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

Online Appendix

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

Appendix A: Additional Results
[NOT FOR PUBLICATION]

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

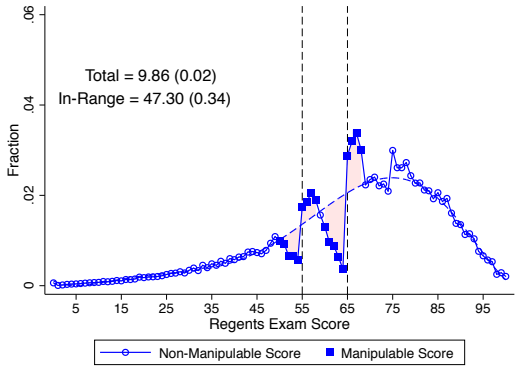
June 2009 English Exam -- Manipulable Scores Shown in Bold

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

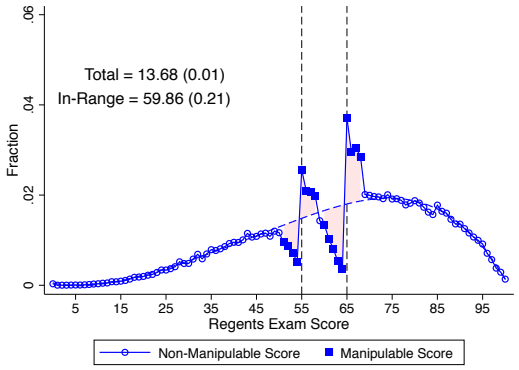
Note: This figure displays the official conversion chart for the English Language Arts Regents Exam for June 2009. For expositional purposes, the scale scores corresponding with essay points 2-14 and 20-23, and those corresponding with 19-24 multiple choice items correct, are omitted and represented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

Appendix Figure A2: Results by Subject, 2004-2010

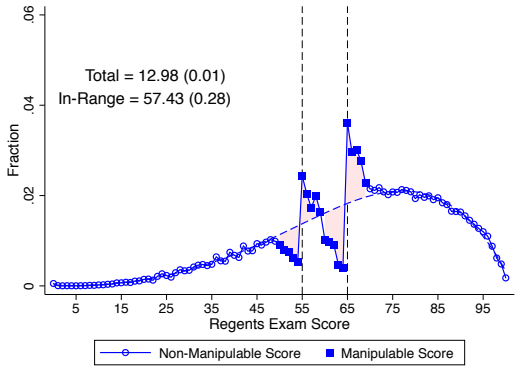
(A) English Language Arts, 2004-2010



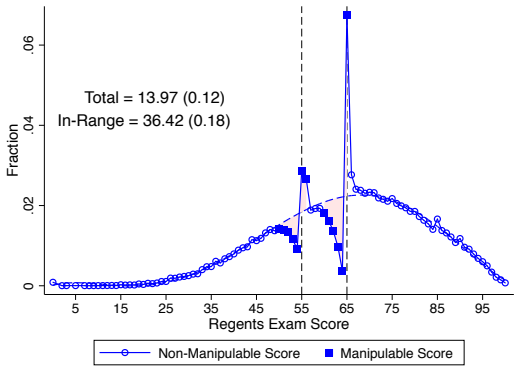
(B) Global History, 2004-2010



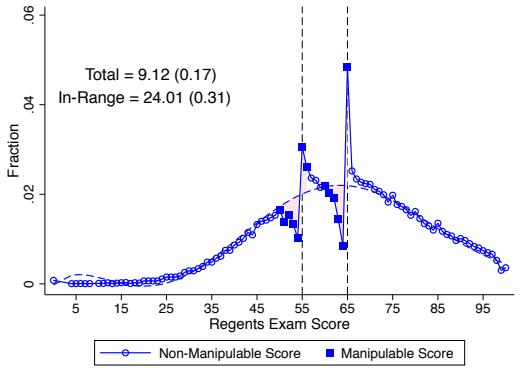
(C) U.S. History, 2004-2010



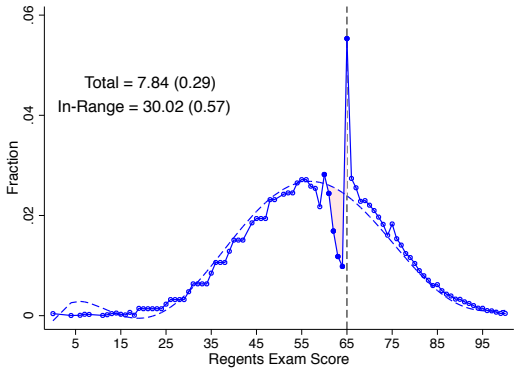
(D) Living Environment, 2004-2010



(E) Math A, 2004-2008



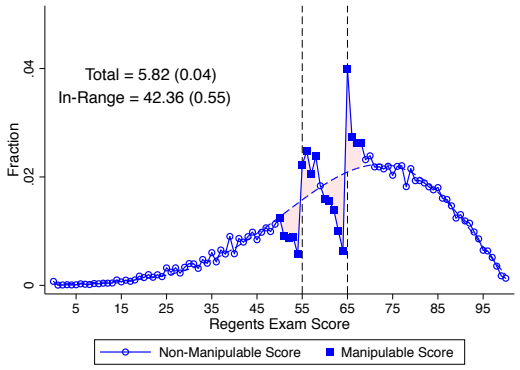
(f) Integrated Algebra, 2008-2010



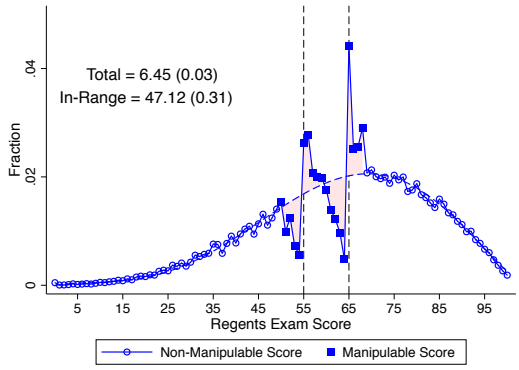
Note: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject x year specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near each cutoff. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III and detailed in Appendix Table A3. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

Appendix Figure A3: Results by Year

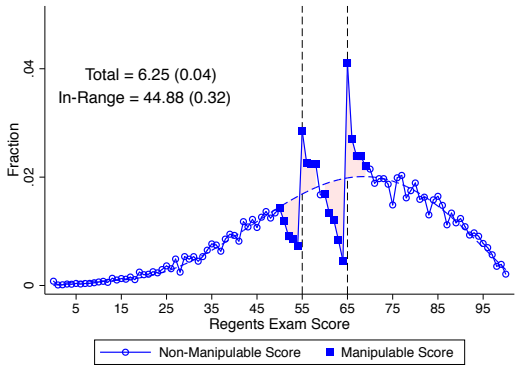
(A) 2004 Core Exams



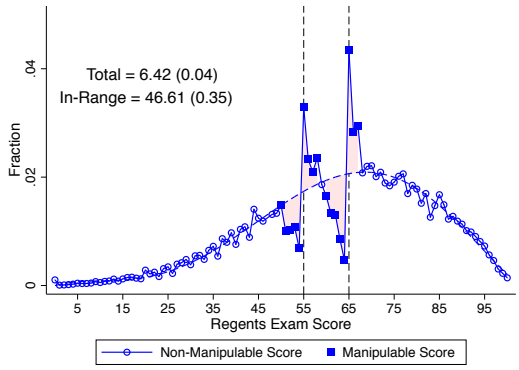
(B) 2005 Core Exams



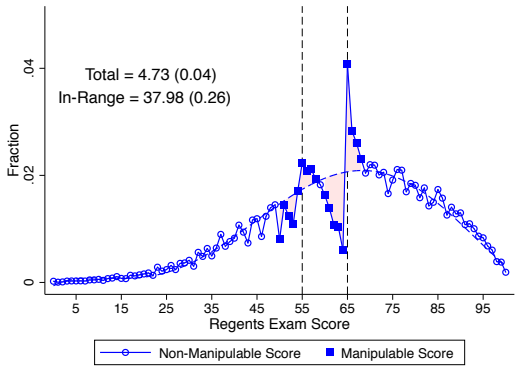
(C) 2006 Core Exams



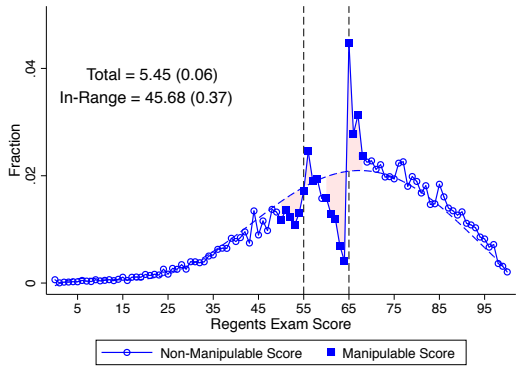
(D) 2007 Core Exams



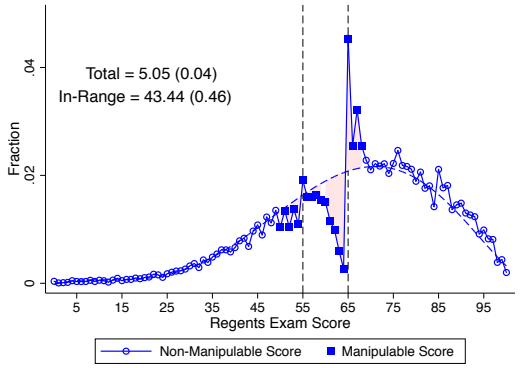
(E) 2008 Core Exams



(F) 2009 Core Exams

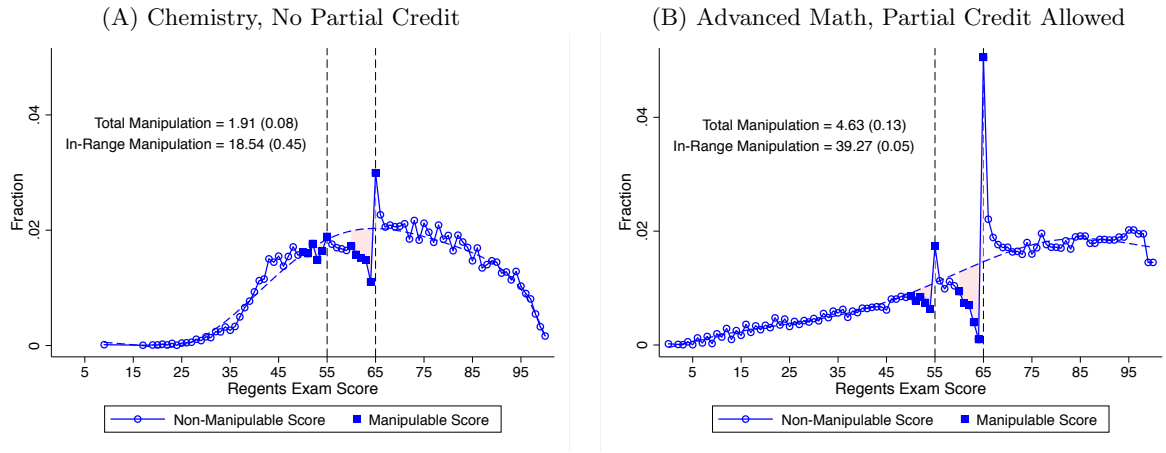


(G) 2010 Core Exams



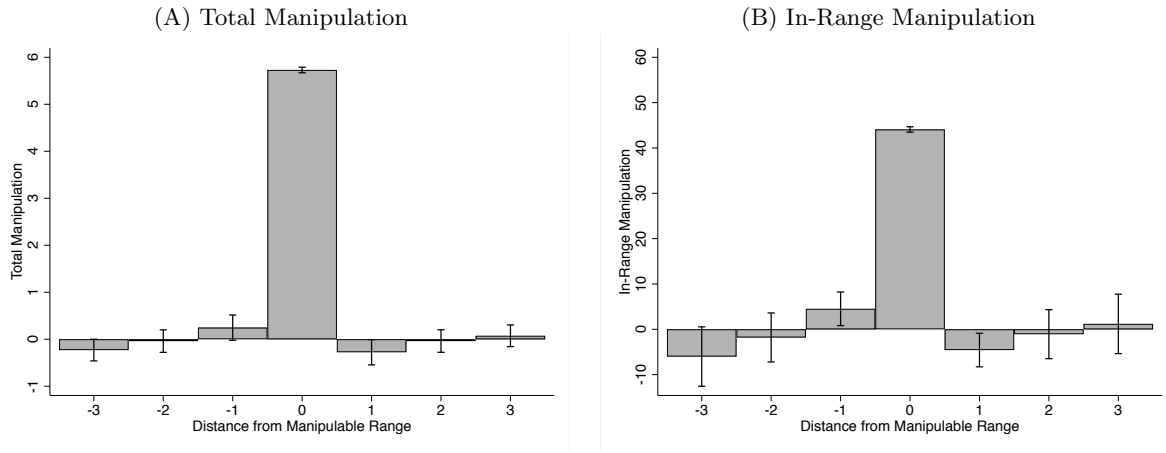
Note: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject x year specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near each cutoff. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III and detailed in Appendix Table A3. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

Appendix Figure A4: Results for June 2001 Elective Exams With and Without Partial Credit



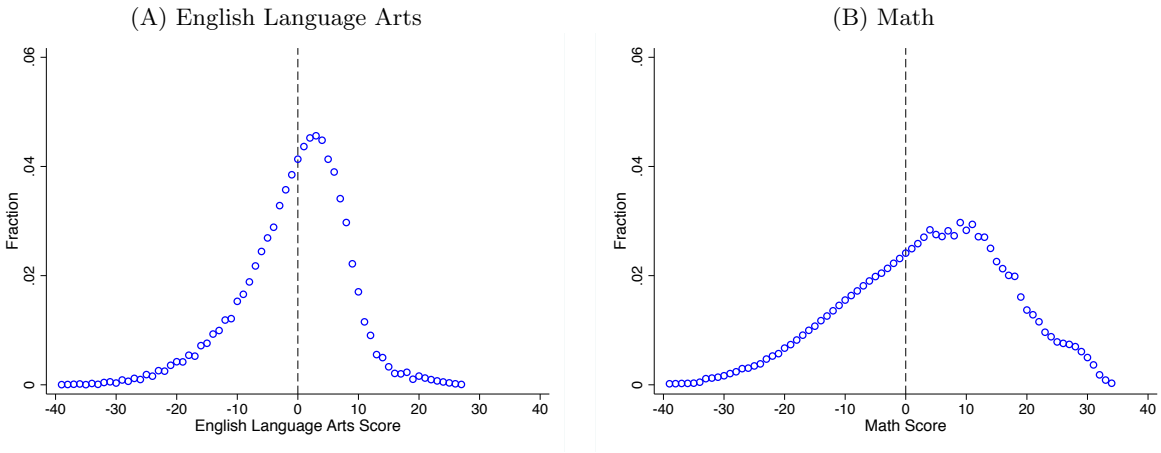
Note: These figures show the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers in June 2001. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject x year specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near each cutoff. The shaded area represents either the missing or excess mass for manipulable scores as we define based on the scoring guidelines described in Section III and detailed in Appendix Table A3. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

Appendix Figure A5: Estimates of Bunching Just Outside of the Manipulable Range



Note: These figures show estimated manipulation at points below, at, and above our specified manipulable ranges around the 55 and 65 score cutoffs for New York City high school core exams between 2004-2010. Core exams include English Language Arts, Global History, U.S. History, Math A/Integrated Algebra, and Living Environment. We include the first test in each subject for each student in our sample. We estimate the counterfactual distribution of scores using a subject x year specific sixth-degree polynomial fitted to the empirical distribution, excluding scores in the manipulable range and scores within three points of this range near each cutoff. The shaded bars represent either the missing or excess mass for these excluded scores, defined as the difference between the counterfactual and empirical distribution. Distance zero corresponds to scores within the manipulable range. Positive distance denotes scores above the manipulable range, negative distance denotes scores below. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

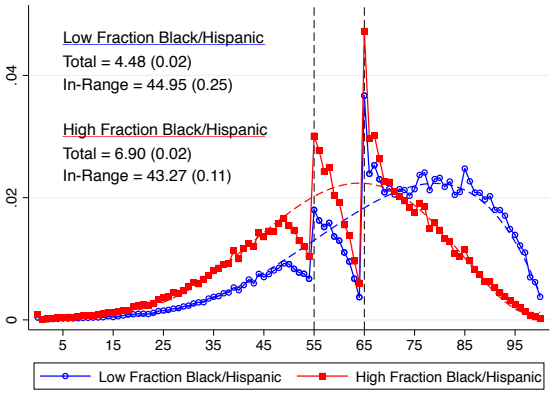
Appendix Figure A6: Test Score Distributions for Centrally Graded Exams in Grades 3-8



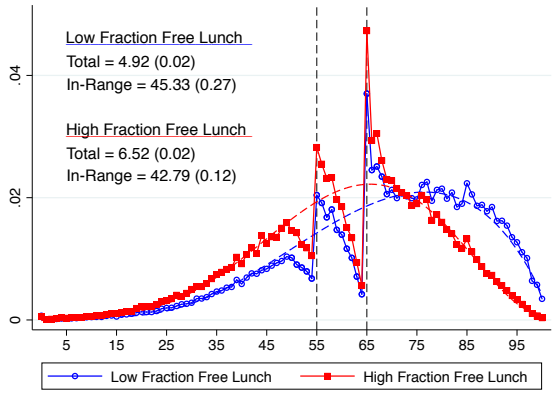
Note: These figures show the test score distribution around the proficiency score cutoff for New York City grade 3-8 test takers between 2004-2010. Each point shows the fraction of test takers in a score bin. See the data appendix for additional details on the variable definitions.

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

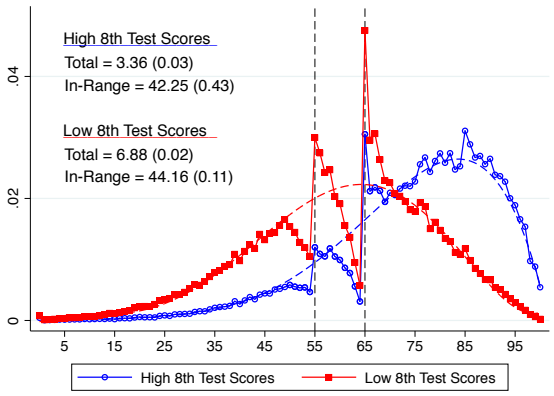
(A) High vs. Low Black/Hispanic Enrollment



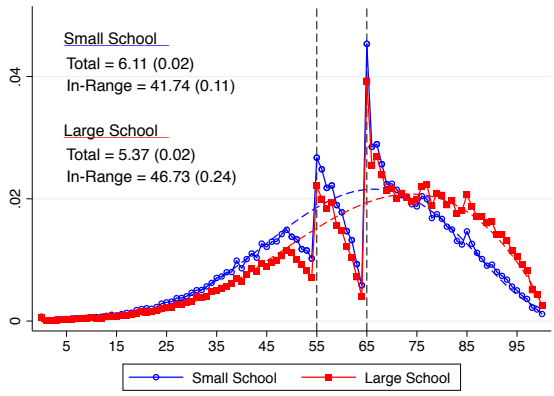
(B) High vs. Low Free Lunch Enrollment



(C) Low vs. High 8th Test Scores



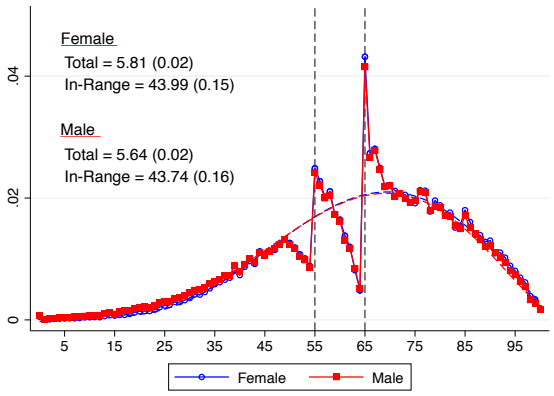
(D) Small vs. Large Schools



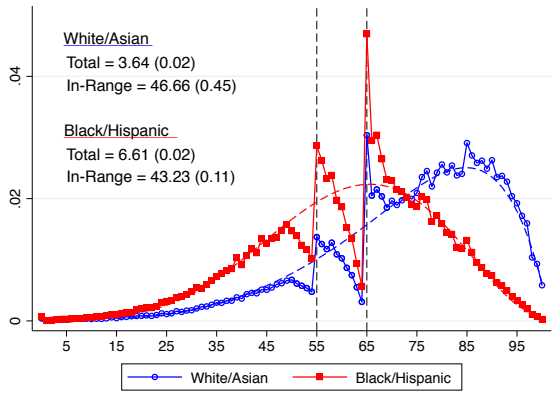
Note: These figures show the test score distribution for core Regents exams around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Panel (A) considers exams taken in schools above and below median in the fraction of black/Hispanic students. Panel (B) considers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

Appendix Figure A8: Results by Student Characteristics, 2004-2010

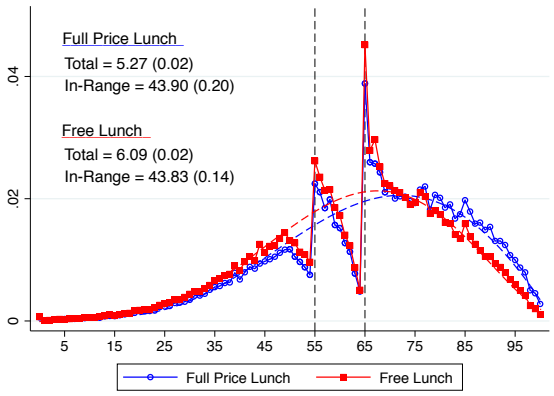
(A) Female vs. Male



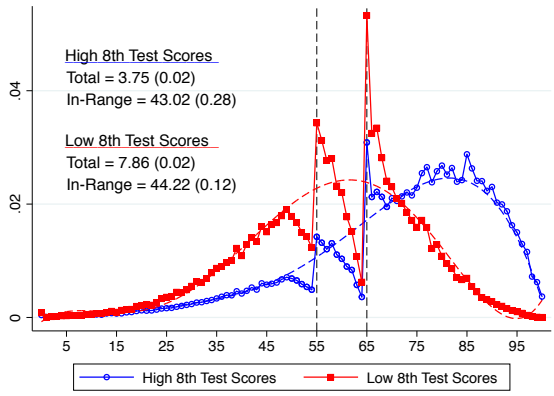
(B) White/Asian vs. Black/Hispanic



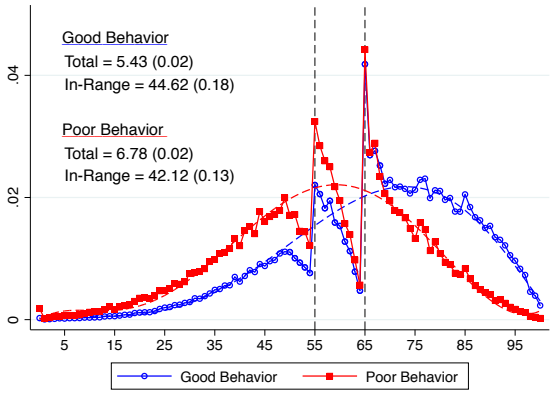
(cC) Full Price vs. Free Lunch



(D) Low vs. High 8th Test Scores

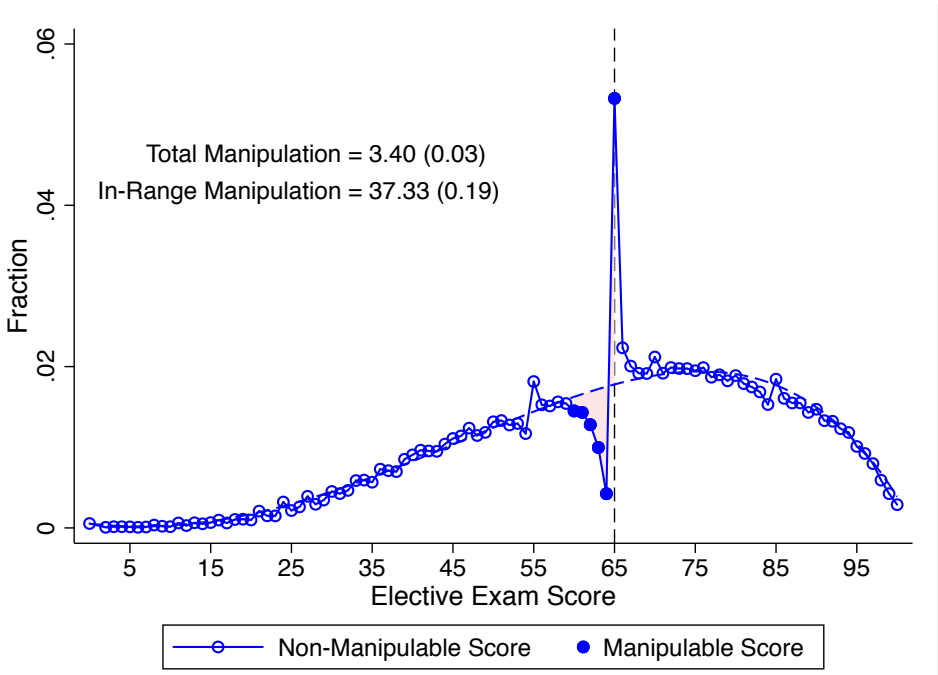


(E) Good vs. Poor Behavior



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

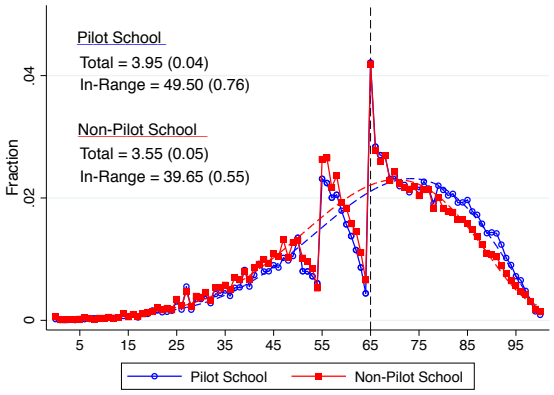
Appendix Figure A9: Results for Elective Regents Exams, 2004-2010



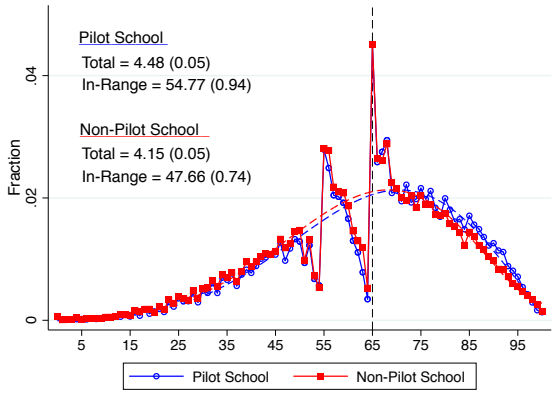
Note: This figure shows the test score distribution around the 65 score cutoff for New York City high school test takers between 2004-2010. Included elective exams include Chemistry, Math B, and Physics. We include the first test in each subject for each student in our sample. Each point shows the fraction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

Appendix Figure A10: Additional Test Score Distributions Before Grading Reforms, 2004-2009

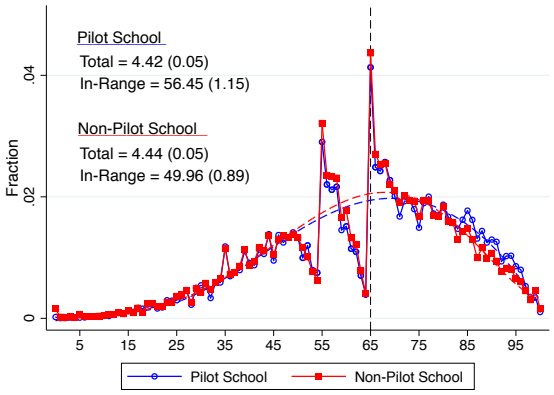
(A) 2004: Re-Scoring and Decentralized Grading



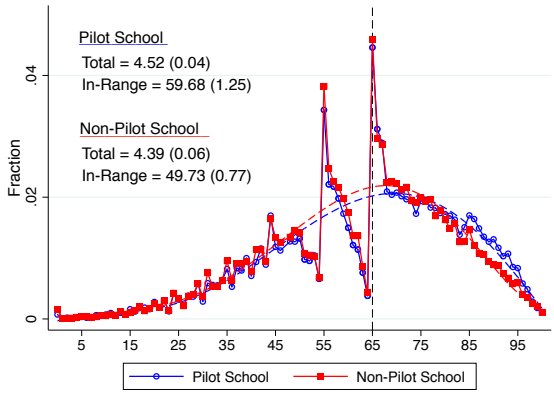
(B) 2005: Re-Scoring and Decentralized Grading



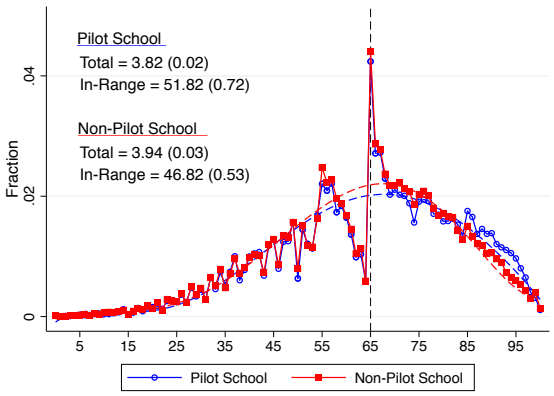
(C) 2006: Re-Scoring and Decentralized Grading



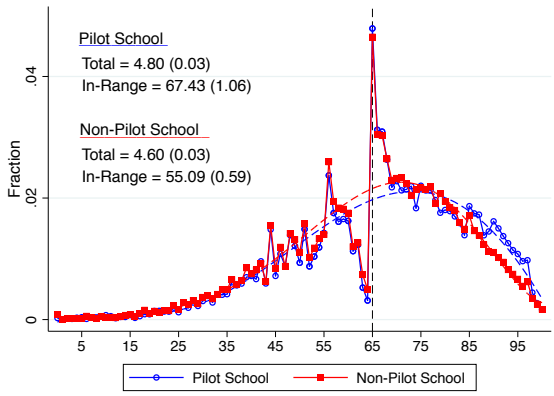
(D) 2007: Re-Scoring and Decentralized Grading



(E) 2008: Re-Scoring and Decentralized Grading

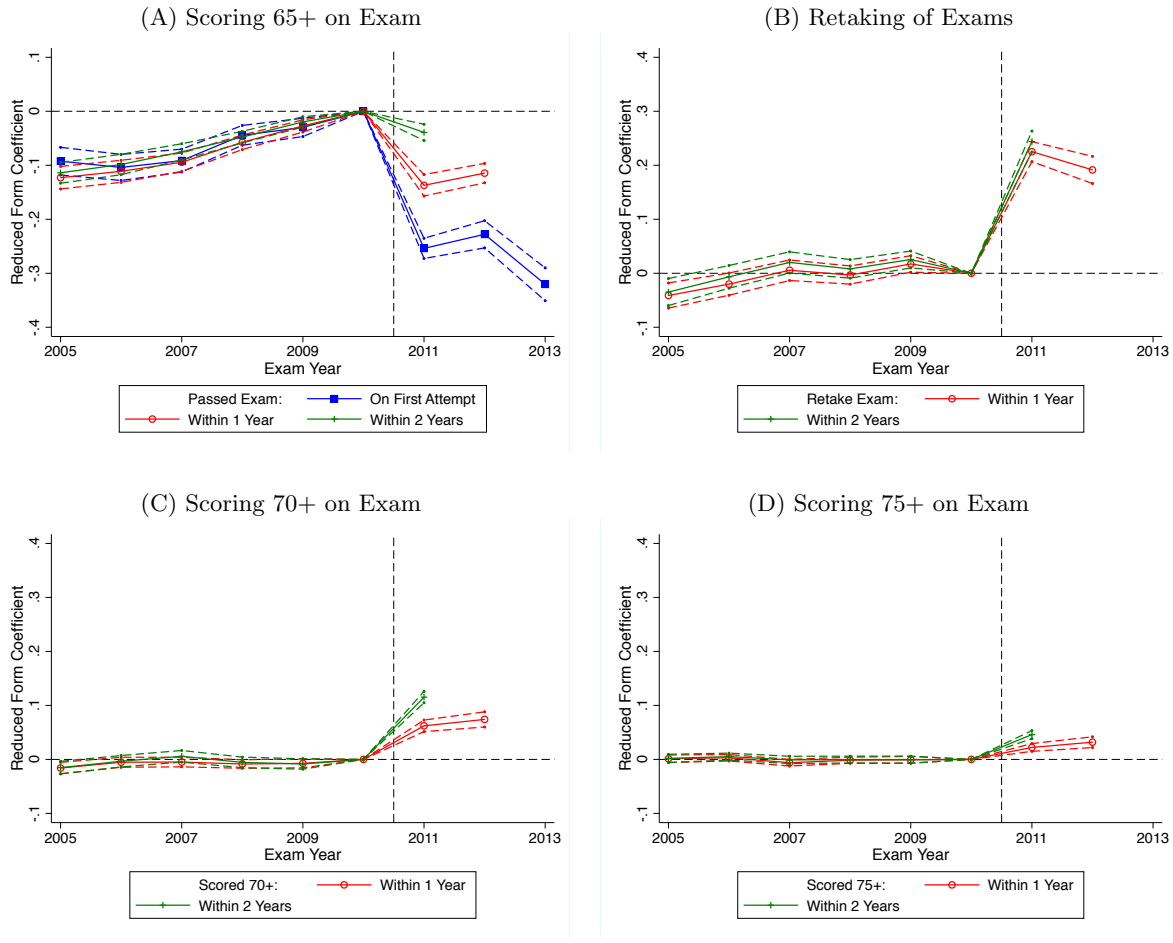


(F) 2009: Re-Scoring and Decentralized Grading



Note: These figures show the test score distribution around the 65 score cutoff for New York City high school test takers between 2004-2009 in June. Included core exams include English Language Arts, Global History, U.S. History, Integrated Algebra, and Living Environment. See the Figure 1 notes for additional details on the sample and empirical specification.

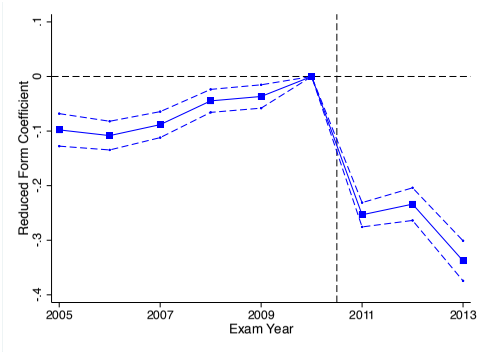
Appendix Figure A11: Regents Outcomes in the Difference-in-Differences Sample



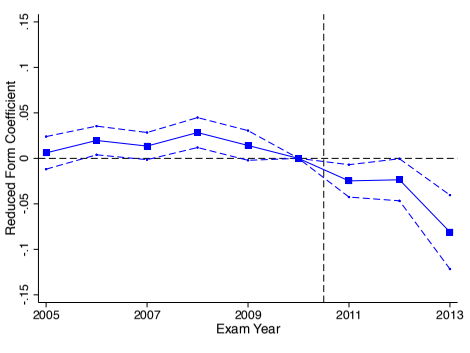
Note: These figures plot the reduced form impact of the Regents grading reforms on Regents outcomes. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. We report reduced form results using the interaction of taking the test in the indicated year and score in the manipulable range around the 65 cutoff. We control for an indicator for scoring between 0-59 in 2011-2013, 10-point scale score effects, and exam x year-of-test effects. We stack student outcomes across the Living Environment, Math A/Algebra, and Global History exams and cluster standard errors at the individual and school levels. See the text for additional details.

Appendix Figure A12: Regents Grading Reforms and Student Outcomes by Student Subgroup

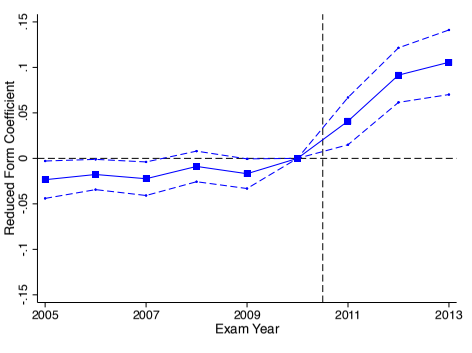
(A) Males: Score 65+ on First Admin.



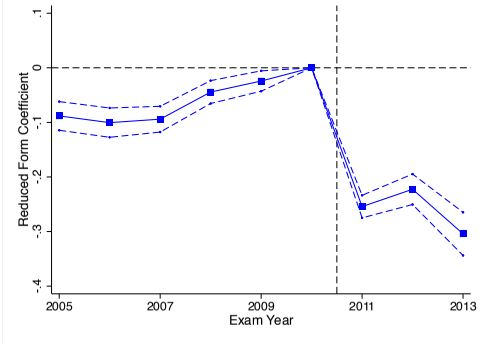
(B) Males: Graduate High School



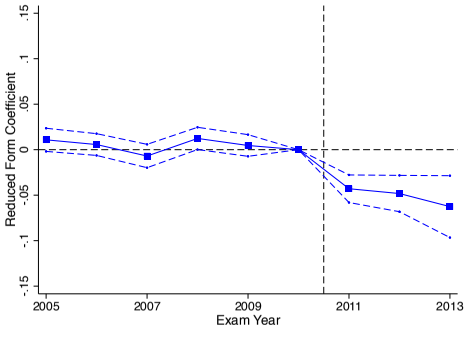
(C) Males: Adv. Regents Req. Taken



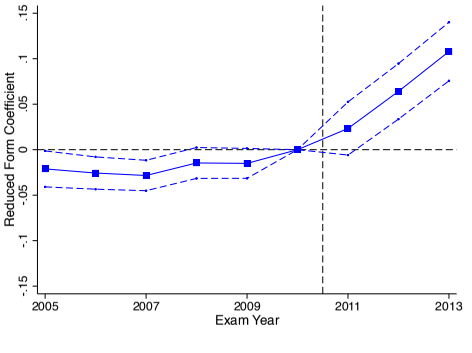
(D) Females: Score 65+ on First Admin.



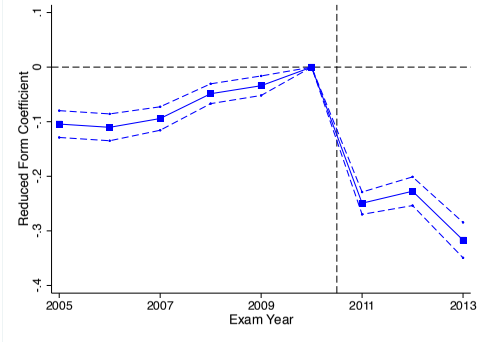
(E) Females: Graduate High School



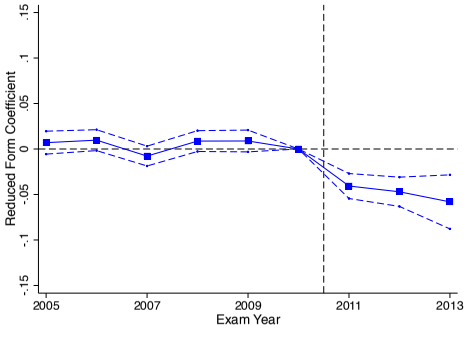
(F) Females: Adv. Regents Req. Taken



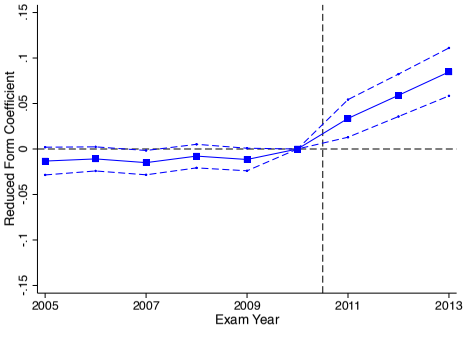
(G) Black/Hispanic: Score 65+ on First Admin.



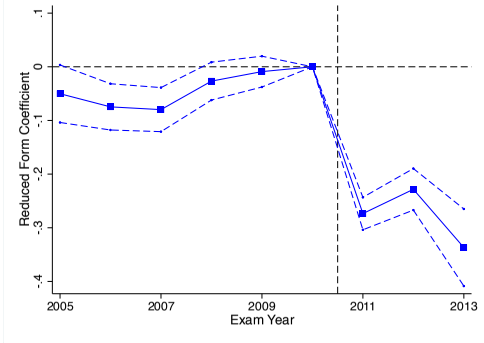
(H) Black/Hispanic: Graduate High School



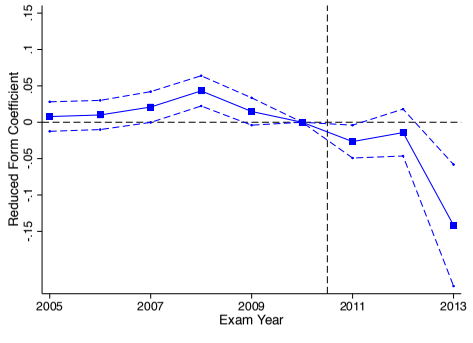
(I) Black/Hispanic: Adv. Regents Req. Taken



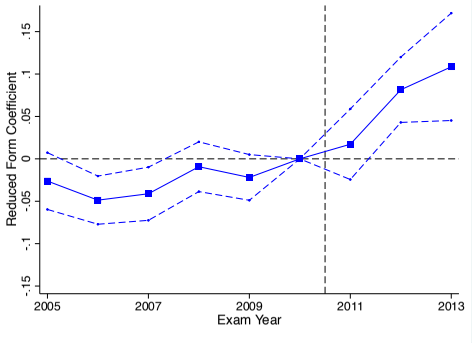
(J) White/Asian: Score 65+ on First Admin.



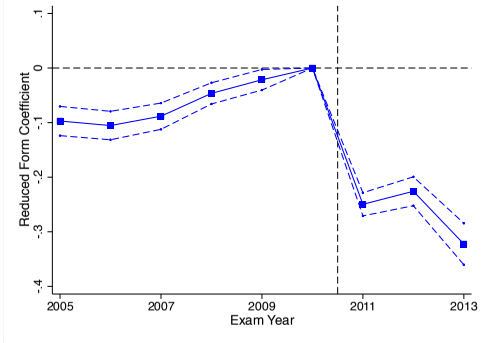
(K) White/Asian: Graduate High School



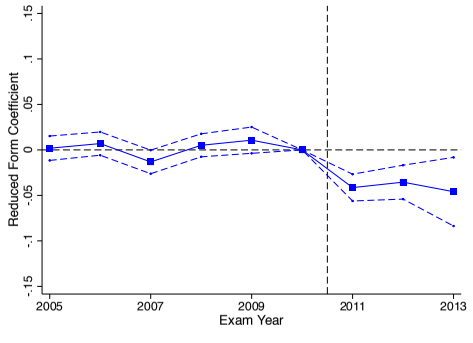
(L) White/Asian: Adv. Regents Req. Taken



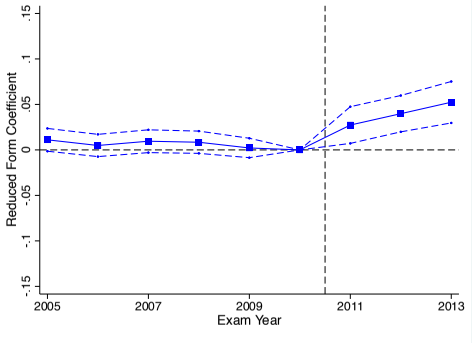
(M) Low 8th Score: Score 65+ on First Admin.



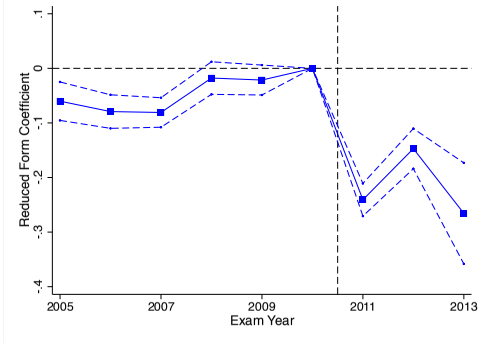
(N) Low 8th Score: Graduate High School



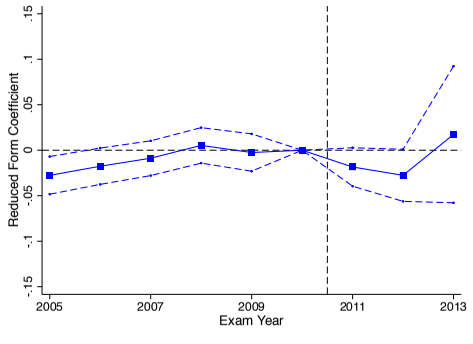
(O) Low 8th Score: Adv. Regents Req. Taken



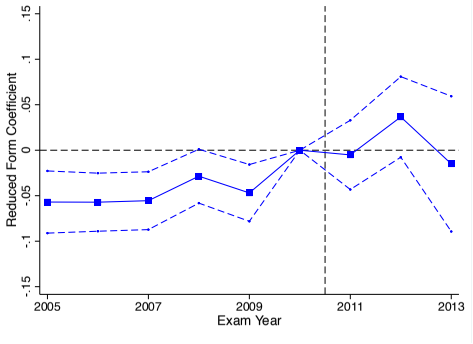
(P) High 8th Score: Score 65+ on First Admin.



(Q) High 8th Score: Graduate High School



(R) High 8th Score: Adv. Regents Req. Taken



Note: These figures plot the reduced form impact of the Regents grading reforms on selected outcomes by student subgroup. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. We report reduced form results using the interaction of taking the test in the indicated year and score in the manipulable range around the 65 cutoff. See the text for additional details.

Appendix Table A1: Regents Exam Requirements by Diploma Type and Cohort

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

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

Appendix Table A2: AIC for Different Polynomial Choices

Polynomial Order	AIC
(1)	(2)
First	51410.0
Second	50119.7
Third	48320.8
Fourth	48025.1
Fifth	47980.9
Sixth	47888.6
Seventh	47927.6

Note: This table reports Akaike Information Criteria for a series of regressions of test score frequency on various polynomials of test score. All polynomials are fully interacted with indicators for each test-year combination.

Appendix Table A3: Manipulable Scores by Test Subject x Year

	Comp. English	Living Env.	Math A	U.S. History	Global History	Int. Algebra
June 2004:	(1)	(2)	(3)	(4)	(5)	(6)
55 Cutoff	50-58	50-56	50-55	50-58	50-58	
65 Cutoff	60-67	60-65	60-65	60-68	60-68	
June 2005:						
55 Cutoff	50-57	50-56	50-55	50-59	50-58	
65 Cutoff	60-68	60-65	60-65	60-68	60-68	
June 2006:						
55 Cutoff	50-57	50-55	50-56	50-58	50-58	
65 Cutoff	60-67	60-65	60-65	60-69	60-68	
June 2007:						
55 Cutoff	50-58	50-55	50-56	50-58	50-57	
65 Cutoff	60-67	60-65	60-65	60-67	60-67	
June 2008:						
55 Cutoff	50-57	50-55	50-55	50-58	50-58	50-56
65 Cutoff	60-66	60-65	60-65	60-68	60-67	60-65
June 2009:						
55 Cutoff	50-57	50-55		50-58	50-58	
65 Cutoff	60-67	60-65		60-68	60-67	60-65
June 2010:						
55 Cutoff	50-58	50-55		50-59	50-58	
65 Cutoff	60-67	60-65		60-67	60-68	60-65
June 2011:						
65 Cutoff	60-68	60-65		60-68	60-68	60-65
June 2012:						
65 Cutoff	60-68	60-65		60-67	60-67	60-65
June 2013:						
65 Cutoff	60-69	60-65		60-67	60-67	60-65

Note: This table reports the manipulable scores around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2013. See the text for details.

Appendix Table A4: Estimates by Test Subject x Year x Month

	Comp. English	Living Env.	Math A	U.S. History	Global History	Int. Algebra
January 2004:	(1)	(2)	(3)	(4)	(5)	(6)
Total Manipulation	5.07		4.62			
	(0.07)		(0.11)			
In-Range Manipulation	48.11		24.82			
	(2.27)		(0.20)			
	20325		22781			
June 2004:						
Total Manipulation	6.25	6.90	5.48	5.86	5.97	
	(0.06)	(0.15)	(0.11)	(0.02)	(0.01)	
In-Range Manipulation	55.86	32.71	29.55	50.25	51.34	
	(2.49)	(0.14)	(0.13)	(0.61)	(0.44)	
	26699	41875	31158	38106	48934	
January 2005:						
Total Manipulation	5.51		4.45			
	(0.02)		(0.12)			
In-Range Manipulation	50.54		27.16			
	(0.97)		(0.19)			
	23838		24449			
June 2005:						
Total Manipulation	7.68	6.98	5.05	6.86	7.34	
	(0.02)	(0.09)	(0.14)	(0.02)	(0.01)	
In-Range Manipulation	70.49	33.23	26.84	56.91	62.81	
	(1.35)	(0.07)	(0.21)	(0.79)	(0.77)	
	24052	43572	31905	35387	47447	
January 2006:						
Total Manipulation	4.50		3.01			
	(0.02)		(0.17)			
In-Range Manipulation	42.28		16.24			
	(0.67)		(0.53)			
	27808		28171			
June 2006:						
Total Manipulation	6.13	7.73	5.05	7.00	7.88	
	(0.02)	(0.09)	(0.20)	(0.03)	(0.02)	
In-Range Manipulation	56.65	35.98	24.47	57.72	66.97	
	(0.89)	(0.04)	(0.37)	(1.06)	(1.06)	
	24483	41348	28267	36798	47147	
January 2007:						
Total Manipulation	5.91		4.18			
	(0.02)		(0.13)			
In-Range Manipulation	54.49		21.79			
	(1.02)		(0.28)			
	29929		27671			
June 2007:						
Total Manipulation	6.19	7.85	4.07	7.19	7.55	
	(0.02)	(0.13)	(0.15)	(0.02)	(0.02)	
In-Range Manipulation	57.03	36.68	19.12	60.32	65.53	
	(1.07)	(0.05)	(0.37)	(0.94)	(0.97)	
	22403	40932	27248	37687	44551	
January 2008:						
Total Manipulation	3.19		3.91			
	(0.02)		(0.19)			
In-Range Manipulation	32.88		21.93			
	(0.54)		(0.43)			
	27915		26352			

	Comp. English	Living Env.	Math A	U.S. History	Global History	Int. Algebra
June 2008:	(1)	(2)	(3)	(4)	(5)	(6)
Total Manipulation	3.94	5.85	4.99	5.46	6.37	3.03
	(0.02)	(0.08)	(0.26)	(0.02)	(0.02)	(0.12)
In-Range Manipulation	40.68	36.84	21.76	52.98	56.55	22.13
	(0.66)	(0.03)	(0.52)	(1.02)	(0.69)	(0.39)
	23618	42073	18044	38289	44951	34185
January 2009:						
Total Manipulation	3.86					3.84
	(0.03)					(0.20)
In-Range Manipulation	38.53					30.55
	(0.87)					(0.34)
	27547					10489
June 2009:						
Total Manipulation	4.10	5.32		7.49	6.44	4.33
	(0.03)	(0.16)		(0.02)	(0.02)	(0.20)
In-Range Manipulation	40.89	32.41		69.91	57.17	34.08
	(0.92)	(0.18)		(1.56)	(0.76)	(0.23)
	23697	41261		39470	43283	39513
January 2010:						
Total Manipulation	3.46					3.78
	(0.04)					(0.10)
In-Range Manipulation	38.10					38.75
	(1.30)					(0.17)
	27099					13956
June 2010:						
Total Manipulation	3.50	6.10		5.60	6.32	4.07
	(0.04)	(0.14)		(0.03)	(0.03)	(0.15)
In-Range Manipulation	38.49	36.42		53.41	58.77	31.11
	(1.32)	(0.07)		(1.22)	(1.20)	(0.23)
	22771	41477		37435	42707	34132

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

Appendix Table A5: Summary Statistics for School x Subject In-Range Manipulation

	In-Range Manipulation			Within-School Correlation				
	Obs.	Mean	S.D.	U.S. History	Global History	English	Math	Living Env.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
U.S. History	259	54.2	33.7	1.00				
Global History	263	59.1	36.3	0.78	1.00			
English	258	47.1	23.4	0.30	0.25	1.00		
Math	271	26.2	15.3	0.21	0.22	0.21	1.00	
Living Environment	273	36.1	16.8	0.06	0.07	0.14	0.25	1.00

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

Appendix Table A6: Comparison of Pilot and Non-Pilot High Schools

	Pilot Schools	Non-Pilot Schools	Difference
Characteristics:	(1)	(2)	(3)
Male	0.484	0.466	0.018
White	0.197	0.107	0.090**
Asian	0.206	0.204	0.003
Black	0.276	0.302	−0.026
Hispanic	0.315	0.383	−0.068*
Free Lunch	0.651	0.699	−0.048
8th Grade Test Scores	0.199	0.161	0.038
Core Regents Performance:			
Comprehensive English	76.890	75.215	1.675
Living Environment	74.932	74.569	0.364
Int. Algebra	68.795	69.484	−0.689
U.S. History	77.513	76.542	0.971
Global History	72.184	70.781	1.403
Students	54,852	73,416	

Note: This table reports summary statistics for students in New York City taking a core Regents exam in 2010-2011. Column 1 reports mean values for students enrolled in a school that is in the distributed scoring pilot program. Column 2 reports mean values for students not enrolled in a school that is in the distributed scoring pilot program. Column 3 reports the difference in means with standard errors clustered at the school level. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table A7: Difference-in-Differences Placebo Estimates

	Pre-Reform Mean	Reduced Form	
Panel A: Characteristics	(1)	(2)	(3)
Male	0.470	0.0100	0.0092
	(0.499)	(0.0074)	(0.0066)
White	0.144	−0.0004	−0.0049
	(0.351)	(0.0058)	(0.0034)
Asian	0.177	0.0049	0.0001
	(0.382)	(0.0072)	(0.0054)
Black	0.321	−0.0121	0.0013
	(0.467)	(0.0101)	(0.0054)
Hispanic	0.352	0.0101	0.0057
	(0.478)	(0.0091)	(0.0061)
Free Lunch	0.582	0.0225**	0.0245***
	(0.493)	(0.0092)	(0.0076)
Above Median 8th Score	0.546	−0.0027	−0.0063
	(0.498)	(0.0068)	(0.0064)
Panel B: Predicted Outcomes			
Predicted Graduation	0.795	−0.0002	−0.0017
	(0.137)	(0.0019)	(0.0017)
Predicted Regents Requirements	0.885	−0.0005	−0.0010
	(0.054)	(0.0007)	(0.0007)
Predicted Adv. Regents Requirements	0.369	0.0001	−0.0040
	(0.247)	(0.0035)	(0.0030)
Observations	1,002,804	1,002,804	1,002,804
Student Controls	–	No	No
Year x Score Trends	–	Yes	Yes
School Fixed Effects	–	No	Yes

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

Appendix Table A8: Difference-in-Differences Results by Subject

	All Core Exams	Living Env.	Int. Algebra	Global History	English	U.S. History
Panel A: High School Grad.	(1)	(2)	(3)	(4)	(5)	(6)
Graduate High School	0.215***	0.181***	0.138**	0.182***	0.315***	0.292***
	(0.023)	(0.037)	(0.055)	(0.025)	(0.078)	(0.047)
Pre-Reform Mean	0.799	0.826	0.759	0.795	0.800	0.835
Panel B: Diploma Requirements						
Regents Req. Taken	0.037	0.019	−0.054	0.053**	0.082*	0.043
	(0.032)	(0.050)	(0.095)	(0.026)	(0.048)	(0.029)
Pre-Reform Mean	0.896	0.912	0.861	0.905	0.884	0.938
Adv. Regents Req. Taken	−0.106**	−0.132**	−0.453***	0.087*	−0.212**	−0.064
	(0.044)	(0.059)	(0.101)	(0.049)	(0.101)	(0.059)
Pre-Reform Mean	0.366	0.415	0.327	0.374	0.350	0.373
Panel C: Advanced Science and Math Exams						
Take Physical Science Exam	0.028	0.008	−0.055	0.038	0.013	0.076
	(0.027)	(0.046)	(0.075)	(0.030)	(0.079)	(0.050)
Pre-Reform Mean	0.722	0.786	0.674	0.725	0.697	0.744
Take Adv. Math Sequence	−0.066*	−0.077*	−0.367***	0.135***	−0.181*	−0.069
	(0.038)	(0.045)	(0.092)	(0.046)	(0.098)	(0.059)
Pre-Reform Mean	0.389	0.430	0.352	0.398	0.376	0.401
Observations	1,674,762	301,881	367,517	333,406	376,374	295,584
Student Controls	Yes	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table reports two-stage least squares estimates of the effect of test score manipulation by subject. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. We use the interaction of taking the test between 2011-2013 and scoring in the manipulable range around the 65 cutoff as an instrument for scoring 65+ on the first administration. All specifications include the baseline characteristics from Table 1, an indicator for scoring below the manipulable range in 2011-2013, 10-point scale score x subject effects, year x subject effects, and 10-point scale score x subject linear trends. Standard errors are clustered at both the student and school level. The pre-reform sample mean for each subgroup is reported in brackets. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table A9: Difference-in-Differences Results for Additional Outcomes

	Pre-Reform Mean	Reduced Form		2SLS	
Panel A: Diploma Requirements	(1)	(2)	(3)	(4)	(5)
Regents Requirements Met	0.662	−0.136***	−0.139***	0.499***	0.512***
	(0.473)	(0.015)	(0.014)	(0.056)	(0.054)
Adv. Regents Requirements Met	0.229	0.035***	0.027**	−0.127***	−0.101**
	(0.420)	(0.012)	(0.012)	(0.046)	(0.044)
Panel B: Advanced Science and Math					
Take Physical Science Exam	0.724	0.001	−0.003	−0.004	0.011
	(0.447)	(0.010)	(0.008)	(0.035)	(0.028)
Take Advanced Math Sequence	0.390	0.017	0.011	−0.063	−0.042
	(0.488)	(0.012)	(0.011)	(0.044)	(0.041)
Panel C: Other Attainment Measures					
Years Enrolled in High School	4.124	−0.109***	−0.105***	0.400***	0.387***
	(0.581)	(0.007)	(0.007)	(0.030)	(0.028)
Highest Enrolled Grade	11.837	−0.078***	−0.078***	0.287***	0.286***
	(0.528)	(0.007)	(0.006)	(0.025)	(0.025)
Panel D: High School Graduation					
Graduate in 5 Years	0.849	−0.056***	−0.057***	0.197***	0.203***
	(0.358)	(0.006)	(0.006)	(0.023)	(0.022)
Graduate in 6 Years	0.877	−0.044***	−0.044***	0.153***	0.156***
	(0.328)	(0.005)	(0.005)	(0.020)	(0.019)
Panel E: GED Receipt					
GED Diploma	0.006	0.001	0.001	−0.005	−0.005
	(0.075)	(0.001)	(0.001)	(0.003)	(0.003)
Observations	1,002,804	1,002,804	1,002,804	1,002,804	1,002,804
Student Controls	–	Yes	Yes	Yes	Yes
Year x Score Trends	–	Yes	Yes	Yes	Yes
School Fixed Effects	–	No	Yes	No	Yes

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

Appendix Table A10: Robustness of Difference-in-Differences Results

	2SLS				
Panel A: High School Graduation	(1)	(2)	(3)	(4)	(5)
Graduate High School	0.167***	0.129***	0.213***	0.170***	0.165***
	(0.021)	(0.021)	(0.024)	(0.021)	(0.020)
Panel B: Diploma Requirements					
Regents Requirements Taken	0.016	0.001	0.035	0.014	0.013
	(0.043)	(0.031)	(0.063)	(0.041)	(0.043)
Adv. Regents Requirements Taken	−0.098*	−0.104***	−0.104	−0.103**	−0.104**
	(0.051)	(0.034)	(0.077)	(0.050)	(0.050)
Observations	1,002,804	442,789	407,637	1,002,804	1,002,804
Student Controls	Yes	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes	Yes
Drop [0,59] Scores	No	Yes	Yes	No	No
Drop [c_e+1,80] Scores	No	No	Yes	No	No
Drop [81,100] Scores	No	Yes	No	No	No
IV Year-Specific Interaction	No	No	No	Yes	Yes
IV Pilot School Interaction	No	No	No	No	Yes

Note: This table reports two-stage least squares estimates of the effect of test score manipulation using different instrumental variables for scoring 65+ on the first administration. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. Column 1 replicates our preferred specification from Table 4. Column 2 drops scores below the manipulable range and from 81-100. Column 3 drops scores below the manipulable range and from one plus the upper limit of the manipulable range to 80. Column 4 uses the interactions of scoring in the manipulable range around the 65 cutoff and year-specific indicators for taking the test between 2011-2013 as instruments. Column 5 uses the interactions of scoring in the manipulable range around the 65 cutoff and year-specific indicators for taking the test between 2011-2013 and an indicator for attending a school in the distributed grading pilot program as instruments. All specifications include the baseline characteristics from Table 1, an indicator for scoring below the manipulable range in 2011-2013, 10-point scale score x subject effects, year x subject effects, and 10-point scale score x subject linear trends. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exam subjects and cluster standard errors at the individual and school levels. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

Appendix Table A11: Additional Robustness of Difference-in-Differences Results

	2SLS			
Panel A: High School Graduation	(1)	(2)	(3)	(4)
Graduate High School	0.167***	0.223***	0.247***	0.256***
	(0.021)	(0.028)	(0.031)	(0.031)
Panel B: Diploma Requirements				
Regents Requirements Taken	0.016	0.007	0.017	0.023
	(0.043)	(0.060)	(0.073)	(0.081)
Adv. Regents Requirements Taken	−0.098*	−0.130*	−0.179**	−0.182*
	(0.051)	(0.072)	(0.089)	(0.098)
Panel C: First Stage Results				
Score 65+ in First Administration	−0.272***	−0.200***	−0.166***	−0.153***
	(0.010)	(0.009)	(0.008)	(0.007)
Observations	1,002,804	1,002,804	1,002,804	1,002,804
Student Controls	Yes	Yes	Yes	Yes
Year x Score Trends	Yes	Yes	Yes	Yes
School Fixed Effects	Yes	Yes	Yes	Yes
Upper Limit of Manipulable Range	Baseline	+ 1	+ 2	+ 3

Note: This table reports two-stage least squares estimates of the effect of test score manipulation using different instrumental variables for scoring 65+ on the first administration. The sample includes students entering high school between 2003-2004 and 2010-2011 and taking core Regents exams between 2004-2013. Column 1 replicates our preferred specification from Table 4. Column 2 increases the upper end of the manipulable region by 1 point. Column 3 increases the upper end of the manipulable region by 2 points. Column 4 increases the upper end of the manipulable region by 3 points. All specifications include the baseline characteristics from Table 1, an indicator for scoring below the manipulable range in 2011-2013, 10-point scale score x subject effects, year x subject effects, and 10-point scale score x subject linear trends. We stack student outcomes across the Living Environment, Math A/Integrated Algebra, and Global History exam subjects and cluster standard errors at the individual and school levels. *** = significant at 1 percent level, ** = significant at 5 percent level, * = significant at 10 percent level. See the data appendix for additional details on the sample construction and variable definitions.

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

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

A. Historical and Structural Details

The original Regents exams were administered as high school entrance exams for 8th grade students as early as 1865. These entrance exams were phased out relatively quickly, however, and in 1878 the Regents began offering advanced academic exams in various subjects to be used in college admissions. A fuller accounting of the Regents first 100 years can be found in Tinkelman (1965). Among the most important changes in recent years was the introduction of a new minimum competency test in the late 1970s that students were required to pass in order to graduate from high school. This competency test was replaced in the late 1990s by graduation requirements tied to the more demanding, end-of-course Regents Examinations we examine in this paper (Chudowsky et al. 2002).

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

Scoring of regents exams followed very explicit policies. For the English and social studies exams, principals are required to designate a scoring coordinator who is responsible for managing the logistics of scoring, assigning exams to teachers, and providing teachers with necessary training. For essay questions, the materials available to support this training include scoring rubrics and pre-scored "anchor papers" that provide detailed commentary on why the example essays merited different scores. For open-ended questions, the materials include a rubric to guide scoring. A single qualified teacher grades the open-ended questions on the social science exams. In the math exams, the school must establish a committee of three mathematics teachers to grade the examinations,

and no teacher should rate more than a third of the open-ended questions in mathematics. In the science exams, the school must establish a committee of two science teachers to grade the examinations, and no teacher should rate more than a half of the open-ended questions.

During our primary sample period (2003-2004 to 2009-2010), grading guidelines distributed to teachers typically included the following text explaining this policy: "All student answer papers that receive a scale score of 60 through 64 must be scored a second time to ensure the accuracy of the score. For the second scoring, a different committee of teachers may score the student's paper or the original committee may score the paper, except that no teacher may score the same open-ended questions that he/she scored in the first rating of the paper. The school principal is responsible for assuring that the student's final examination score is based on a fair, accurate and reliable scoring of the student's answer paper." See for example: `https://www.jmap.org/JMAPRegentsExamArchives/INTEGRATEDALGEBRAEXAMS/0610ExamIA.pdf`.

Two exceptions to these grading guidelines that we are aware of are the Chemistry exam in June 2001, which was only based on multiple choice questions, and the Living Environment exam in June 2001, where exams with scale scores from 62 to 68 were to be re-scored.

B. Use in School Accountability

In order to meet requirements for Adequate Yearly Progress (AYP) under the 2002 No Child Left Behind Act, high schools in New York must meet several criteria related to Regents examination participations and performance. First, 95 percent of a school's 12th graders must have taken the Regents Examinations in mathematics and English or an approved alternative (NYSED 2010). Second, the same must be true for all sub-groups with at least 40 students, where subgroups are based on race/ethnicity, poverty status, and program receipt. Third and fourth, a school's performance indices based on the Regents examinations in math and English must meet the statewide objectives for both its overall student population and among accountability sub-groups. The subject-specific performance indices are increasing in the share of students whose scale scores on the Regents Examination exceed 55, with students whose scores exceed 65 having twice the impact on this index. Specifically, the performance index equals 100*[(count of cohort with scale scores $\geq$ 55 + count of cohort with scale scores $\geq$ 65) / cohort size] (NYSED 2010). Thus, the performance index ranges from 0 (i.e., all students have scale scores below 55) to 200 (i.e., all students have scale scores of 65 or higher). These state-mandated performance objectives increased annually in order to meet NCLB's mandated proficiency goals for the school year 2013-2014. The fifth measure relevant to whether a high school makes AYP is whether its graduation rate meets the state standard, which is currently set at 80 percent. Like the other criteria, this standard is also closely related to the Regents Examinations, since eligibility for graduation is determined in part by meeting either the 55 or 65 scale score thresholds in the five core Regents Examinations.

New York City's separate accountability system awarded grades (A to F) to high schools starting in 2007. To form the school grades, the NYCDOE calculated performance within three separate elements of the progress report: school environment (15 percent of the overall score), student

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

C. Grading Appeals

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

Appendix C: Data Appendix
[NOT FOR PUBLICATION]

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

A. Data Sources

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

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

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

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

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

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

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

B. Sample Restrictions

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

1. We only include "core" Regents exams taken after 2003-2004. Exams taken before 2003-2004 cannot be reliably linked to student demographics. The core Regents exams during this time period include: Integrated Algebra (from 2008 onwards), Mathematics A (from 2003-2008), Living Environment, Comprehensive English, U.S. History and Global History. These exams make up approximately 75 percent of all exams taken during our sample period. Occasionally we extend our analysis to include the following "elective" Regents exams: Math B, Chemistry, and Physics. We do not consider foreign language exams due, in part, to the lack of score conversion charts for these years. We also do not consider Sequential Math exams, as these exams were typically taken before 2003. We also focus on exams taken in the regular test period. This restriction drops all core exams taken in August and the Living Environment, U.S. History, and Global History exams taken in January. We also drop all elective exams taken in January and August. However, the patterns we describe in the paper also appear in the these test administrations. Following this first set of sample restrictions, we have 2,470,187 exams in our primary window of 2003-2004 to 2009-2010.

2. Second, we drop observations with scale scores that are not possible scores for that given exam. This sample restriction leaves us with 2,453,437 remaining exams.

3. Third, we only consider a student's first exam in each subject to avoid any mechanical bunching around the performance thresholds due to re-taking behavior. This sample restriction leaves us with 1,977,221 remaining exams.

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

5. Fifth, we drop all exams originating from schools where more than five percent of core exam scores contain reporting errors. This is to eliminate schools with systematic mis-grading. This sample restriction leaves us with 1,728,043 remaining exams.

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

C. Adjustments to Raw Frequency Counts

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

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

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

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

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

D. Misc. Data Cleaning

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

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