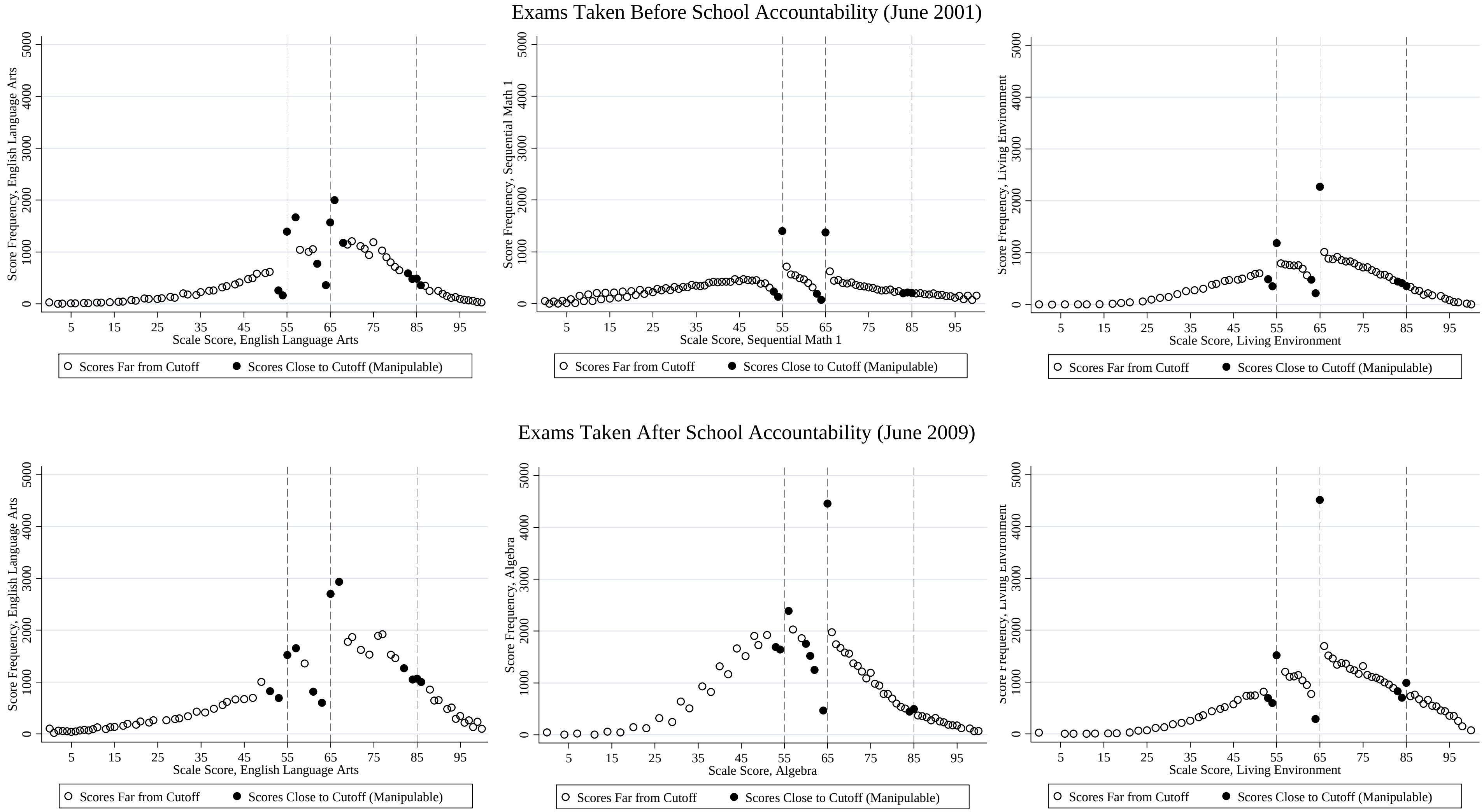
Source: Authors' calculations using data on all New York State Regents exams taken in June, 2009. The x-axis shows scale scores. Score frequencies for Algebra and Living Environment exams are adjusted by dividing the raw frequencies by the number of raw points mapping into each scale score. Interpolation estimates are made in the score region where manipulation is likely to occur, while the prediction lines show local linear estimates of the density of scores to the left and right of the region of interpolation. See text for more details.

Figure 6: Score Manipulation in Elective Exams



Source: Authors' calculations from data on Regents exams taken by New York City students in June, 2009. The x-axis shows students' scale scores, from which cutoffs of 55 and 65 are used for graduation requirements and a cutoff of 85 is used for other accolades. Adjusted frequencies are calculated by dividing the raw frequencies by the number of raw points mapping into each scale score, and are only relevant for Algebra and Living Environment exams. Manipulable scores, shown by solid circles, are those which we assert to be most at risk of manipulation by teachers seeking to move students past a cutoff point. See text for details

Figure 7: Score Manipulation Before and After the Rise of School Accountability Systems

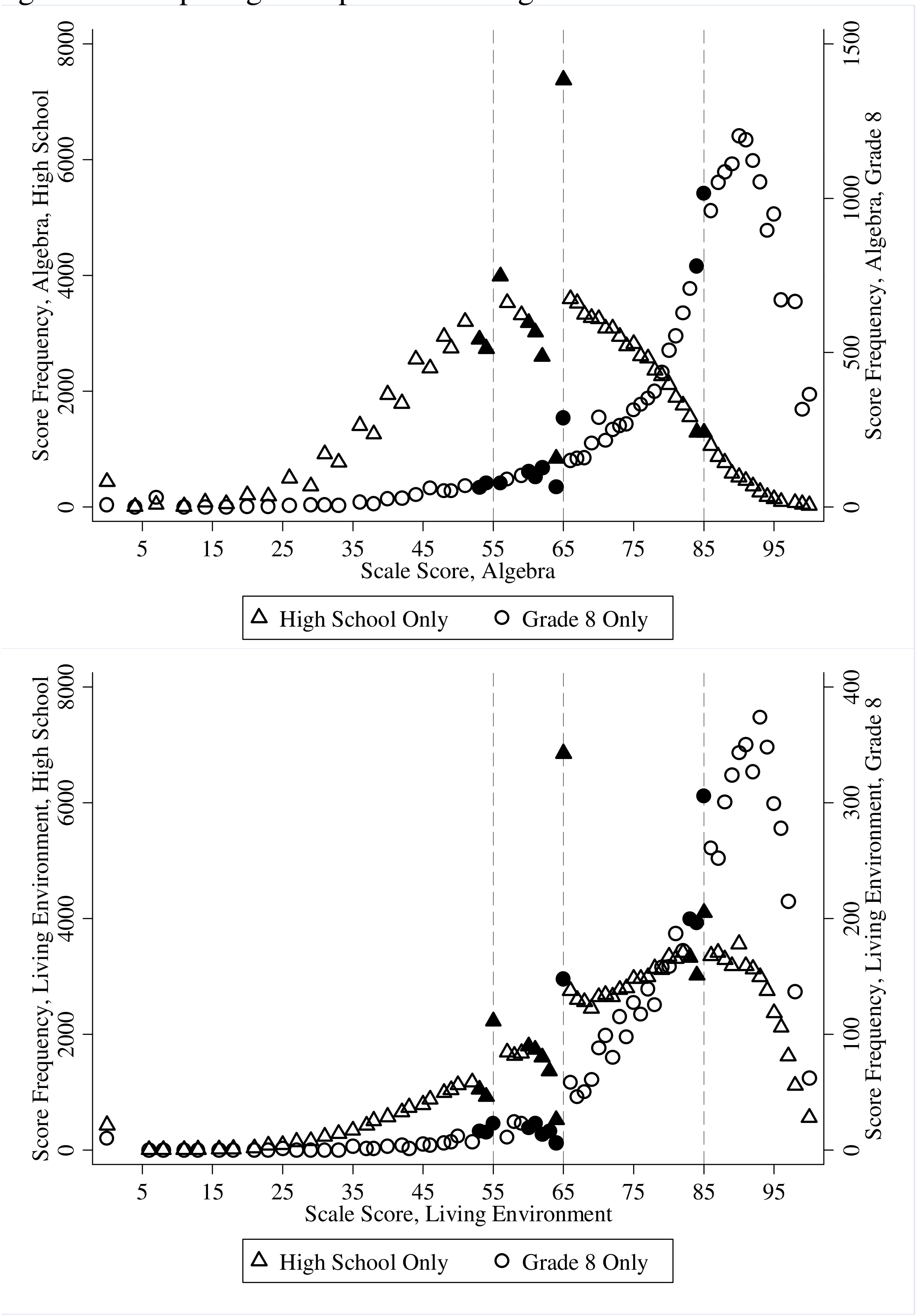


Exams Taken Before School Accountability (June 2001)

Exams Taken After School Accountability (June 2009)

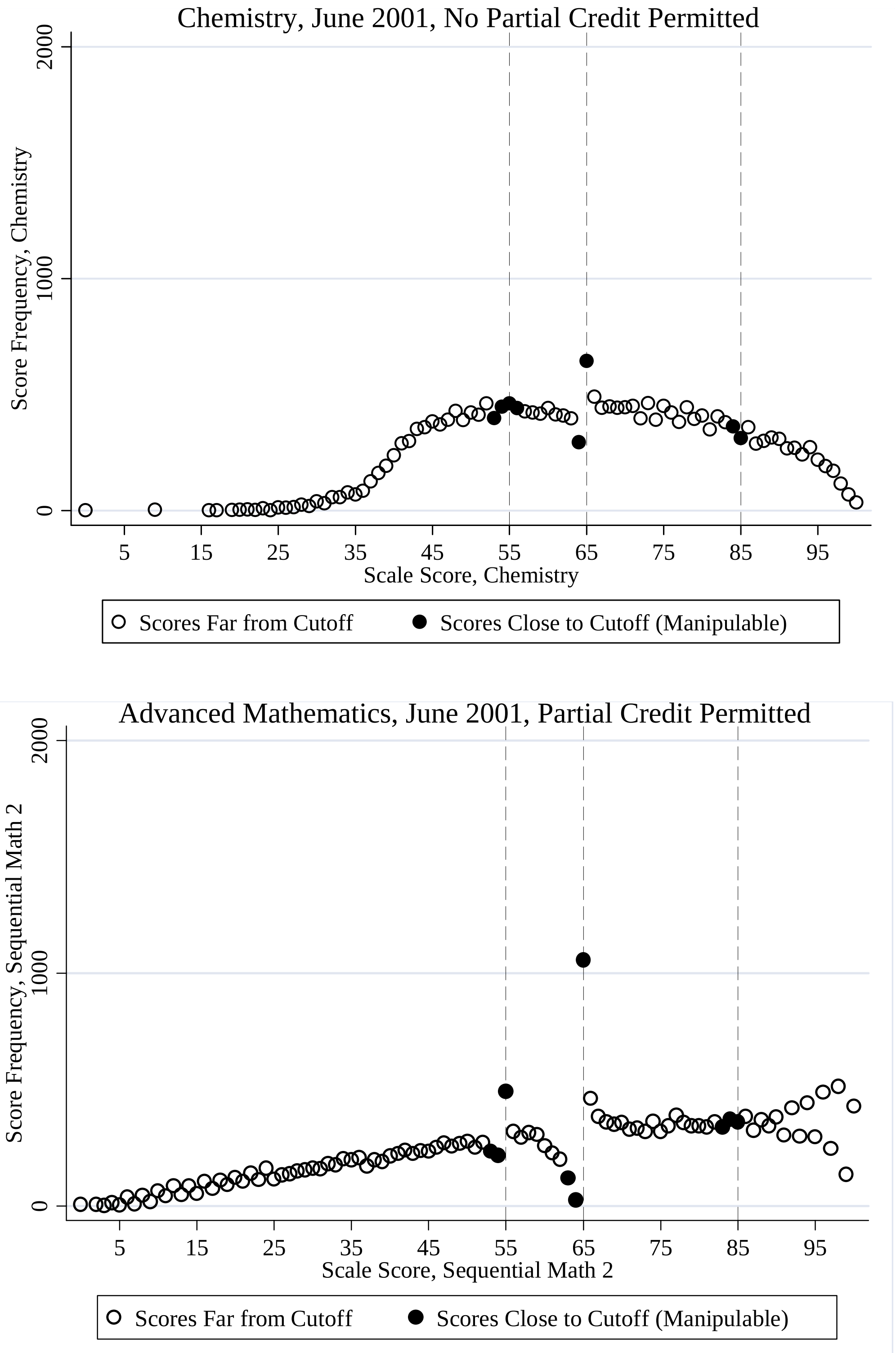
Source: Authors' calculations from data on Regents exams taken by New York City students in June, 2001 and June, 2009. The x-axis shows students' scale scores, from which cutoffs of 55 and 65 are used for graduation requirements and a cutoff of 85 is used for other accolades. Adjusted frequencies are calculated by dividing the raw frequencies by the number of raw points mapping into each scale score, and are only relevant for Algebra and Living Environment exams. Manipulable scores, shown by solid circles, are those which we assert to be most at risk of manipulation by teachers seeking to move students past a cutoff point. See text for details.

Figure 8: Comparing Manipulation in High Schools vs. Grade 8 Schools



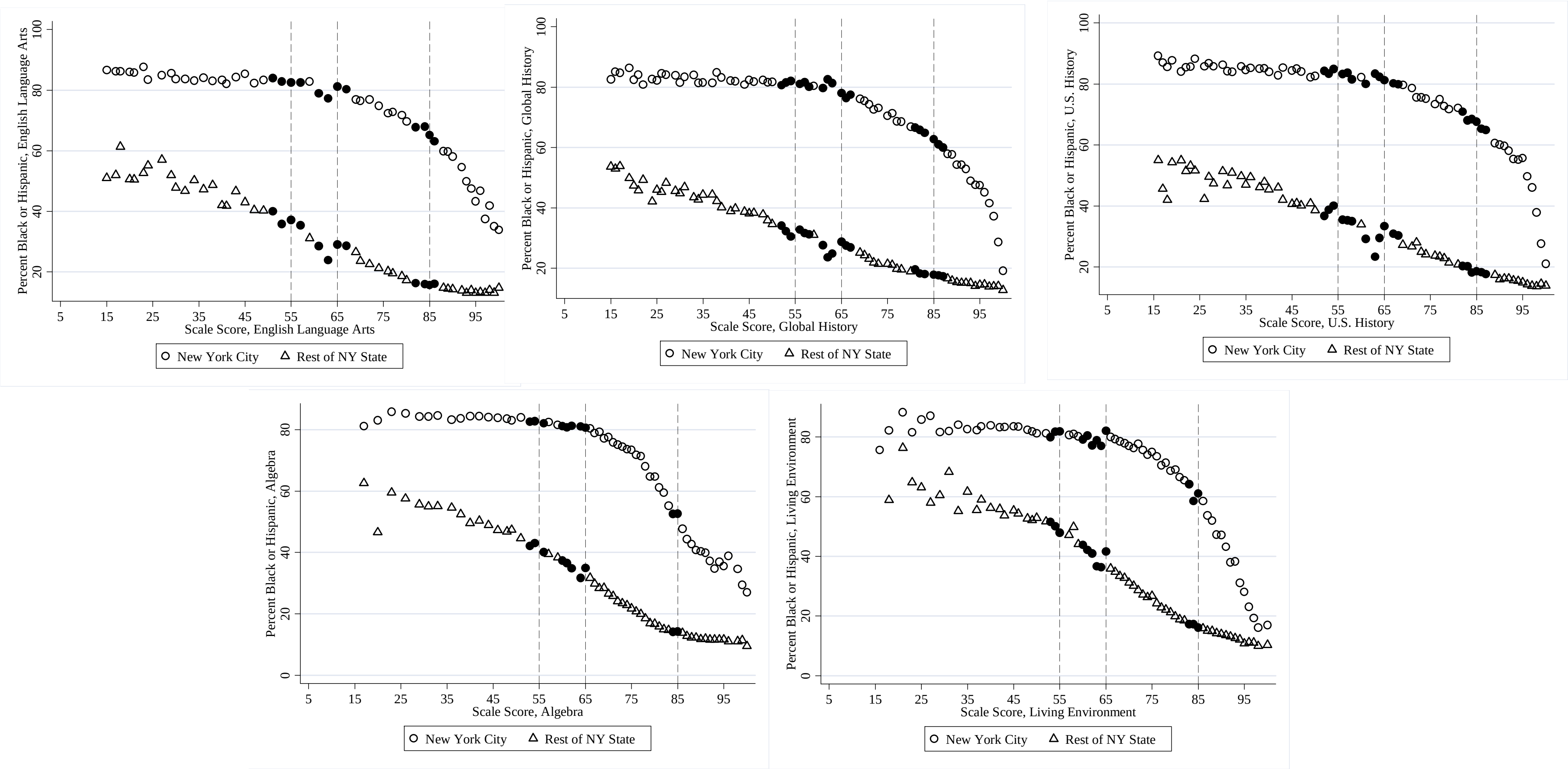
Source: Authors' calculations from data on New York State Regents exams taken in June, 2009. The x-axis shows students' scale scores, from which cutoffs of 55 and 65 are used for graduation requirements and a cutoff of 85 is used for other accolades. The y-axis shows adjusted frequencies, calculated by dividing the raw frequencies by the number of raw points mapping into each scale score. "High School Only" refers to schools serving students in grades 9-12 but not serving students in grade 8 or below. "Grade 8 Only" refers to school serving students in grade 8 and below.

Figure 9: Comparing Manipulation With and Without Partial Credit



Source: Authors' calculations from data on Regents exams in Chemistry and Sequential Math 2 (i.e., an optional, advanced math exam) taken by New York City students in June, 2001. The x-axis shows students' scale scores, from which cutoffs of 55 and 65 are used for graduation requirements and a cutoff of 85 is used for other accolades. Adjusted frequencies are calculated by dividing the raw frequencies by the number of raw points mapping into each scale score. Manipulable scores, shown by solid circles, are those which we assert to be most at risk of manipulation by teachers seeking to move students past a cutoff point. See text for more details.

Figure 10: School-wide Percent Minority by Scale Score on Core Regents Exams



Source: Authors' calculations from data on all New York State Regents exams taken in June, 2009, merged with school level data on student demographics. The x-axis shows students' scale scores, from which cutoffs of 55 and 65 are used for graduation requirements and a cutoff of 85 is used for other accolades. The percentage of minority students is defined as the percentage Black or Hispanic, as measured in New York State school reports. Manipulable scores, shown by solid circles and solid triangles, are those which we assert to be most at risk of manipulation by teachers seeking to move students past a cutoff point. See text for details.

Appendix Table 1: Regents Exam Requirements by Diploma Type and Cohort

Year of 9th Grade Entry	Local Diploma	Regents Diploma	Advanced Regents Diploma
Fall 2001-2004	55+ in 5 core subjects	65+ in 5 core subjects	65+ in 5 core subjects, 65+ in Advanced Math, Physical Science, Language*
Fall 2005	65+ in 2 core subjects, 55+ in 3 core subjects		
Fall 2006	65+ in 3 core subjects, 55+ in 2 core subjects		
Fall 2007	65+ in 4 core subjects, 55+ in 1 core subject		65+ in English, Life Science, Physical Science, US History, Global History, Language*; 65+ in one of the following Math sequences: (Math A/Math B; Math A/Algebra II; Integrated Algebra/Geometry/Algebra II)
Fall 2008-2010	Available only to Students with Disabilities		

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

Appendix Table 2: Conversion of Multiple Choice Items and Essay Ratings to Scale Scores

English Exam, June 2009	Cumulative Essay Rating									
Number Correct on Multiple Choice Items	0	1	…	15	16	17	18	19	…	24
0	0	1	…	30	34	38	41	45	…	65
1	1	1	…	32	36	40	43	47	…	67
2	1	1	…	34	38	41	45	49	…	69
3	1	2	…	36	40	43	47	51	…	70
4	1	2	…	38	41	45	49	53	…	72
5	2	2	…	40	43	47	51	55	…	74
6	2	2	…	41	45	49	53	57	…	76
7	2	3	…	43	47	51	55	59	…	77
8	2	3	…	45	49	53	57	61	…	79
9	3	4	…	47	51	55	59	63	…	80
10	3	5	…	49	53	57	61	65	…	82
11	4	6	…	51	55	59	63	67	…	84
12	5	7	…	53	57	61	65	69	…	85
13	6	8	…	55	59	63	67	70	…	86
14	7	9	…	57	61	65	69	72	…	88
15	8	10	…	59	63	67	70	74	…	89
16	9	11	…	61	65	69	72	76	…	90
17	10	13	…	63	67	70	74	77	…	92
18	11	14	…	65	69	72	76	79	…	93
…	…	…	…	…	…	…	…	…	…	…
25	21	24	…	77	80	84	86	89	…	99
26	23	27	…	79	82	85	88	90	…	100

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