Manipulation in the Grading of New York's 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@law.berkeley.edu

February 1, 2011
IN-PROGRESS DRAFT

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 grading 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. Moreover, we find that the rates of test manipulation in NYC were roughly twice as high as those in the entire state. We estimate that roughly 6 to 10 percent of NYC students who scored above the passing threshold for a Regents Diploma actually had scores below the state requirement.

* 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 the data used in this analysis. We would also like to thank Don Boyd and Jim Wyckoff for helpful comments. 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, Cullen and Reback 2002). Jacob and Levitt (2003) also document instances of test-score manipulation on the part of teachers.

In this study, we present new evidence of manipulation in the grading of student scores on New York's signature high-school assessment, the Regents Examinations. These exams have important consequences for students because they determine eligibility for graduation and for schools because they drive how schools are evaluated under New York's school accountability system. Using student-level data from the 2009 administration of the Regents Examinations, we document sharp discontinuities in the distribution of student scores that align closely with the performance levels relevant for student and school consequences. Our study is organized as follows. In Section 2, we describe the Regents Examinations, the consequences linked to these tests and their state-mandated grading procedures. In Section 3, we discuss the available prior evidence on the quality of these grading practices. In Section 4, we describe the data and methodologies used in our analysis while Section 5 provides a discussion of our results. 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 its modern incarnation, the public high-school students in New York who have met the performance thresholds for the Regents Examinations in designated subjects receive a Regents Diploma (or a Regents Diploma with advanced designation) in lieu of a local diploma. More specifically, in order to receive a Regents Diploma today, students are required to pass Regents

Examinations (i.e., receive a scale score of 65 or higher) in each of five subjects: English, Mathematics, Science, U.S. History and Government, and Global History and Geography.[1]

In the late 1970's, New York also introduced a minimum competency test, the Regents Competency Test (RCT), which students were required to pass in order to receive even a local high-school diploma. However, in the late 1990s, the state of New York began phasing out the RCT and replacing it with local-diploma graduation requirements tied to the more demanding, end-of-course Regents Examinations (Chudhowsky et al. 2002). Currently, the option of even receiving a local diploma is being eliminated entirely, beginning with the graduating class of 2012 (NYSED 2010). In other words, students who entered the 9th grade in the fall of 2008 will be required to meet the rigorous Regents requirements (i.e., a score of 65 or higher in each of the five subjects listed above) in order to obtain any type of high school diploma in New York State.

Table 1 shows how these graduation requirements have changed in recent years. Students who entered the 9th grade in the fall of 2007 (the class of 2011) still have the option of earning a local diploma as long as they achieve passing scores on four Regents examinations and a scale score between 55 and 64 on a fifth exam. The cohort that entered 9th grade in the fall of 2006 (the class of 2010) needs scale scores of at least 65 on three Regents Examinations and 55 on two other exams in order to graduate with a local diploma. And students who entered 9th grade in the fall of 2005 need 65 or higher on two exams and at least 55 on three to receive a local diploma (NYSED 2010). It should be noted that students with disabilities can earn local diplomas with scale scores on Regents Examinations of 55 or higher. Also, New York recently approved performance alternatives to the Regents Examinations; however, these are based exclusively on comparable performance thresholds in Advanced Placement, International Baccalaureate and SAT II exams (NYSUT 2010).

2.2 Regents Examinations and School Accountability

The stakes associated with student performance on the Regents Examinations also apply to schools as well as students because of how high schools are currently 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 based on the Regents Examinations, particularly in mathematics and English. First, New York's accountability system for high schools has two testing-participation criteria (one for mathematics and another for English) that must be met in order for a school to make AYP. Under this standard, 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, whether a New York high school achieves AYP also depends on how its students perform on the mathematics and English Regents examinations (both overall and among accountability sub-groups with at least 30 members). Specifically, the math and English performance criteria require that indices based on the Regents examinations in these two subjects meet statewide objectives. These state-mandated performance objectives increase annually in

[1] Students from all of these cohorts could also earn a Regents Diploma with advanced designation by passing Regents Examinations in other designated subjects (i.e., 7 or more in total). Additionally, students whose scores exceed 85 on designated exams receive an "Annotation of Mastery" on their Regents Diploma. We find some evidence of manipulation around this performance threshold as well. However, it is less common and not the focus of this study.

order to meet NCLB's mandated proficiency goals for the 2013-14 school year. The subject-specific performance indices are increasing in the share of students whose scale scores on the Regents Examination that exceed of 55. However, students whose scores exceed 65 have twice the impact on this index.[2]

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 test-participation and performance criteria, this standard is also closely related to the Regents Examinations. Specifically, as detailed above (Table 1), eligibility for graduation is determined in part by meeting either the 55 or 65 scale-score thresholds in the five core Regents Examinations.

2.3 The Design and Grading of Regents Examinations

Regents Exams are administered within schools in January, June, and August of each calendar year. Students typically take the exam at the end of the corresponding course, so that most students take the exams in June. Each exam period spans 3 hours except for the comprehensive English exam, which is conducted in two 3-hour sessions on separate days.

The test content is explicitly based on the state's core curriculum in each of the five subjects. Importantly, all of the Regents Exams contain both multiple-choice and open-response (or essay) components. 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, and literary texts. 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 Regents Examinations for mathematics (i.e., Integrated Algebra) and science (i.e., Living Environment) also include open-ended questions in addition to multiple-choice questions.

Unlike most other standardized exams, teachers grade the Regents Examinations for students in their own school (both the multiple-choice as well as the open-response and essay components). The State Education Department of New York provides explicit guidelines for how the teacher-based scoring of each Regents Examination should be organized (e.g., NYSED 2009). For each of the five exams, the materials provided to schools include the correct answers to multiple-choice questions, which are either hand scored or machine scored within schools.

However, the state-required approaches to scoring the open-ended and essay questions differ by subject. For the English and two social-studies exams, principals are required to designate a scoring coordinator who is responsible for managing the logistics of grading each exam and the necessary rater training. The materials available to support this training include scoring rubrics and pre-scored "anchor papers" that provide explanatory commentary on why the example essays merited different scores. A single qualified teacher grades the open-ended questions on the social-science exams. However, the scoring coordinator assigns *two* qualified teachers to rate each English (and social-science) essay independently.[3] The coordinator takes

[2] 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.

[3] Sometimes, the ratings submitted by two teachers for a given essay do not agree. If they are contiguous, the scores are averaged and, when the raw scores across multiple essays are aggregated, any remainders are rounded up. However, if they are neither congruent nor contiguous, a third rater evaluates the essay. If two of the three scores are the same, then this "modal" score is recorded. If each of the three scores are different, then the median score is recorded.

the raw scores from the multiple-choice, open-ended, and essay questions and, using a conversion chart, identifies and records the student's final scale score for the given Regents examination.

In the math and science, the school must establish a committee of three mathematics (or two science) teachers to grade the Regents Examinations. The materials available to these graders include the correct answers to the multiple-choice questions and a rating guide that includes a rubric to guide the scoring of open-ended questions. Only one teacher initially scores each open-ended response, and the state procedures suggest that no teacher should rate more than a third of the open-ended questions in mathematics or more than half of those in 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. If a student's final scale score is from 60 to 64 (i.e., just failing), the entire exam is to be rescored but with a stipulation that a different teacher rates the open-ended responses on the exam in question. Principals also have the discretion to mandate that the committee rescore exams with initial scale scores from 50 to 54.

3. Prior Evidence on Grading Inaccuracies & Potential Manipulation

New York's increasingly demanding graduation requirements have intensified the high stakes tied to student performance on the Regents Examinations, and raised concerns about possible increases in the dropout rate (Medina 2010). With the passage of NCLB, Regents Exam scores have become increasingly high-stakes for teachers and schools as well. These pressures, combined with the unusually decentralized grading procedures for the Regents Exams, have fueled concern that school staff may be manipulating student exam results.

In 2009, the NYS Comptroller released the results of an audit it conducted to determine whether oversight of local scoring practices was adequate to assure the accuracy of Regents scores (DiNapoli 2009). They concluded the oversight by the NY State Education Department (SED) was not adequate, and identified a number of specific shortcomings in the procedures used to score the exams.

Perhaps most disturbingly, the audit makes clear that SED has known about widespread scoring inaccuracies for years. SED conducts periodic statewide reviews in which trained experts from the department rescore randomly selected exam from a sample of schools throughout the state. In a review of June 2005 exams, for example, SED found that, on rescoring, 80% 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 likely large enough to result in a substantial difference – perhaps as much as10 scale score points – in a student's final grade. The audit notes that an earlier analysis covering the 2003-04 school year found similar patterns.

While the 2009 audit report and earlier departmental reviews clearly suggest the presence of grading inaccuracies, they have several important limitations. First, they are limited to an extremely small number of students, schools and subjects, and thus do not allow one to determine the extent of the problem. Similarly, they are not able to determine whether the likelihood and/or magnitude of scoring inaccuracies vary across the state. For example, are such inaccuracies more prevalent in high-poverty inner-city schools or more affluent suburban schools? Second, they do not provide a clear sense of the magnitude of the mis-grading. For example, what fraction of students who were given passing grades would have failed a particular exam, and perhaps more importantly, would have failed to graduate high school, with more objective grading?

4. Data and Methodology

In order to examine the existence and magnitude of scoring inaccuracies on the NYS Regents Examinations, we analyzed data on the set of exams administered in 2009 (i.e., associated with the 2008-2009 school year). The data file we use contains scale scores for every student who one of the five "core" Regents Examinations listed above, during any one of the testing dates in 2009, along with the student's school and district.[4] For some analyses, we also utilize publicly available data on student demographics, staff characteristics and school performance on the NYS accountability system.

The analysis results presented below focus on the main administration of each exam, which occurs in June, although the patterns we describe also appear in the January and August administrations of the exam (see Appendix Figure 1).

5. Results

Figure 1 shows the distribution of student scores in New York State for each of the June 2009 exams. The height of each dot represents the number of students in NYS who received the corresponding scale score on this exam. Note that some dots appear to be "missing" because certain scale scores are not possible on some exams due to the conversion between raw and scale scores. Importantly, the horizontal lines at 55 and 65 indicate the passing thresholds that are relevant to being eligible for a local and Regents diploma respectively (Table 1).

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. That is, one would not expect very large jumps between two adjacent scale scores. However, it is obvious from these figures that there are substantial discontinuities in the data. In particular, there appear to be large "jumps" in the number of students scoring right at or above 55 and 65. Indeed, the scores immediately to the left of these cutoffs appear to be less frequent than one would expect from a well-behaved statistical distribution, as if the scores just below the passing thresholds were "shifted" to just above the passing thresholds. This pattern is striking in all five subjects and at both the 55 and 65 passing thresholds.

In several of the subjects (e.g., Integrated Algebra and Living Environment), there also appears to be two somewhat parallel lines, particularly to the right of the Regents threshold of 65 scale points. We strongly suspect that this odd feature is due to the conversion of the raw-score components of each subject exam into a subject-specific scale score. It is common in standardized achievement exams for multiple raw scores to map into the same scale score, such that there are more possible ways to obtain certain scale scores than others. For example, in the June 2009 Living Environment exam it appears that scores 78, 80, 81 and 83 were nearly twice as frequent as scores of 79, 82 and 84. While these patterns look odd in the figures, they do not have any impact on the analysis we conduct below because they occur further away from the critical passing thresholds that we study.[5]

While the patterns around the passing thresholds in Figure 1 are strongly suggestive of manipulation - particularly because these sharp discontinuities occur across five different

[4] Several of those who took part of the English exam in January 2009 were unable to complete the second day of testing because of inclement weather; they were allowed to complete their exam in June.

[5] One can also see a similar, though much less pronounced pattern, between scores of 25 and 50 on the Integrated Algebra exam. But, again, because these patterns occur outside of our focus area, they will not materially impact the analysis we present here. And, if one created identical figures using the underlying raw scores rather than the scale scores, these odd patterns would presumably disappear.

subjects - it is still conceivable that these patterns may simply have occurred by chance. In order to test this, we conducted a series of statistical tests, shown in Table 2. One of the most straightforward ways to examine this issue is to simply compare the number of students scoring just below and exactly at the passing cutoff. This is typically at 64 and 65 for the Regents diploma and 54 and 55 for the local diploma (Table 1). However, these cutoffs vary slightly across subjects due to the raw-scale score conversion issue discussed above, which occasionally makes one of the threshold scores impossible to obtain (e.g., on the English Language Arts exam the equivalent threshold scores are 53 and 55 and 63 and 65 because the exam does not allow scores of 54 or 64).

Table 2 shows the number of students scoring immediately below and exactly at the cutoff for each exam, as well as the difference in frequencies between these points. Looking at the results for U.S. History and Geography around the 65 threshold, we see that 6,412 students scored at 65 while only 395 students received a score of 64, so that the difference is 6,017. A simple t-test and the corresponding p-value are shown below. In all cases, the p-values shown in this table indicate that the differences we observe in the data are indeed statistically significant – that is, they are extraordinarily unlikely to have occurred by chance.

However, in all of the figures, the frequency of scores appears to be increasing even before these suspicious thresholds, 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.

By calculating differences between each of the possible adjacent scores, one can determine whether the jumps one sees as 55 and 65 are statistically different from the patterns in the rest of the data. As is obvious from a casual inspection of the pictures in Figure 1, we confirm that these jumps are statistically different.

Having confirmed that the patterns we see in Figure 1 are not simply due to chance, it is valuable to gauge the magnitude of the test score inflation we see. In order to do this, one must determine what the "true" distribution should have been in the absence of any score manipulation or grading inaccuracies. To do so, we first identified a range of scores near the two performance thresholds that appeared to be subject to potential score manipulation. Note that this range is actually much larger than the two points we have been discussing above (i.e., 64-65 and 54-55). Looking at Figure 1, it appears that graders shifted students within roughly 5 points below each of the cutoffs. For example, in English Language Arts, the frequency of students drops from 49, through 53 before jumping at 55 (note that the scores of 50, 52 and 54 are not possible). Similarly, the number of scores drops from 59 to 63 before jumping up at 65 (note that the scores of 60, 62 and 64 are not possible). For each subject, we denote these potentially suspect regions with dashed vertical lines in Figure 2.

Next, we interpolate the distribution in this suspect region. Our prediction is 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 can estimate the number of scores that have

been inflated. Table 3 presents a variety of estimates that speak to the magnitude of the likely score inflation in the June 2009 Regents Examinations.

We estimate that between 5,000 and 11,000 of students taking each exam had their scores inaccurately inflated. This translates to between 2.5 and 5.2 percent of students taking the exams. However, one might legitimately argue that students whose true scores were very high or very low were not "at-risk" of having their scores inflated. Indeed, we estimate that between 14 and 25 percent of students in the suspect region had their scores inflated (Table 2).

Test score inflation appears to be least prevalent in the Integrated Algebra exam, where we estimate that only 2.5 percent of students had inflated scores (14.1 percent of students in the suspect region). This may be due to the fact that a greater fraction of the exam consists of multiple-choice items, or it is more difficult to inflate scores on open-response math problems than on English or History essays. It also may be due to the fact that NYS guidelines mandate an automatic rescoring of all math and science Regents exams scoring between 60 and 64, and allows principals discretion to rescore math and science exams scoring between 50 and 54. The prevalence of score inflation is greatest is U.S. History and Government and Global History and Geography, in which 4.7 and 5.2 percent of exams appear to have inflated scores.

It is also possible to estimate the number of student scores affected at each of the different passing thresholds. Rows 6 and 7 in Table 3 show respectively the number and fraction of students who scored 65 or higher on each exam that we estimate actually should have scored below 65. The numbers range from roughly 4,200 to nearly 8,200 students. To be specific, these numbers are calculated as the difference between the actual frequency of students scoring at or above 65 within the suspect range minus the number of students we predict should have attained these same scores (as shown by the heavy dashed line in Figure 2). For example, for the English Language Arts exam, the number of students we estimate to have inappropriately received scores at or above the 65 threshold is 4,295, which we calculate as the sum of the vertical distances between the open dots at 65 and 67 and the heavy dashed line at these same points. To get the corresponding fraction of 3.9 percent, we divide this number by the sum of all of the actual frequencies associated with scores 65 or higher – that is, the sum of all of the open dots from 65 through 100.

On the Global History and Geography exam, we estimate that 5.3 percent of all students scoring 65 or higher should have scored below this mark. Of course, this includes many extremely high-achieving students – e.g., all students scoring above 90 on the exam. Among students who scored 65 or higher who were close to this passing threshold, we estimate that 38.8 percent should have scored below 65. Similarly, on the Integrated Algebra exam, we estimate that 2.8 percent of all students scoring 65 or higher should have scored below this mark. Among students scoring at or above 65 but in the suspect region, we estimate that 51.1 percent had inflated scores that pushed them over the passing threshold inappropriately. The number of students estimated to have been inappropriately pushed over the 55 point threshold are substantially lower, but still significant.

In Tables 4 and 5, we replicate these estimates using only exam scores from high schools in New York City and schools outside of New York City respectively. These results suggest that the extent of manipulation of Regents scores was roughly twice as large in New York City as in other areas of the state. For example, the fraction of exam scores predicted to have been inflated (row 4 of Table 4) ranges from 4.0 (Integrated Algebra) to 8.3 (U.S. History and Government). The fraction of students scoring at or above 65 predicted to score below this passing threshold (row 7 of Table 4) ranges from 5.4 to 10.8 percent.

6. Conclusions

The state of New York has been in 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, end-of-course tests tied to the state's learning standards in five core academic subjects. Student scores on these exams also have high stakes for schools and teachers because these tests determine how schools are evaluated under the state school-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 grading 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 grading 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 implications of this manipulation for the overall number of students designated as eligible to graduate are more modest but still non-trivial. More specifically, our estimates imply that 3 to 5 percent of the exam scores that met the requirement for high-school graduation actually fell below the state's performance threshold. Interestingly, we find that this pattern is noticeably more common in New York City's high schools than in high schools located in the rest of the state.

It should be noted that we explicitly do not characterize the grading manipulation documented here as definitive evidence of teacher *cheating*. This is partly because the intent and awareness of teachers whose grading moves students across a performance threshold are not clear. The state-level design of the Regents grading procedures may also encourage exactly the sort of scoring manipulation documented here. For example, suppose that, when teachers are rescoring an exam, there is unconscious bias towards a "hold harmless" approach (i.e., to only revise grades upward). In this scenario, the compulsory rescoring of math and science scores just below the performance thresholds may contribute to the discontinuities documented here.

Furthermore, 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, these results are also likely to be viewed as policy relevant because they indicate that the state's current grading procedures and the resulting manipulation of marginal scores are undermining the intent of the learning standards embodied in the Regents Examinations. This manipulation may also raise substantive issues of fairness (i.e., "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 ongoing research is examining this issue by evaluating the extent of the school-level variation in grading manipulation as well as its determinants.

Regardless, our results do suggest that there may be attractive, sensible reforms to New York's grading 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 grading 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 grading 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 grading responsibilities from schools to a central office.[6] 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.

[6] 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.

Table 1 – Regents Exam Requirements by Diploma Type and Cohort

Cohorts by 9th Grade Entry	Local Diploma	Regents Diploma
Fall 2002-2004	55 or higher in 5 subjects	65 or higher in 5 subjects
Fall 2005	65 or higher in 2 subjects, 55-64 in 3 subjects	65 or higher in 5 subjects
Fall 2006	65 or higher in 3 subjects, 55-64 in 2 subjects	65 or higher in 5 subjects
Fall 2007	65 or higher in 4 subjects, 55-64 in 1 subject	65 or higher in 5 subjects
Fall 2008	Not Available	65 or higher in 5 subjects

Notes: The five core Regents-Examination subjects are English, Mathematics, Science, U.S. History and Government, Global History and Geography. An advanced-designation Regents diploma is available to students in who pass Regents Examinations in additional subjects. Students with disabilities can earn a local diploma over this period with a score of 55 on at least 1 Regents Examination.

Table 2 – Distribution of Scale Scores on the June 2009 NYS Regents Examinations

	U.S. Hist & Govt	Global Hist & Geo	English Language Arts	Integrated Algebra	Living Environ. (Science)
Panel A: Performance Threshold at 55					
Number of students scoring just below cutoff	825	1,612	1,187	3,263	1,100
Number of students scoring exactly at cutoff	3,354	4,485	2,607	4,651	3,132
Difference	2,529	2,873	1,420	1,388	2,032
P-value of difference	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001
Panel B: Performance Threshold at 65					
Number of students scoring just below cutoff	395	1,016	1,315	1,100	1,227
Number of students scoring exactly at cutoff	6,412	7,687	6,547	8,962	8,216
Difference	6,017	6,671	5,232	7,862	6,989
P-value of difference	<0.00001	<0.00001	<0.00001	<0.00001	<0.00001

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

Table 3 – Estimated Magnitudes of Testing Inaccuracies on the June 2009 NYS Regents Examinations

	U.S. Hist & Govt	Global Hist & Geo	English Language Arts	Integrated Algebra	Living Environ. (Science)
(1) Total number of tested students	186,245	207,489	136,591	222,941	206,291
(2) Number of students with scores near performance thresholds	34,234	46,542	26,048	40,142	29,779
(3) Number of students near performance thresholds predicted to have inflated exam scores	8,691	10,826	5,239	5,679	6,674
(4) Fraction of all students predicted to have inflated exam scores	0.047	0.052	0.038	0.025	0.032
(5) Fraction of students near performance thresholds predicted to have inflated exam scores	0.254	0.233	0.201	0.141	0.224
(6) Number of students who scored 65 or higher predicted to have a scored below 65	6,548	8,179	4,295	4,576	4,925
(7) Fraction of students who scored 65 or higher predicted to have scored below 65	0.043	0.053	0.039	0.028	0.028
(8) Fraction of students who scored 65 or higher (but still near the performance threshold) predicted to have scored below 65	0.411	0.388	0.341	0.511	0.599
(9) Number of students who scored 55 or higher predicted to have scored below 55	2,558	2,936	983	854	492
(10) Fraction of students who scored 55 or higher predicted to have scored below 55	0.015	0.017	0.008	0.005	0.003

Notes: Students scoring near the performance thresholds are defined as follows: ELA (scores above 49 and below 69); Global History (scores above 50 and below 69); Integrated Algebra (scores above 51 and below 66); Living Environment (scores above 50 and below 66); U.S. History (scores above 47 and below 69).

Table 4 – Estimated Magnitudes of Testing Inaccuracies on the June 2009 NYS Regents Examinations: New York City High Schools

	U.S. Hist & Govt	Global Hist & Geo	English Language Arts	Integrated Algebra	Living Environ. (Science)
(1) Total number of tested students	60,986	73,516	48,015	79,998	71,091
(2) Number of students with scores near performance thresholds	18,311	22,115	14,787	20,767	17,892
(3) Number of students near performance thresholds predicted to have inflated exam scores	5,072	5,956	3,560	3,190	4,300
(4) Fraction of all students predicted to have inflated exam scores	0.083	0.081	0.074	0.040	0.060
(5) Fraction of students near performance thresholds predicted to have inflated exam scores	0.277	0.269	0.241	0.154	0.240
(6) Number of students who scored 65 or higher predicted to have a scored below 65	3,736	4,441	3,315	2,543	2,989
(7) Fraction of students who scored 65 or higher predicted to have scored below 65	0.092	0.101	0.108	0.054	0.059
(8) Fraction of students who scored 65 or higher (but still near the performance threshold) predicted to have scored below 65	0.466	0.464	0.474	0.549	0.621
(9) Number of students who scored 55 or higher predicted to have scored below 55	1,726	2,561	1,140	511	442
(10) Fraction of students who scored 55 or higher predicted to have scored below 55	0.036	0.047	0.031	0.009	0.007

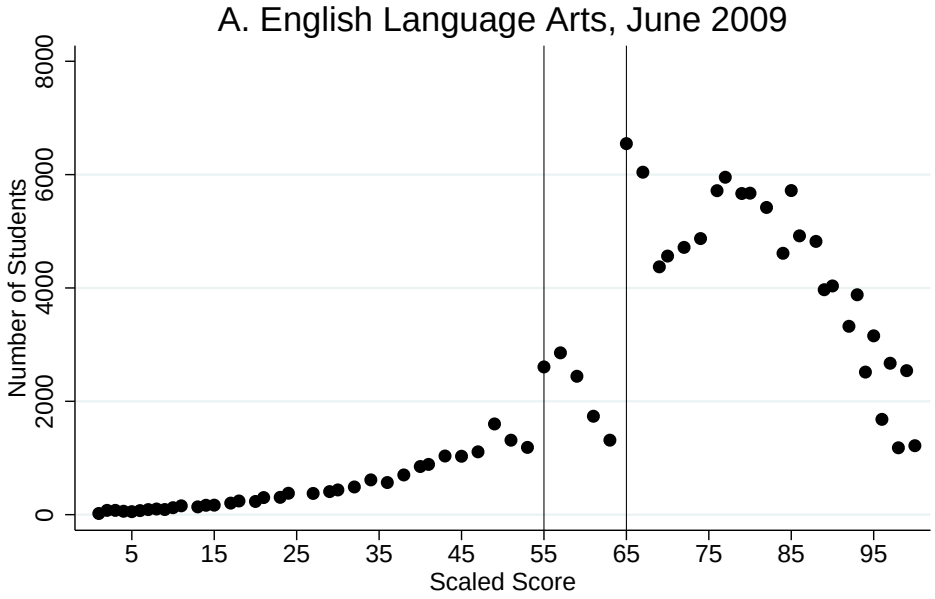
Notes: Students scoring near the performance thresholds are defined as follows: ELA (scores above 49 and below 69); Global History (scores above 50 and below 69); Integrated Algebra (scores above 51 and below 66); Living Environment (scores above 50 and below 66); U.S. History (scores above 47 and below 69).

Table 5 – Estimated Magnitudes of Testing Inaccuracies on the June 2009 NYS Regents Examinations: Schools Outside of New York City

	U.S. Hist & Govt	Global Hist & Geo	English Language Arts	Integrated Algebra	Living Environ. (Science)
(1) Total number of tested students	125,259	133,973	88,576	142,943	135,200
(2) Number of students with scores near performance thresholds	15,923	24,427	11,261	19,375	11,887
(3) Number of students near performance thresholds predicted to have inflated exam scores	3,595	5,047	1,679	2,685	2,289
(4) Fraction of all students predicted to have inflated exam scores	0.029	0.038	0.019	0.019	0.017
(5) Fraction of students near performance thresholds predicted to have inflated exam scores	0.226	0.207	0.149	0.139	0.193
(6) Number of students who scored 65 or higher predicted to have a scored below 65	2,758	3,836	1,123	2,075	1,952
(7) Fraction of students who scored 65 or higher predicted to have scored below 65	0.024	0.035	0.014	0.018	0.016
(8) Fraction of students who scored 65 or higher (but still near the performance threshold) predicted to have scored below 65	0.348	0.333	0.201	0.479	0.574
(9) Number of students who scored 55 or higher predicted to have scored below 55	728	901	187	311	107
(10) Fraction of students who scored 55 or higher predicted to have scored below 55	0.006	0.008	0.002	0.002	0.001

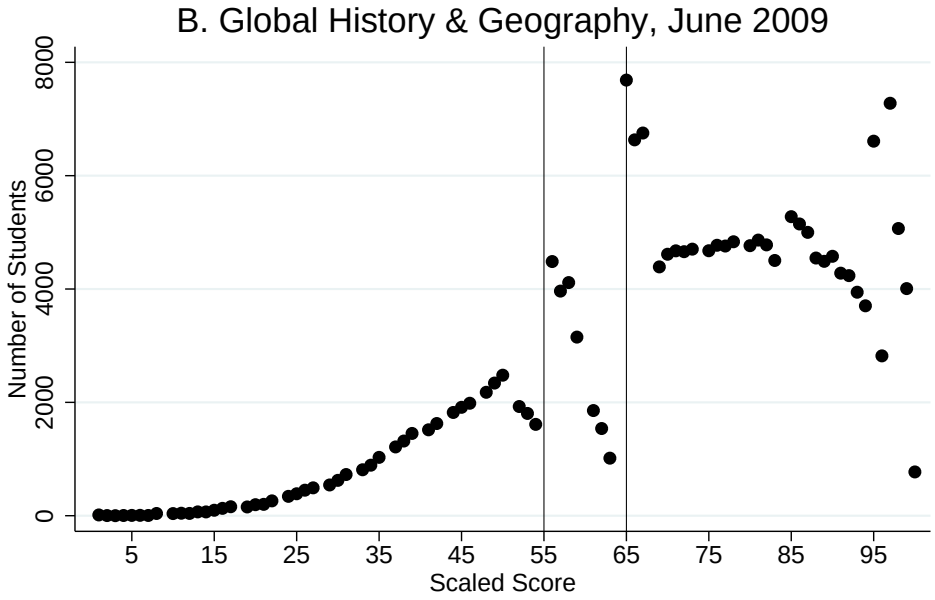
Notes: Students scoring near the performance thresholds are defined as follows: ELA (scores above 49 and below 69); Global History (scores above 50 and below 69); Integrated Algebra (scores above 51 and below 66); Living Environment (scores above 50 and below 66); U.S. History (scores above 47 and below 69).

FIGURE 1. DISTRIBUTION OF STUDENT SCORES ON REGENTS EXAM



Source: Author calculations based on data from New York Regents.

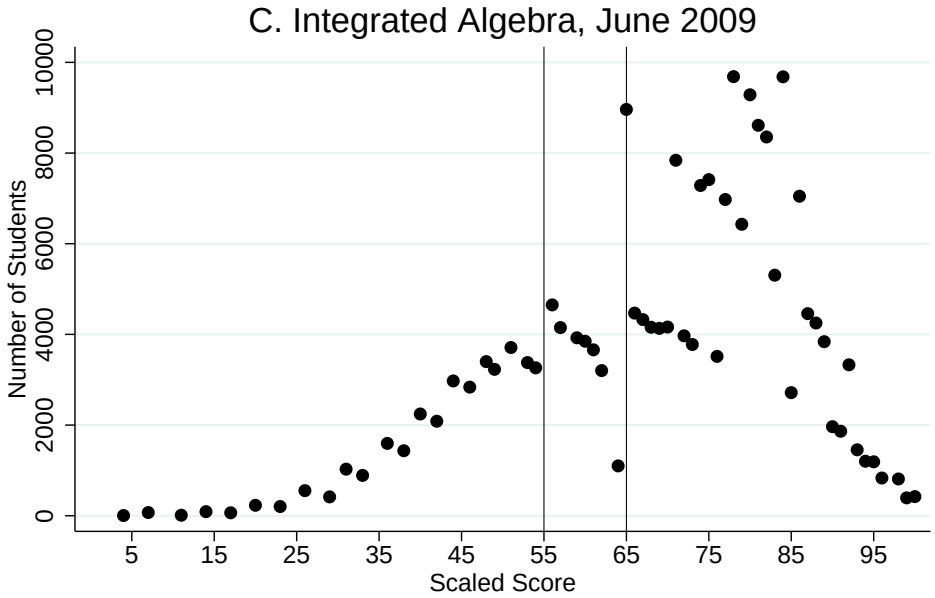
Notes: Graph shows number of students attaining each scaled score.



Source: Author calculations based on data from New York Regents.

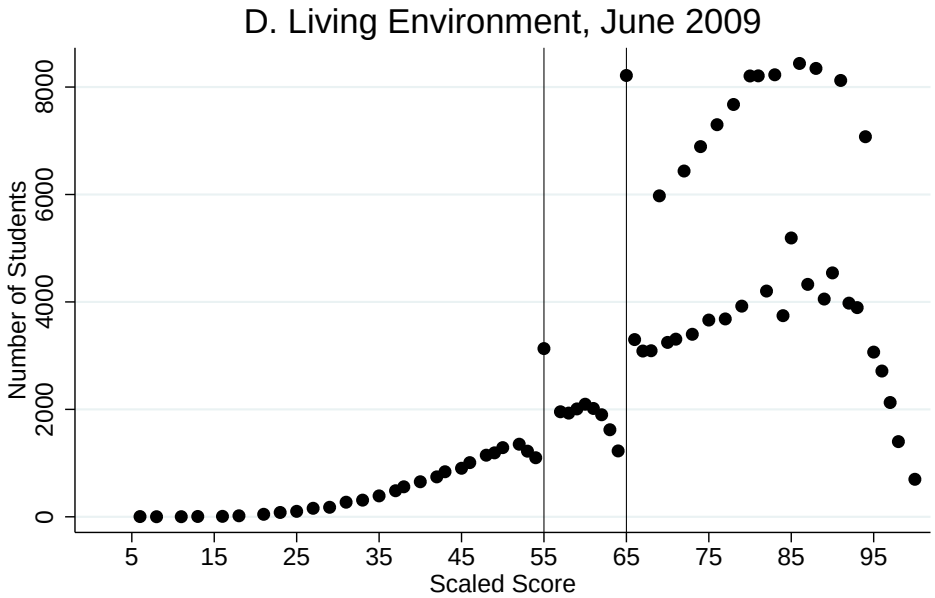
Notes: Graph shows number of students attaining each scaled score.

FIGURE 1 (CONTINUED). DISTRIBUTION OF STUDENT SCORES ON REGENTS EXAM



Source: Author calculations based on data from New York Regents.

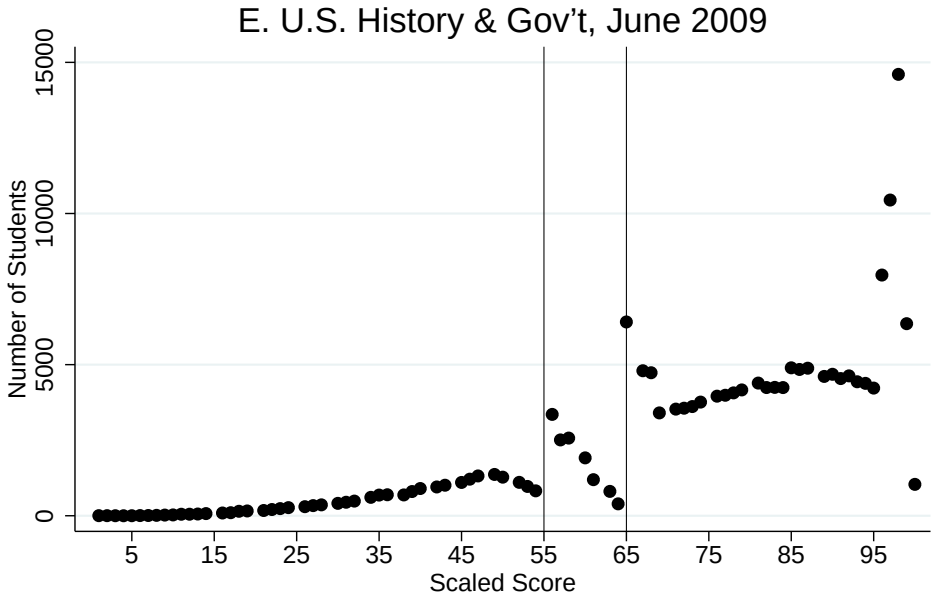
Notes: Graph shows number of students attaining each scaled score.



Source: Author calculations based on data from New York Regents.

Notes: Graph shows number of students attaining each scaled score.

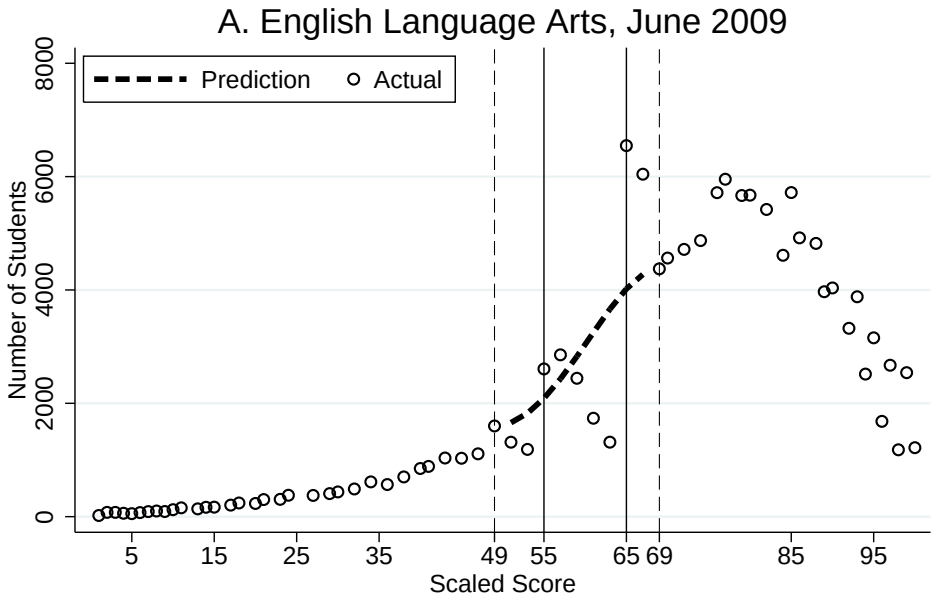
FIGURE 1 (CONTINUED). DISTRIBUTION OF STUDENT SCORES ON REGENTS EXAM



Source: Author calculations based on data from New York Regents.

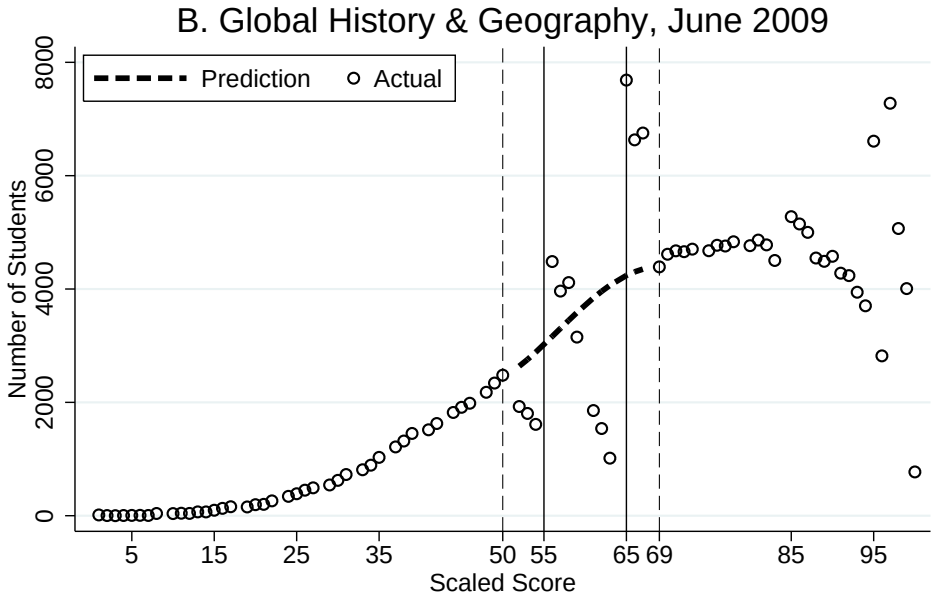
Notes: Graph shows number of students attaining each scaled score.

FIGURE 2. WHAT STUDENT SCORES PROBABLY WERE



Source: Author calculations based on data from New York Regents.

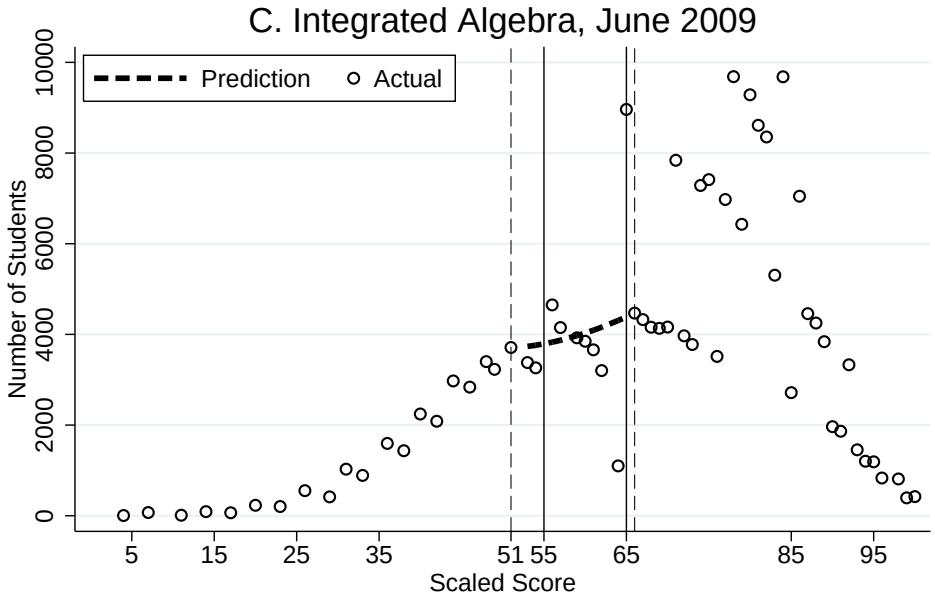
Notes: Dashed curve interpolates between 49 and 69. See text for details.



Source: Author calculations based on data from New York Regents.

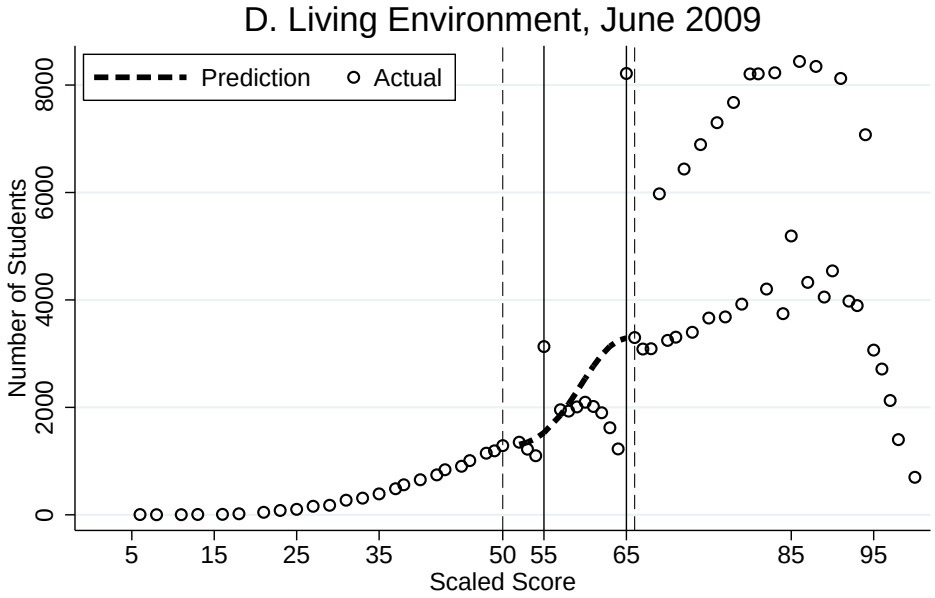
Notes: Dashed curve interpolates between 50 and 69. See text for details.

FIGURE 2 (CONTINUED). WHAT STUDENT SCORES PROBABLY WERE



Source: Author calculations based on data from New York Regents.

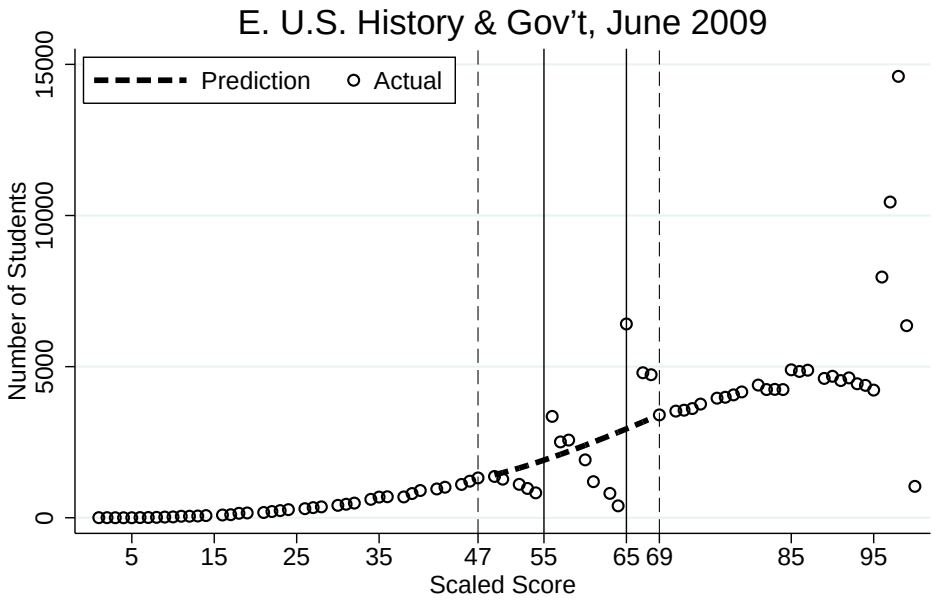
Notes: Dashed curve interpolates between 51 and 66. See text for details.



Source: Author calculations based on data from New York Regents.

Notes: Dashed curve interpolates between 50 and 66. See text for details.

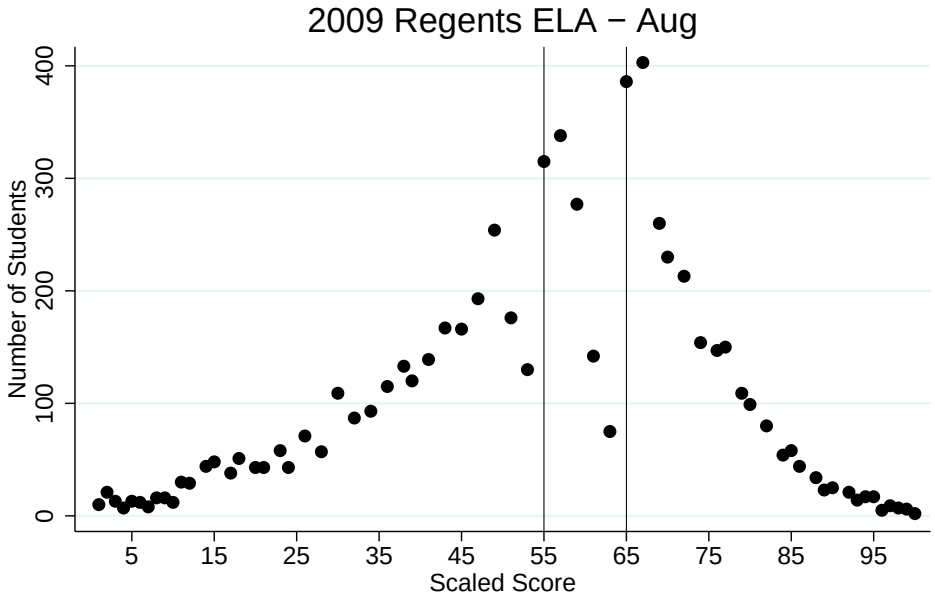
FIGURE 2 (CONTINUED). WHAT STUDENT SCORES PROBABLY WERE



Source: Author calculations based on data from New York Regents.

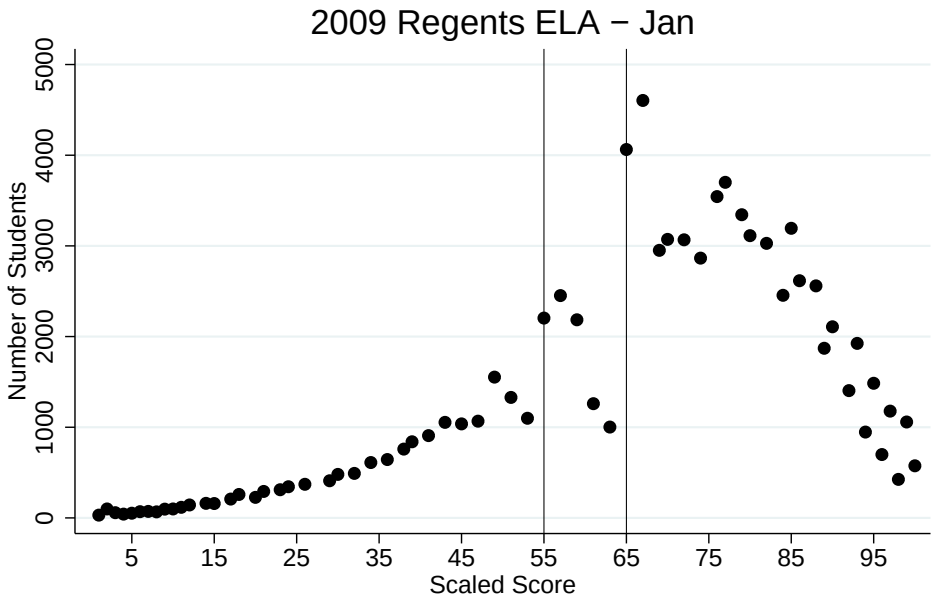
Notes: Dashed curve interpolates between 47 and 69. See text for details.

Appendix Figure 1. Distribution of Student Scores, Other Exams



Source: Author calculations based on data from New York Regents.

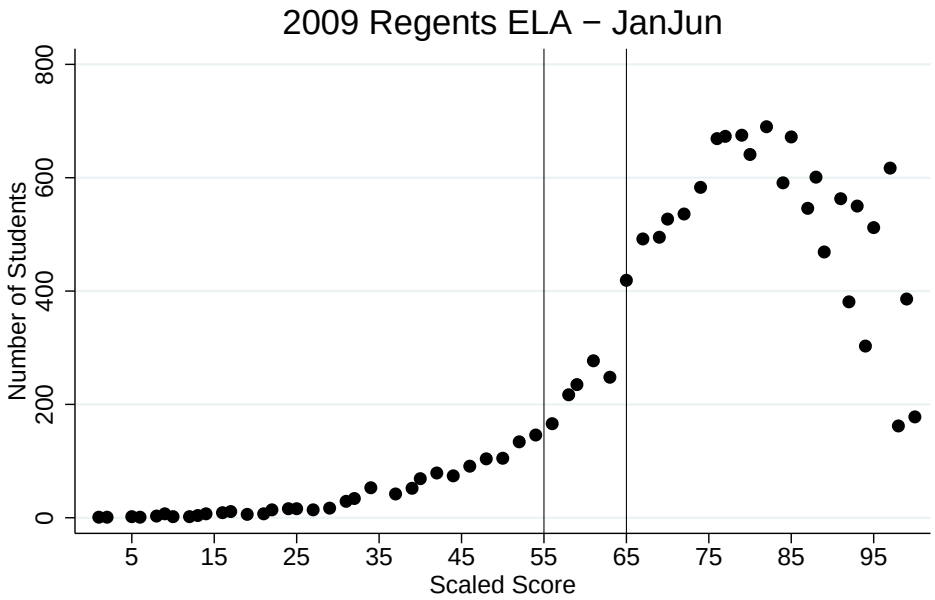
Notes: Graph shows number of students attaining each scaled score.



Source: Author calculations based on data from New York Regents.

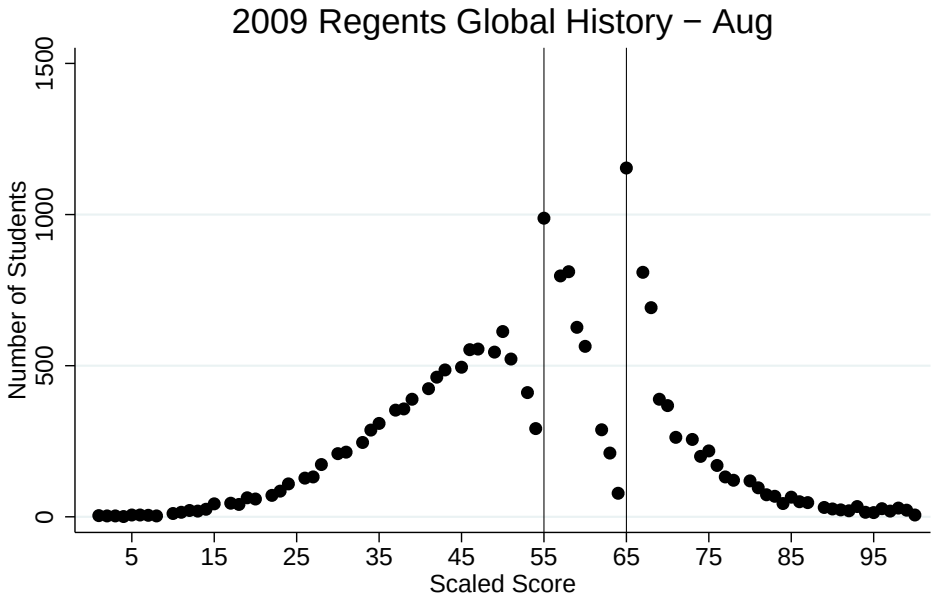
Notes: Graph shows number of students attaining each scaled score.

Appendix Figure 1 (cont'd). Distribution of Student Scores, Other Exams



Source: Author calculations based on data from New York Regents.

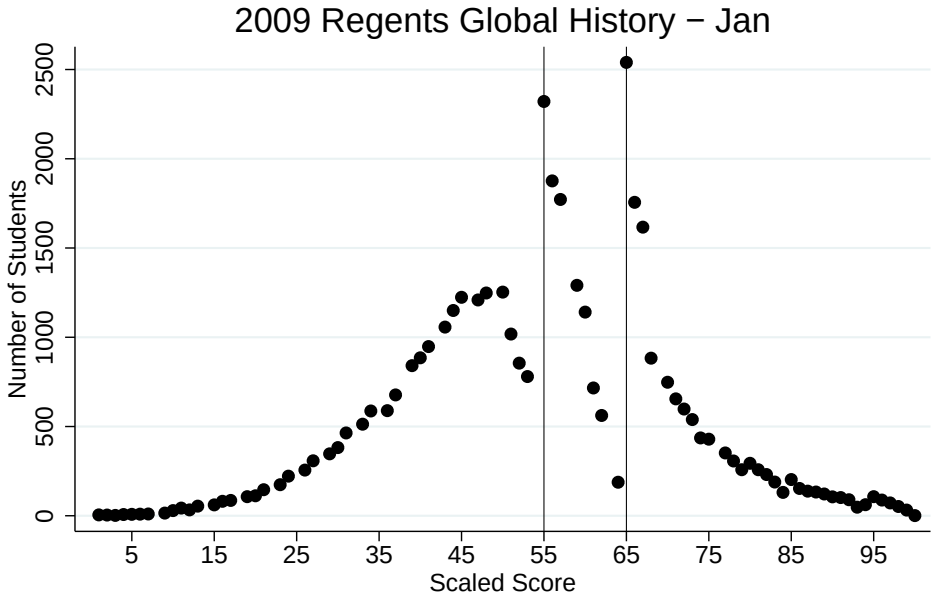
Notes: Graph shows number of students attaining each scaled score.



Source: Author calculations based on data from New York Regents.

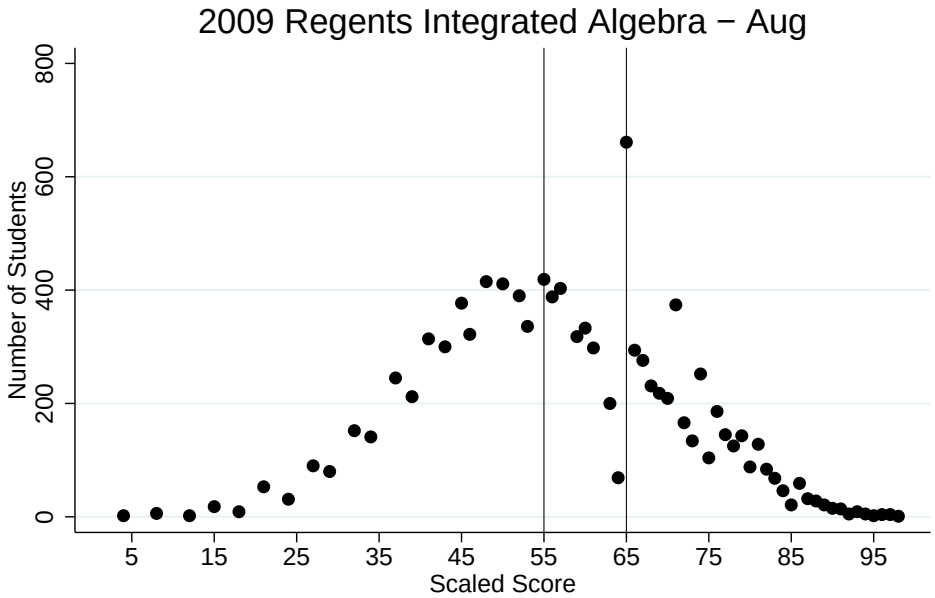
Notes: Graph shows number of students attaining each scaled score.

Appendix Figure 1 (cont'd). Distribution of Student Scores, Other Exams



Source: Author calculations based on data from New York Regents.

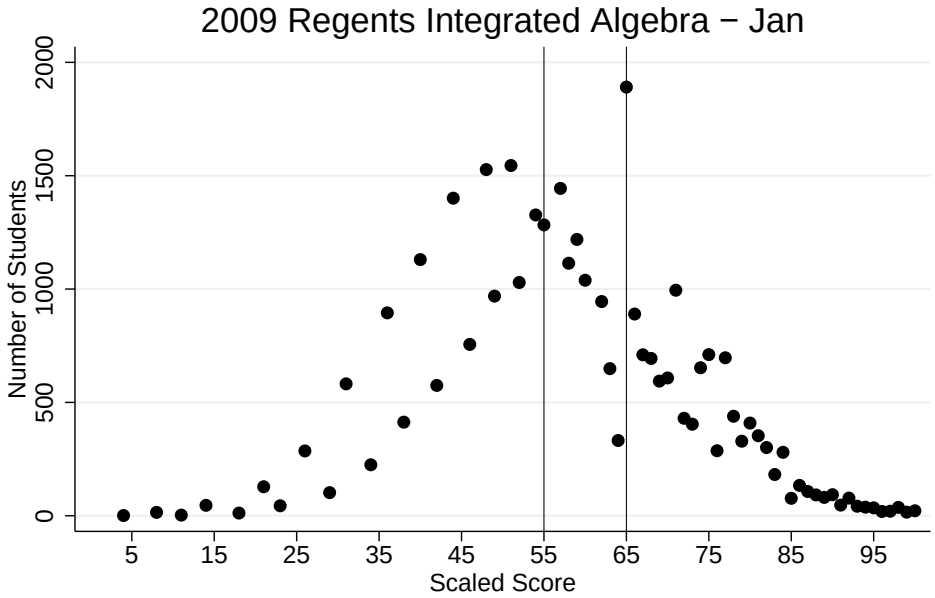
Notes: Graph shows number of students attaining each scaled score.



Source: Author calculations based on data from New York Regents.

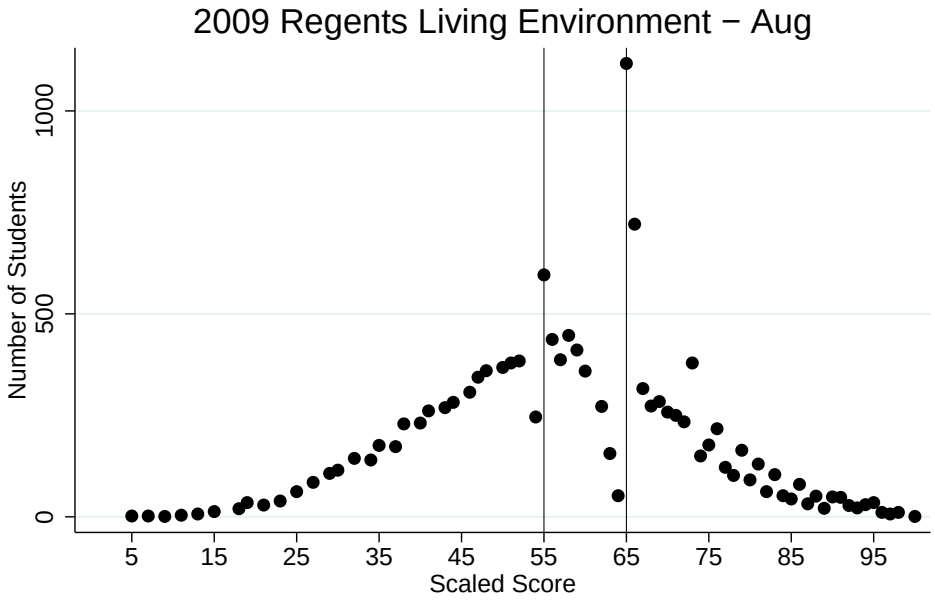
Notes: Graph shows number of students attaining each scaled score.

Appendix Figure 1 (cont'd). Distribution of Student Scores, Other Exams



Source: Author calculations based on data from New York Regents.

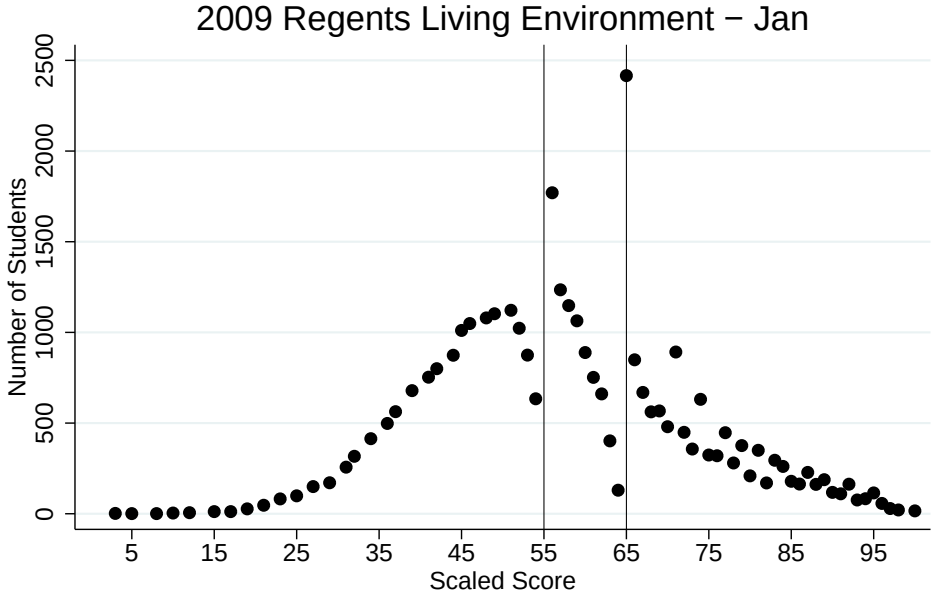
Notes: Graph shows number of students attaining each scaled score.



Source: Author calculations based on data from New York Regents.

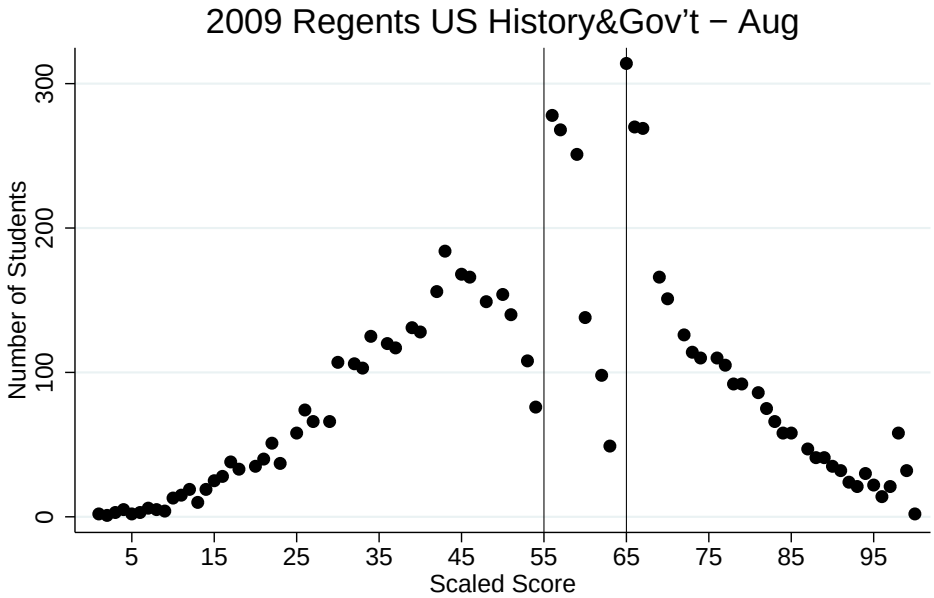
Notes: Graph shows number of students attaining each scaled score.

Appendix Figure 1 (cont'd). Distribution of Student Scores, Other Exams



Source: Author calculations based on data from New York Regents.

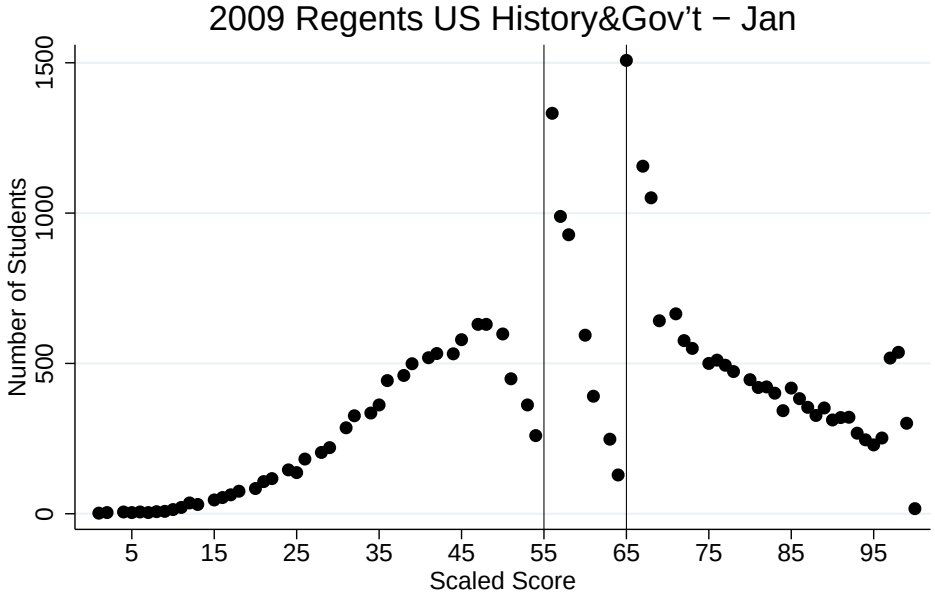
Notes: Graph shows number of students attaining each scaled score.



Source: Author calculations based on data from New York Regents.

Notes: Graph shows number of students attaining each scaled score.

Appendix Figure 1 (cont'd). Distribution of Student Scores, Other Exams



Source: Author calculations based on data from New York Regents.

Notes: Graph shows number of students attaining each scaled score.