Grading Leniency and Educational Choices: Evidence from a Blind Grading Regime

Gonçalo Lima *

October 5, 2023

Abstract

How do positive signals of academic ability change educational choices and future performance? I study the consequences of receiving a higher grade in a national standardized test on students' outcomes. For identification, I exploit the random assignment of graders to anonymized tests in the end of middle school, using administrative data from Portugal. First, I show that there is substantial grade manipulation in this blind grading context, consistent with graders being lenient. I then use bunching methods to identify the plausibly causal effect of leniency on students' choices. I find that low performers who benefit from lenient grading are significantly less likely to repeat the same school grade, and encouraged to enrol in a more demanding high school track. However, I cannot reject that there is no impact on academic achievement in the medium- and long-term. Finally, although graders do not observe test takers characteristics, I show that leniency in grading is selective on characteristics that may be indirectly inferred from the test. In particular, girls are marginally more likely to be bumped up.

JEL Classification: H75, I21, I28.

Keywords: Grading, manipulation, leniency, bunching, education, student outcomes.

*Contact: goncalo.lima@eui.eu. PhD Candidate at the Department of Economics of the European University Institute. I am grateful to Andrea Ichino, Sule Alan, Alessandro Tarozzi, Alex Solís, Ellen Greaves, Emilia del Bono, Fabrizia Mealli, Filipe B. Caires, João Firmino, Kosuke Imai, Libertad González, Marta Korczak and Pedro Freitas for helpful dicussions. Without the administrative data provided by the Portuguese Directorate General for Education and Science Statistics (DGEEC) this work would not have been possible. I am also grateful to the Economics of Education Knowledge Center, at Nova School of Business and Economics, in Portugal, for access to the data.

I. Introduction

Individuals often make human capital investment decisions in environments of incomplete information about their own academic ability (Manski, 1989; Ertac, 2005; Stange, 2012; Stinebrickner and Stinebrickner, 2012). Schooling, however, provides many opportunities to learn about one's aptitude and comparative advantage. By getting feedback on academic performance over time—primarily through grades—individuals can frequently update beliefs about their ability (Ertac, 2005; Bobba and Frisancho, 2022; Li and Xia, 2022). Based on the signaling value of grades, students may then change their self-perception, study effort and human capital investment decisions, even when these signals are noisy or biased.

Does a positive signal of ability change educational choices and future academic achievement? The causal effect of such a signal is theoretically ambiguous. On the one hand, perceived ability and future effort may be complements, so that an improvement in academic self-concept increases the motivation to exert effort (e.g. Li and Xia, 2022; Delavande et al., 2022). On the other hand, a higher level of perceived ability may also increase the temptation to 'rest on one's laurels'. If the marginal benefit of effort is decreasing in ability, individuals may reduce effort when feeling more talented. In this sense, a positive signal may induce excess self-confidence in one's capacity to translate effort into achievement (Bénabou and Tirole, 2002, 2003). Importantly, a perception of higher academic ability may also change human capital investment incentives.

In this paper, I study how middle school students respond to receiving a positive signal of ability through a higher grade in a standardized national exam. I do so by exploiting idiosyncratic features of Portugal's education system, using administrative panel data on the universe of students enrolled in public schools in the country. In particular, I investigate the effect of a higher grade on high school track choice and long-term achievement, given their relevance for predicting labor market and nonpecuniary outcomes in adulthood (Zax and Rees, 2002; Oreopoulos and Salvanes, 2011; Hanushek et al., 2015; Heckman et al., 2018)[1]. To disentangle the impact of a higher grade from other factors that affect both performance in the exam and future choices and outcomes, I exploit the fact that graders are lenient in this context, bumping up some students from lower to higher grades.

The type of ability signal received by students is crucial for my empirical investigation. Exam grades are reported in an integer 1-to-5 coarse scale, while being scored on a 0-100 more granular scale. In my main analysis, the coarse grades are the only signals students get about their performance in the exam. Individuals with very similar performance (in the granular scale) are thus assigned different, noisy signals of performance. Thus, this setting provides a natural experiment in which some graders effectively award a coarse, positive signal of ability, without any change in the student's performance in the exam. Moreover—and contrary to other studied settings (e.g. Diamond and Persson, 2016; Dee et al., 2019)—national exams

[1] In Portugal, test score and educational data has never been matched with individual labor market information.

are centrally handled, anonymized and randomly allocated to graders.

To identify signaling effects of a higher grade previous research has mostly relied on regression discontinuity (RD) designs (Tan, 2022; Li and Xia, 2022; McEwan et al., 2021; Anelli, 2020; Canaan and Mouganie, 2018; Avery et al., 2018; Goodman et al., 2017; Smith et al., 2017; Main and Ost, 2014), based on the assumption of no manipulation of the running variable (McCrary, 2008). Instead, my identification strategy relies on comparing observed outcomes with the counterfactual outcomes of students had they not been bumped up by a lenient grader. Importantly, students are blind to whether they benefited from an inflated grade, as they only observe the coarse grade. I use the estimator developed in Diamond and Persson (2016)—and has since been used in other contexts (e.g. Chen et al., 2021; Coviello et al., 2022)—to estimate the effects on the outcomes of interest. To impute the counterfactual outcomes, I rely on continuity assumptions, akin to those in RD settings. Importantly, by redefining the estimand of interest, I forgo the need of relying exclusively on partial identification of an RD parameter given manipulation of the running variable (see, e.g., Gerard et al., 2020; Ishihara and Sawada, 2022). Therefore, instead of focusing on the effect of being located at or close to the cutoff, I focus on identifying a local average treatment effect (LATE) of being manipulated across the cutoff. Because I do not observe graders I have no quasi-experimental variation in leniency to which a student is potentially exposed. However, using bunching methods, I can reliably estimate a first stage as the proportion of potentially manipulated students that are actually manipulated. Potentially manipulated students are those falling in regions of the test score distribution where some manipulation occurs. As in Diamond and Persson (2016), I use the relationship between outcomes and test scores excluding the data from manipulation regions to predict the counterfactual relationship between outcomes and test scores in the range of test scores that are potentially manipulated. Combining the counterfactual outcome in each test score bin with the counterfactual test score distribution allows to estimate the expected outcome in a counterfactual world without manipulation. The difference between observed and counterfactual outcomes identifies an intention-to-treat (ITT) estimate. Relying on a plausible monotonicity assumption, I scale the ITT by the first stage to identify the LATE (Imbens and Angrist, 1994).

I present five sets of findings. First, I find that about 10% of all Language and 5% of all Math test scores are manipulated. Of all students on the margin of being manipulated 58% in Language and 23% in Math receive an inflated test grade. To measure the extent of manipulation in exam scores, I employ 'bunching' methods (Kleven, 2016). Bunching has been used to estimate the extent of manipulation in given distributions, to study, for instance, behavioral responses of consumers around tax bracket cutoffs (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013). Estimating the extent of manipulation in test score distributions is just a natural extension of this type of methods (as in Diamond and Persson, 2016; Dee et al., 2019). I also uncover suggestive evidence that graders are more likely to be lenient in exam

items where awarding a full score is more open to grader's discretion. Given the institutional context, I argue that this type of behavior is consistent with graders having other-regarding preferences, which motivate them to be lenient.

Second, I find limited to no effects of grader's leniency on students' short-term educational choices. In the transition from middle to high school, students have to opt between an academic and a vocational track. Importantly, admission to any of the tracks in Portugal is not based on prior academic achievement or formal teacher recommendations. Hence, exam grades do not directly restrict the set of students' educational choices[2]. For the same level of ability, getting a positive signal through a higher grade in the national exam may encourage students to enroll in the more demanding, yet higher-return choice. I find that bumped up students close to the lowest grade cutoff in the Math exam are 6 p.p. more likely to choose the academic track in high school. For all other grade cutoffs, effects are generally positive but considerably small in magnitude and statistically indistinguishable from zero. Among those enrolling in the academic track, I also study whether receiving a higher exam grade in any of the subjects leads to a change in the likelihood of taking the scientific sub-track as the main concentration during their high school studies. I find that students bumped up to receive the highest grade in the Language exam are 5 p.p. less likely to choose the scientific sub-track, or 6% relative to the counterfactual mean.

I show evidence that the results for low-performing students are likely driven by a change in their exposure to schooling in the end of middle school. I investigate whether marginal changes in exam grades lead to changes in educational attainment in the end of middle school. I find that bumped up low-performing students in Language are 13 percentage points (p.p.) less likely to repeat the grade, or by 38% of the control group mean. For the case of Math, students are 6 p.p. less likely to repeat the same school grade (30% of the control group). Therefore, the change induced by a higher grade in track choices for these students is partially explained by the change in grade retention.

Third, I investigate what are the effects of a positive signal of performance in the end of middle school on long-term academic achievement in high school. In particular, I study whether students with bumped up grades in the middle school exam obtain better results in high school. I find no evidence that receiving an exogenously awarded higher grade in middle school exams leads to a consistent change in academic achievement in high school exit exams. The results suggest that, on average, the potential change in perceived academic ability induced by graders' leniency has no significant long-term consequences in academic achievement.

Fourth, I investigate whether student's behavior changes with the type of ability signal received. For the period of my main analysis, students are only reported their integer coarse grade in the exam (1-5), being uninformed about how far they scored from the closest grade cutoff. Separately using a period of

[2] Although it may restrict the timing at which students make these choices, as a low grade in the exam may induce some students in being retained in middle school one more year.

analysis where students are also reported their granular test score (0-100), I exploit this grading report reform to test two competing hypotheses of behavioral responses to grading. Under a full information regime, students observe their 'luck', or how far they were from getting a higher or lower coarse grade. In the case where only the coarser grade is reported, one should expect the update on the marginal students' belief to be stronger, as there is higher uncertainty about the individual's underlying ability. Studying the differences between the effects under these two grading regimes allows me to shed light on whether potential impacts on choices mostly operate through a discontinuous change in perceived ability or an encouragement effect associated with the coarse grade as a label. Most related studies focus on settings where underlying scores are not reported to students (e.g., Li and Xia, 2022; McEwan et al., 2021; Avery et al., 2018). However, effects in contexts where students are reported their underlying test scores are more likely to be driven by the psychological response of a higher performance label (e.g., Murphy and Weinhardt, 2020; Papay et al., 2016; Main and Ost, 2014). Exploiting this variation in grading regimes, I can study both scenarios within the same education system. I find that when granular scores are also reported, low-performing students in Math who are bumped up are still more likely to choose the academic track. Likewise, I also find that these low-performing students are less likely to repeat the school grade, lending credibility to the hypothesis that not repeating the grade is the main driver of a higher likelihood of choosing the academic track. Finally, I also find no consistent evidence of an effect on academic achievement in high school.

Finally, I find evidence that graders are selectively lenient. While graders do not directly observe test takers, leniency may still not be randomly awarded. First, graders can infer student characteristics from the exam. For instance, a grader may be able to identify a student's gender by the handwriting or gender-specific pronouns used in open-response items. Second, abler students may convey their ability in skills that are not to be scored. For instance, graders may be less willing to be lenient with those that are marginally worse in structuring answers, even if this should not marginally affect their test score. I find that girls are significantly more likely to being bumped up across grade cutoffs in the Language exam. However, I cannot reject the hypothesis that girls and boys are being randomly bumped up in the case of Math. Likewise, I find that better students, as measured by their subject grades in school before sitting the exam, are more likely to be manipulated.

This paper contributes to a small, yet growing literature, on the causes and motivations of grader discretion and manipulation on high stakes tests. In line with my findings, manipulation of New York City's (NYC) high school exit exams scores is driven by teachers' 'altruistic' motivation to help students meet given standards, rather than compensating for disadvantage (Dee et al., 2019). On the other hand, in Sweden, graders' leniency is targeted at good students with an idiosyncratic bad performance in the exam (Diamond and Persson, 2016). The institutional design of the education system may also work as extrinsic

motivation for manipulation. Indeed, evidence shows that penalties for poor school performance, school competition, and imperfect monitoring generate incentives for systematic cheating in grading (Jacob and Levitt, 2003; Lavy, 2009; Angrist et al., 2017).

My empirical analysis is closely related to that of four other papers, finding mixed effects of teachers' grade manipulation on human capital. In Sweden, grade inflation in a school subject leads to better performance in other subjects and end-of-year GPA, with effects being larger at the right-tail of the performance distribution. Furthermore, score manipulation in the beginning of lower secondary education raises the likelihood of high school graduation and college enrolment, with low performing students benefiting from some earning gains in the labor market by the time they are 23 (Diamond and Persson, 2016). Others find that grade inflation in the end of high school in New York City (US) raises the likelihood of graduation but discourages students from enrolling in advanced courses (Dee et al., 2019). In Romania, a policy that curbed teacher discretion on grading of high school high-stakes tests decreased college access to disadvantaged students (Borcan et al., 2017). Explicit cheating of teachers in the grading of students in a US school district—by erasing and correcting students' wrong answers—is found to lead to lower future performance and a higher likelihood of dropping out (Apperson et al., 2016). However, extant empirical evidence is based on contexts where grading is decentralized and students are directly observed by graders. To the best of my knowledge, this is the first paper to study the consequences of grading leniency in a setting where both students and graders are blind to each other's identities.

The effect of receiving a higher grade ties with a growing literature on the impact of feedback in learning environments. Empirical evidence showcases mixed results, depending on the stage of the educational career, the type (e.g., Brade et al., 2022; Kinne, 2022), and the timing (e.g., Fischer and Wagner, 2018) of feedback received. Bandiera et al. (2015) find that feedback on past performance has a positive impact on short-term achievement and attainment in a college context. Most other studies focus on the impact of relative performance feedback. While some find overall positive effects of relative performance feedback among college and high school students (Dobrescu et al., 2021; Tran and Zeckhauser, 2012; Azmat and Iriberri, 2010), others find heterogeneous effects across gender and performance level (Goulas and Megalokonomou, 2021; Kajitani et al., 2020), or even negative effects from greater grade transparency (Azmat et al., 2019). I contribute to this literature by studying the long-term effects of receiving positive feedback signal—through a higher grade—while student ability is kept fixed.

The remainder of the paper is structured as follows. Section II presents the relevant institutional background. Section III describes the data used in the analysis, as well as relevant summary statistics. Section IV presents the identification strategy as well as the details of the estimation method. Section V shows the existence and extent of grade manipulation in my setting. Section VI discusses its effects—or lack thereof—among manipulated students. Section VII reports evidence of selective manipulation.

Finally, Section VIII summarizes the main takeaways of the paper, including a discussion about further research.

II. Setting

School System and Grading. Schooling in Portugal is compulsory from ages 6 to 18, or up until the end of high school. Basic education covers the initial nine grades of schooling, and it includes primary and middle school. Throughout middle school, children's academic achievement is evaluated on a 1 to 5 integer grading scale in each subject, where 5 is the highest grade. Students fail a subject if they cannot obtain a grade higher than 2. The final grade in a subject depends on a combination of in-class tests designed by teachers, class participation and national exams. Grades on all subjects contribute to the student's final grade point average (GPA)[3]. A student has to repeat the same year grade if she fails both Math and Portuguese Language, or three subjects. Grade retention is a ubiquitous remedial strategy in the Portuguese education system. Every year, a non-negligible proportion of students is mandated to repeat the same grade[4].

Upon completing basic education, typically at age 15, students transition to high school. At this stage, students self-select into an academically-oriented or a vocationally-oriented track in high school. Students are free to choose the track they prefer, independently of their grades or national exam scores[5]. The academic track is geared towards the pursuit of university degrees upon completion; the vocational track offers professionally relevant curricula, geared toward faster integration in the labor market or vocationally-oriented colleges. Within the academic track, students must decide on one of four curricular offers (or sub-tracks): (i) science and technology; (ii) economic sciences; (iii) humanities; and (iv) arts. To graduate from the scientific sub-track, students must successfully complete high school courses in math, biology, physics or chemistry. Students that graduate from the scientific sub-track typically apply to college degrees for which access is more competitive, such as in STEM fields or medicine.

Students choose the track according to their curricular preferences, aspirations, as well as non-binding recommendations from teachers and school psychologists. The choice of high school track strongly correlates with family background and students' academic ability. High school graduation rates from either track are below 70% (see Section III).

National Exams. All eligible students are required to take national exams in Math and Portuguese Language at the end of ninth grade, typically in June[6]. Grading of national exams is blind: the tests are

[3] Students have to attend classes on twelve mandatory subjects: Math, Portuguese Language, Foreign Language I, Foreign Language II, Physics and Chemistry, Natural Sciences, History, Geography, Technological and Visual Education, Physical Education, Civic Education, and Information and Computer Technologies.

[4] About 9% in ninth grade (See Section III).

[5] In the case of excess demand, a public school has no discretion in rejecting applications. Enrollment is then determined according to a series of ranked criteria. The assignment system of students to schools is residence-based. A student has preferential placement in a high school in the catchment area to which they belong. Access to some private high schools may be dependent on previous academic achievement.

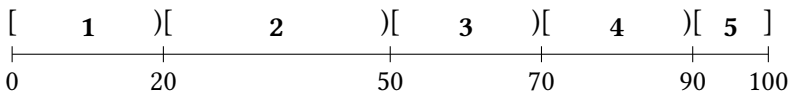
[6] Students with special education needs or requiring some other type of accommodation do not sit these tests. Eligible

randomly assigned to teachers not employed by the school of the students being evaluated (from now on, graders), and no information about students' or school's characteristics is conveyed to the graders.

Exam scoring is based on a set of detailed scoring criteria and solution keys for each exam item[7]. Nonetheless, the test score is ultimately at the grader's discretion. While points awarded to given test items may be hard to manipulate, others may give more leeway for teachers to provide additional points. Ninth grade exams are graded on an integer scale between 0 and 100. Furthermore, grading is not relative, i.e., there is no further standardization of exam grades or grading on a curve. Each score in the exam scale has a counterpart in the integer grade scale.

Figure 1 illustrates the exam grade $p \in \{1, ..., 5\}$ corresponding to each test score interval, $s \in \{0, ..., 100\}$. Scoring at or above 50 points would have the student pass the exam. Grade cutoffs are constant across years and common knowledge.

FIGURE 1. EXAM SCORES AND GRADE LEVELS



Notes. The figure shows the mapping of exam scores, in a 0-100 integer scale (bottom), to exam grades, in a 1-5 integer scale (above).

Exam performance are relevant signals of ability for schools, teachers, parents and students alike. First, since they are centrally administered and blindly graded, they provide arguably less biased evidence of student ability than if these were graded by the students' own teachers. Second, they can have real stakes for some students' educational attainment, as exam grades (p) are 30% of the student's final grade in the subject. Finally, exams are a relevant accountability mechanism in the system. School-level means in national exams determine school rankings at the national level. Parents' school and residence choices are influenced by this type of public reporting (Nunes et al., 2015)[8]. Schools have thus the incentive to make the stakes salient to students, even if test scores do not determine admission to the high school track.

Importantly, for my main period of analysis, only the (coarse) exam grades (p) are reported to students. However, I can observe each student's (granular) exam score (s), allowing me to measure how close each student is relative to each grade cutoff. For three years in the sample, students are also reported their exam score (s). I also leverage this change in reported information to study its differential impact on graders'

students are allowed to sit the exam in two phases. Students that do not sit through the first phase for some justified reason may sit through the second phase. In the sample, over 92% of the students enrolled in the ninth grade have sat the exam in the first phase.

[7] The national exams are developed by groups of teachers and subject experts hired by the Ministry of Education. Evaluation criteria are distributed to and implemented by the graders. Grading criteria and solutions are made available to the public after the tests are taken.

[8] Lower performing schools, particularly private ones, have been shown to be more likely to close since the public release of these type of rankings (Nunes et al., 2015).

leniency and student outcomes.

Graders' incentives. Exam graders do not know the students they evaluate. But even in a blind grading context, lenient graders may derive utility from bumping up students to a higher grade, provided that the costs of such manipulation are sufficiently low. In this context, manipulation costs are relatively limited. First, although students can plead for a re-scoring of the test, the first grader of a given test is never responsible for re-appreciation. It is unlikely that manipulation incentives are thus driven by grader's avoiding complaints and additional work from student re-scoring pleads. Additionally, pleads for re-scoring in first phase ninth grade exams are negligible[9]. Second, grader's do not have any caps or official minimum requirements on the number of students that must be assigned to each grade level (p). It is thus unlikely that manipulation would be driven by administrative concerns. Furthermore, it is also unlikely that a non-negligible portion of graders would grade students down to meet any quotas. Finally, the risk of individual detection is minimal. Although grading criteria are public, there is no specific system set to identify, reprehend or punish leniency of individual graders[10].

III. Data

Data Sources. I combine several administrative data sources from the Ministry of Education of Portugal over the period 2006-2018. While the full dataset covers the universe of students enrolled in public and private schools in the mainland, I restrict the analysis to public school students in the end of middle school for two main reasons. First, as I am interested in educational outcomes later on, I focus the analysis in a period for which I can still recover educational choices and outcomes throughout high school. Second, as private schools are not mandated to report information on student background variables, I exclude all private school students. The population of interest are all non-adult Math or Language exam takers enrolled in public schools in ninth grade between 2007 and 2015[11]. I use score-by-test-by-year data for yearly and overall estimates of manipulation. For all other estimates I use student-by-test level data. In summary, the dataset is a repeated cross-section with information on a total of 652, 829 students, which are all public school ninth grade exam takers in the country between 2007 and 2015 which I am able to track across the panel. Exam takers are enrolled in 1, 133 different public middle schools. The main variables of interest are presented below.

Test Scores. I observe each student's test score in ninth grade Language and Math exams. To avoid bunching around grade cutoffs due to re-scoring, I only consider scores attributed by the first grader that evaluated the exam. Scores (s) vary discretely between 0 and 100. To each test score corresponds a given grade (p), also varying discretely between 1 and 5. Figure 1 illustrates the mapping from test scores to

[9] Only 0.3% and 0.1% of Language and Math ninth grade first phase exams, respectively, are re-evaluated.

[10] In fact, the identification of graders that are potentially manipulating test scores would be hard to judge on a case-by-case basis.

[11] Appendix A presents the data sources, sample restrictions and data cleaning steps in more detail. It also presents more detailed descriptive statistics.

grade levels.

Manipulation Regions. I construct variables identifying whether a test score belongs to a given manipulation region. A manipulation region includes the range of test scores where there is potential grade manipulation. I include score points that are just to the left and just to the right of each of the grade cutoffs. For each year and test subject, I identify the potentially manipulated scores by inspecting the discrete distribution of test scores. Table A.3 in Appendix A documents the length of these regions by year and test subject. Section IV.2 provides further details on the construction of these variables.

Student Characteristics. For each individual I have information on gender and age, as well as whether if they are an immigrant, receive free or reduced price lunch, have computer and Internet at home, have an unemployed dad or mom, as well as whether at least one of the parents or legal guardian of the child completed a college degree. Each of these variables is measured at the beginning of ninth grade, before any of the outcomes and test score measures are realized.

Outcomes. For each individual I record whether they repeated ninth grade; whether they completed basic education; the track at which they enrolled in high school (academic or vocational); whether the student chose to enrol in the scientific sub-track, conditional on choosing the academic track; Math and Language teacher scores in tenth grade; as well as Math and Language exam scores in the end of twelfth grade.

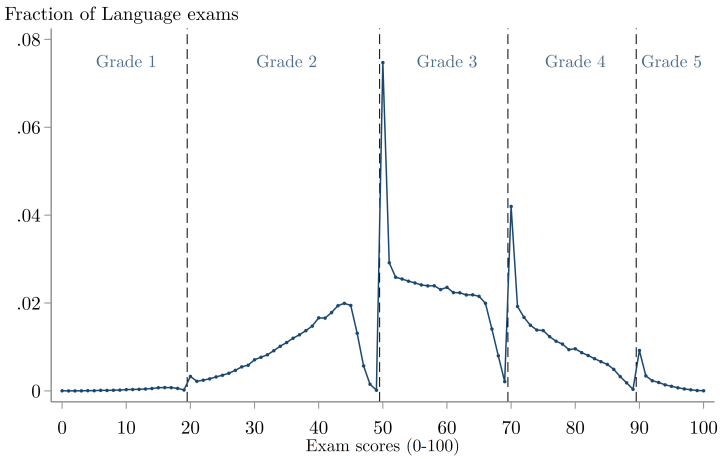
Summary Statistics. Table 1 reports summary statistics of the students that sat at least one of the ninth grade exams. Column 1 presents the number of observations with no missing values. Column 2 presents the mean, or the percentage of individuals for which the indicator variable is turned on for the full sample of exam takers. About 52% of students are female, and 6% are immigrants. Free or reduced price lunches identify students who benefit from means-tested public subsidies. About 35% of the students benefit from this type of social support. About 14% of the students have at least one of the parents or legal guardian unemployed before sitting the exam. Furthermore, for 17% of the students in the sample, at least one of the parents or legal guardian completed college. In terms of short-term educational attainment, 9% of students repeat the ninth grade, whereas nine in every ten students graduate from middle school. Of those that move on to attend high school, 72% choose the academic track. Conditional on enrolling in the academic track, 58% of the students enroll in the science sub-track. Columns 3 to 6 present the same statistics for the subset of students that are at risk of getting their test scores manipulated around each of the grade cutoffs, according to the manipulation regions associated with the Language test. As expected, one finds less socioeconomically disadvantaged students on the top of the test score distributions. Furthermore, virtually all students close to the highest grade cutoff graduate from basic education. Higher grades in either of the ninth grade exam also predict enrolment in high school academic track, particularly in the science sub-track.

Table 1. Summary Statistics of Exam Takers by Manipulation Region

	Exam Takers		Manipulation Region Around Threshold			
	Observations	Mean / %	$s \in [17, 20]$	$s \in [45, 53]$	$s \in [66, 72]$	$s \in [86, 91]$
	(1)	(2)	(3)	(4)	(5)	(6)
Socio-economic Characteristics						
Female	652,829	52.3	35.0	49.0	59.2	69.4
Immigrant	652,829	6.0	10.9	5.8	4.3	3.0
Age	652,829	14.5	15.1	14.6	14.4	14.3
Free or Reduced Price Lunch	652,829	34.6	54.8	38.7	25.6	14.2
Unemployment in the Household	639,187	14.4	18.3	15.5	12.1	8.4
Higher Education in the Household	627,292	17.3	5.1	12.1	23.6	44.1
Baseline School Performance						
School Grade (Language) [1-5]	551,012	3.2	2.5	3.0	3.6	4.5
School Grade (Math) [1-5]	555,965	3.0	2.4	2.8	3.5	4.4
Outcomes						
Repeated Ninth Grade	652,829	8.9	36.3	9.9	2.1	0.1
Graduated from Ninth Grade	649,757	90.5	60.6	89.6	97.8	99.9
Chose Academic Track	626,828	72.0	29.2	66.3	87.3	97.5
Chose Science Sub-track	451,019	58.2	39.5	51.2	64.5	77.3
School Performance in Tenth Grade (Language) [1-20]	469,693	12.5	10.2	11.4	13.4	16.3
School Performance in Tenth Grade (Math) [1-20]	327,495	11.8	10.4	10.5	12.5	15.9
Exam Score Percentile Twelfth Grade (Language)	394,905	49.6	16.2	36.6	57.8	82.6
Exam Score Percentile Twelfth Grade (Math)	228,550	50.2	20.7	36.5	53.4	75.1

Figure 2 presents the empirical distribution of test scores in the Language exam for all years in the sample, which represents clear descriptive evidence of manipulation around relevant grade cutoffs, illustrated by the dash lines. Scores just below each of the cutoffs are substantially less frequent then one would expect given a smooth empirical distribution of test scores.

Figure 2. Empirical Distribution of Language Exam Test Scores (2007-2015)



Notes. The figure shows the empirical distribution of test scores in the Language exam, pooling all years in the sample. The vertical dash lines represent grade cutoffs.

IV. Empirical Strategy

In this section I detail the derivation of the main measures and estimates used in the analysis. To impute counterfactual test score distributions and outcomes, I employ bunching methods (e.g. Saez, 2010; Kleven and Waseem, 2013). In particular, I estimate manipulation as the average of total excess and missing mass close to each grade cutoff. Counterfactual outcomes and characteristics of students in the counterfactual state of the world without manipulation are obtained by extrapolating the smooth relationship between these variables and test scores in unmanipulated regions of the test score distribution into each of the manipulation regions (as in Diamond and Persson, 2016). The difference between observed and counterfactual outcomes gives an intention-to-treat estimate. Relying on a plausible monotonicity assumption, I scale the ITT by the first stage to identify the effect of manipulation on those that are manipulated to a higher grade.

IV.1. Estimands of Interest

Consider a student i who sits the middle school exit exam in subject j[12]. According to their exam score s_i in the exam they receive a coarse grade p_i. I am interested in identifying the causal effect of getting a higher p_i on different future outcomes of the student, Y_i. As outcomes I consider students' human capital investment decisions, as well as medium- to long-term educational outcomes.

To disentangle the impact of a higher grade in the exam from other factors that affect both the grade and outcomes, I exploit bunching around grade cutoffs. As the empirical distribution of test scores suggests (see Figure 2), close to each coarse grade p cutoff, some students have their test scores manipulated to receive a higher grade. In this case, each student has an observed score, s_i, that can be higher than their *true* exam score, s_i^*. The setting is akin to a natural experiment, with a higher grade being granted to some students through graders' leniency, rather than an underlying difference in academic ability. However, in this empirical setting there are two main missing data limitations. First, graders are not identified, which does not allow me to estimate each grader's leniency parameter. Second, I cannot identify which students were effectively manipulated. To overcome these limitations, I use a strategy first developed in Diamond and Persson (2016).

Potential Outcomes. I consider potential outcomes in two possible states of the world: with and without grader manipulation. In a state of the world with grader manipulation ($M_i = 1$), the unobserved true exam scores (s_i^*) are potentially different than the observed ones (s_i). In a state of the world without grader manipulation ($M_i = 0$), true and observed scores are the same $\left(s_i = s_i^*, \forall i\right)$. For students who are ineligible to manipulation, it is always the case that $s_i = s_i^*$, in any of the two states of the world. Eligibility is determined by whether students' observed test scores (s_i) are in ranges where manipulation is feasible $\left(\mathcal{P}_p, \forall p\right)$.

[12] I omit the subscript j from the discussion for parsimony.

ITT. In a potential outcomes framework, each eligible student $\left(s_i \in \mathcal{P}_p\right)$ has a given outcome when assigned to a state of the world where manipulation is feasible, $Y_i(M_i = 1)$, or where manipulation is not possible, $Y_i(M_i = 0)$. Analogous to an intention-to-treat (ITT) effect, I define the impact of manipulation close to each grade (p) cutoff as:

$$\tau_{Y,p} := \mathbb{E}\left[Y_i(M_i = 1) - Y_i(M_i = 0) | s_i \in \mathcal{P}_p\right]. \tag{1}$$

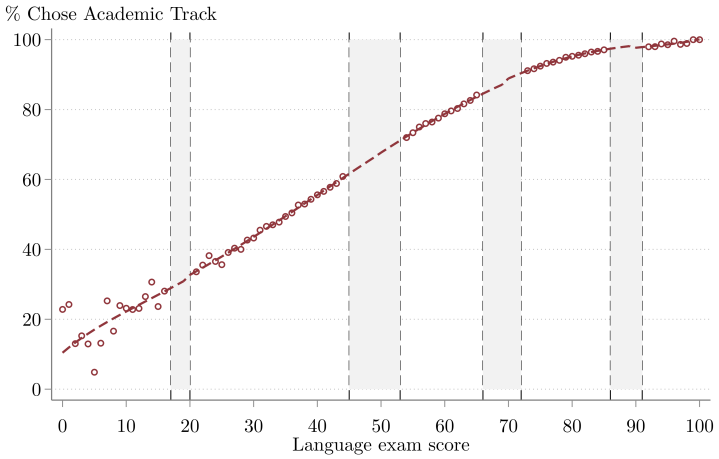
Equation 1 represents the difference in means between units in each state of the world. However, I have no experimental variation in assignment, M_i, as I can only observe students in a state of the world where manipulation exists, i.e., $M_i = M = 1, \forall i$. Thus, I cannot exploit random assignment to each state of the world. Instead, I rely on a continuity assumption to impute the average counterfactual outcomes in a state of the world without manipulation, akin to the assumption in bunching methods (Kleven, 2016). Specifically, I assume that the functional form that describes the relationship between test scores and outcomes—$Y(s)$—outside the manipulation regions $\left(s \notin \mathcal{P}_p\right)$ can be extrapolated to inside the manipulation regions $\left(s \in \mathcal{P}_p\right)$, describing the average outcome in each test score bin in a world without manipulation ($M = 0$).

Main Assumption. Figure 3 illustrates the main implication of this identifying assumption. The gray areas identify the manipulation regions, $\mathcal{P}_p$, taken as given for illustration. The red dash line represents the relationship between outcomes (e.g., choice of academic track in high school) and test scores that would occur in a world without manipulation. The red hollow circles are the observed average outcomes associated with each test score outside the manipulation regions. Average counterfactual outcomes in each manipulation region, $Y_i(M_i = 0), s_i \in \mathcal{P}_p$, are the values predicted by the function that relates test scores and outcomes outside the manipulation regions. The estimation problem—for which details are provided in Section 2—consists in finding the true underlying relationship between test scores and outcomes when $M = 0$.

Latent Groups. Importantly, Equation 1 cannot distinguish groups of students that are *actually* manipulated from those that are not. Although individuals are blindly graded, graders may still selectively manipulate some type of students, according to information inferred from the exam. Compliance with the manipulation assignment is, thus, potentially confounded. Crucially, each student with observed scores in any of the manipulation regions may take an observed score at or above each grade cutoff ($\bar{s}_p$), $T_i = \mathbb{1}\{s_i \geq \bar{s}_p\}$. As in an instrumental variables (IV) setting (Imbens and Angrist, 1994), I can identify individuals with observed test scores in each manipulation region $\mathcal{P}_p$ as: always-takers, never-takers, compliers and defiers.

In this setting, compliers are those those individuals in each manipulation region that are induced by grader manipulation to move from below to above the cutoff, i.e., $T_i(M_i = j) = j, \forall j \in \{0, 1\}$. In a world

Figure 3. Imputation of Counterfactual Outcomes



Notes. The figure shows the stylized relationship between choice of academic track (measured in percentage points) and test scores in the Language exam, pooling all years in the sample. The hollow red dots represent the average outcome of students who have a given test score. The red dash line represents a 7th degree polynomial fit, controlling for potential discontinuities in each grade cutoff, using only data from outside the manipulation regions $\left(s \notin \mathcal{P}_p\right)$, according to the methods described in Sections 2 and VI. The gray areas, in between vertical dash lines identify manipulation regions $\left(s \in \mathcal{P}_p\right)$. Observed outcomes in the manipulation regions are not represented in the figure.

without manipulation, their score would be below the grade cutoff. Compliance is, thus, determined by the decision of the grader in bumping up the grade of individual i. Always-takers are those that would have their score above the cutoff independently of the state of the world, $T_i(M_i = j) = 1, \forall j \in \{0, 1\}$. Analogously, never-takers, would always have their score below the cutoff, $T_i(M_i = j) = 0, \forall j \in \{0, 1\}$. Importantly, Proposition 1, in Appendix B excludes the possibility that graders would grade down students if their true score (s_i^*) would be above the cutoff. Therefore, I assume that manipulation is unidirectional—so that there are no defiers—and monotonicity is verified (Imbens and Angrist, 1994).

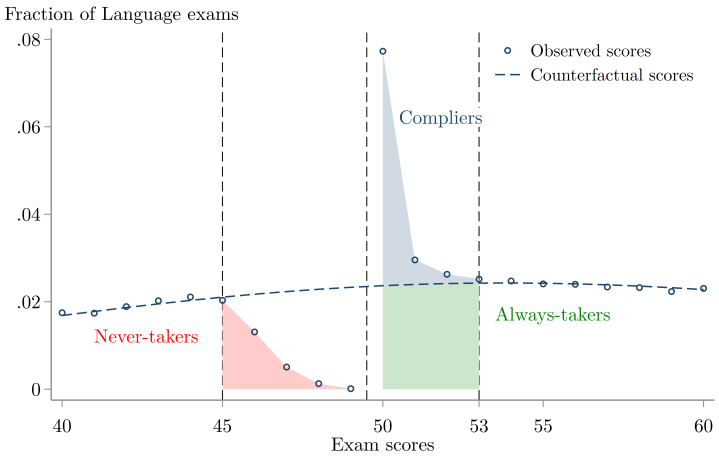
First Stage. The first stage proportion of compliers among eligible students is given by:

$$\tau_{T,p} := \mathbb{E}\left[T_i(M_i = 1) - T_i(M_i = 0) | s_i \in \mathcal{P}_p\right]. \tag{2}$$

To disentangle compliers from always-takers, in a state of the world with manipulation, I impute a counterfactual distribution of test scores F_s^* which indicates the proportion of individuals in each raw test score s^*, according to bunching methods (see, e.g., Kleven, 2016). This allows me to identify the proportion of individuals in each test score bin when $M = 0$, or $\mathbb{E}_i\left[T_i(0)\right]$. Section IV.2 details the estimation of this quantity.

Figure 4 illustrates the idea underlying the identification of each latent group. The left-most and right-most vertical dash lines delimit the manipulation region, in this case, $\mathcal{P}_2 = [45, 53]$. The center vertical dash line represents cutoff $\bar{s}_2 = 50$. The hollow circles identify the fraction of exam scores in each discrete exam score s. In this case, there is clear bunching above the cutoff. The blue dash line represents the imputed counterfactual distribution in the absence of manipulation. The group of compliers is illustrated

FIGURE 4. IDENTIFICATION OF LATENT GROUPS GIVEN MANIPULATION



Notes. The figure shows the distribution of test scores (hollow circles) in the Language exam, pooling all years in the sample. The blue dash line represents the imputed counterfactual distribution $\hat{F}_s^*$. The counterfactual distribution is computed by fitting a 10th degree polynomial using bunching methods, as described in Sections 2 and VI. The left-most and right-most vertical dash lines delimit manipulation region $\mathcal{P}_2 = [45, 53]$. The middle vertical dash line represents cutoff $\bar{s}_2 = 50$. The blue area gives a stylized representation of the group of compliers. The red area gives a stylized representation of the group of never-takers. The green area gives a stylized representation of the group of always-takers.

by the blue area. Compliers are those above the cutoff who are in excess of what would be predicted by the imputed counterfactual fractions F^*. These are the individuals who, if not manipulated, would be below the cutoff. The proportion of always-takers is given by the area below the counterfactual distribution at or above the cutoff and $s_i \in \mathcal{P}_p$. Never-takers are those that are not manipulated even in a state of the world with manipulation ($M_i = 1$).

LATE. As in Diamond and Persson (2016), I identify the effect of manipulation on the manipulated as:

$$\tau_p = \frac{\tau_{Y,p}}{\tau_{T,p}} = \frac{\mathbb{E}_i[Y_i(1) - Y_i(0)|s_i \in \mathcal{P}_p]}{\mathbb{E}_i[T_i(1) - T_i(0)|s_i \in \mathcal{P}_p]}. \tag{3}$$

Assuming that there is no effect of manipulation on never-takers and always-takers, τ_p identifies a local average treatment effect (LATE) for the group of manipulated students.

IV.2. Estimation

Measuring Manipulation. I start by denoting the fraction of students with an observed exam score of s as F_s. On the other hand, the fraction of students with a given test score in a counterfactual state of the world in which there is no manipulation ($M_i = 0, \forall i$) I denote by F_s^*.

The proportion of manipulated individuals at each cutoff can be re-written as the excess mass in the observed distribution relative to the counterfactual distribution of scores at or above each of the cutoffs. Equivalently, it can also be written as the missing mass of the observed distribution relative to the counterfactual one in below each of the cutoffs (Kleven, 2016):

$$B_p = \underbrace{\mathbb{E}\left[\sum_{s \geq \bar{s}_p} \left(F_s - F_s^*\right)\right]}_{\text{Excess mass}} = \underbrace{\mathbb{E}\left[\sum_{s < \bar{s}_p} \left(F_s^* - F_s\right)\right]}_{\text{Missing mass}}, \quad s \in \mathcal{P}_p. \tag{4}$$

As defined in Section IV, $\mathcal{P}_p$ are the regions of potentially manipulated scores, around each grade cutoff $\bar{s}_p$. Moreover, the total fraction of manipulated scores is defined as the sum of manipulated scores across each grade cutoff, $B = \sum_p B_p$.

Estimating the Counterfactual Distribution. To estimate the counterfactual distribution of scores, for each year and exam subject in the sample, I fit a polynomial to the fraction of exams in each discrete test score bin. I exclude the data in each manipulation region $\mathcal{P}_p$ by including indicator variables for each exam score in these regions. Specifically, I run OLS specifications of the form:

$$F_s = \sum_{q=0}^{Q} \theta_q \cdot s^q + \sum_{j \in \mathcal{P}} \lambda_j \cdot \mathbb{1}\{s = j\} + \varepsilon_s. \tag{5}$$

F_s is the the fraction of observed scores at each exam score level s, $\{\theta_q\}_{q=0}^{Q}$ are the coefficients of a Qth order polynomial, $\{\lambda_j\}_{j \in \mathcal{P}}$ are the coefficients associated with potentially manipulated scores, and ε_s is the sampling error. The coefficients of interest are the estimated parameters of the polynomial function $\{\hat{\theta}_q\}_{q=1}^{Q}$ and the coefficients of each dummy $\{\hat{\lambda}_j\}_{j \in \mathcal{P}}$. I recover the estimated counterfactual distribution as $\hat{F}_s^* = \sum_{q=0}^{Q} \hat{\theta}_q \cdot s^q$. To decide on the polynomial order Q, I use an optimal mean squared error criterion, based on a cross-validation method[13].

Defining the Manipulation Regions. A key step in the estimation of the counterfactual distribution is the correct specification of the manipulated scores $s \in \mathcal{P}_p$. I identify each manipulation region by a combination of testing regulations, monotonicity assumptions and visual inspection. First, I collapse the data to the score-by-test-by-year level. For each year and exam subject, I inspect the distribution of scores. I define the lower bound score of each manipulation region according to the simultaneous satisfaction of two criteria: (i) it is the first score (s)—approaching each grade cutoff—for which there is a decrease in mass relative to the previous score ($s-1$) higher than it would be predicted by just sampling error or noise; (ii) it is the first score for which there are always decreases in mass until the score just below each grade cutoff score $\bar{s}_p$. The noticeable accumulation of mass in $\bar{s}_p$ can still be followed by potentially manipulated scores. I define the upper bound of each manipulation region as the first score for which the decrease in mass relative to the previous score can be interpreted as sampling error or noise. As the distributions of scores—excluding those close to grade cutoffs—are relatively smooth, the visual identification of the potentially manipulated scores is straightforward. I argue that, in this case, the seemingly ad hoc judgement of what

[13] Appendix C provides a detailed description of the estimation procedure.

can be interpreted as noise or sampling error becomes less open to interpretation[14].

Estimating the First Stage. To ensure that the estimated counterfactual distribution is a valid probability mass function: (i) $\hat{F}_s^* \leq 1, \forall s$, and (ii) $\sum_s \hat{F}_s^* = 1$. OLS presents some limitations for both conditions to be verified. First, it cannot ensure that the fitted values will be weakly positive. Crucially, though, this limitation only bites in the tails of the distribution, where there is little observed mass, and the counterfactual fitted value may sometimes be negative. Second, the integration constraint may not be strictly verified. In particular, the integration constraint also implies that the estimated excess and missing masses around each cutoff should add to zero, i.e., $\sum_{s \geq \bar{s}_p} \hat{\lambda}_s + \sum_{s < \bar{s}_p} \hat{\lambda}_s = 0, s \in \mathcal{P}_p, \forall p \in \{2, ..., 5\}$.

Given the delimitation of the manipulation regions and the data, the method should ensure both values are statistically indistinguishable. Because of irreducible sampling error, $\hat{\varepsilon}_s$, the discreteness of s, and imperfect definition of manipulation regions, both values need not be exactly equal. To address these concerns and increase power, I compute the estimated manipulation at each grade level as the average of the absolute values of the excess and missing masses (as in Dee et al., 2019; Diamond and Persson, 2016):

$$\hat{B}_p = \frac{1}{2}\left[\sum_{s \geq \bar{s}_p} |\hat{\lambda}_s| + \sum_{s < \bar{s}_p} |\hat{\lambda}_s|\right], \quad s \in \mathcal{P}_p. \tag{6}$$

The estimates $\{\hat{B}_p\}_{p=2}^5$ identify the fraction of exam scores that were manipulated at each grade cutoff $\bar{s}_p$. The estimate of the total proportion of manipulated exam scores is $\hat{B} = \sum_p \hat{B}_p$.

A crucial estimate of interest will be the proportion of compliers. As discussed in Section IV, it measures the fraction of students that were manipulated among those in each manipulation region:

$$\hat{\tau}_{T,p} = \frac{\hat{B}_p}{\sum_s \hat{F}_s}, \quad s \in \mathcal{P}_p. \tag{7}$$

Which is the first stage estimate for the estimand in Equation 2 (Section IV). A second estimate of interest will be in-range manipulation (as defined in Dee et al., 2019), measuring the proportion of those with raw scores below the cutoffs that were effectively manipulated:

$$\hat{b}_p = \frac{\hat{B}_p}{\sum_{s < \bar{s}_p} \hat{F}_s}, \quad s \in \mathcal{P}_p. \tag{8}$$

I compute estimates for each combination of exam subject, year and cutoff. I also compute measures summing across all cutoffs and averaging across the years, weighting by the number of students in each year[15]. For inference, I use block bootstrapped standard errors. I define the blocking group as the students' class in the school. Since each grader receives batches of exams which mostly include students from the

[14] In other contexts, iterative procedures for the determination of the manipulation windows are sometimes used, either due to the multiplicity of test score distributions (e.g. Diamond and Persson, 2016) or due to a visually unidentifiable upper bound (e.g. Chetty et al., 2010).

[15] See Appendix C.

same class, it is likely that grade inflation is highly correlated within class. I draw with replacement blocks of students' classes from the original data at the student-by-test level to generate bootstrap versions of the data[16].

Estimating the Effects of Manipulation. To estimate intention-to-treat (ITT) effects I start by deriving the relationship between the outcome variables (Y_i) and the test scores that would be observed in a counterfactual state of the world without manipulation (s_i^*). I impute the expected counterfactual outcomes by fitting a polynomial to the data, excluding observations with test scores in the manipulation regions. Specifically, I run OLS specifications of the form:

$$Y_i = \underbrace{\sum_{q=0}^{Q'} \gamma_q \cdot s_i^q + \sum_{p=2}^{5} \delta_p \cdot \mathbb{1}\{s_i \geq \bar{s}_p\}}_{:=g^Y(s_i,\boldsymbol{\gamma},\boldsymbol{\delta})} + \omega_i, \quad s_i \notin \mathcal{P}_p. \tag{9}$$

In this case, $\{\gamma_q\}_{q=1}^{Q'}$ are the coefficients of a Q'th order polynomial, determined by an optimal mean squared error criterion, based on a cross-validation method[17]. Importantly, the specification also allows discontinuities (δ_p) in individual outcomes at each of the grade cutoffs, identifying the effects of being above each of the cutoffs in a world without manipulation. Including these discontinuities takes into account the effect a higher grade would have on the always-takers in a world without manipulation. The error term is ω_i.

The estimated counterfactual outcomes in the manipulation region are the fitted values of $g^Y(\cdot)$ inside the manipulation regions $\hat{g}^Y\left(s_i, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}\right), s_i \in \mathcal{P}_p$. The imputed expected outcome around each cutoff $p \in \{2, ..., 5\}$ in a state of the world without manipulation is:

$$\hat{Y}_p^* := \sum_{s \in \mathcal{P}_p} \hat{g}^Y\left(s_i, \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\delta}}\right) \cdot \frac{\hat{F}_s^*}{\sum_{s \in \mathcal{P}_p} \hat{F}_s^*}. \tag{10}$$

As defined above, $\hat{F}_s^*$ is the estimated counterfactual fraction of exams at each test score s. Therefore, $\hat{Y}_p^*$ is the weighted average of the counterfactual outcomes in $\mathcal{P}_p$, with weights given by the counterfactual fraction of students in each test score bin s. Analogously, I denote the observed mean outcome inside the manipulation region in a world with manipulation as $\bar{Y}_p$. The difference between the observed and counterfactual is the ITT estimate associated with each grade level $p \in \{2, ..., 5\}$: $\hat{\tau}_{Y,p} = \bar{Y}_p - \hat{Y}_p^*$.

Finally, the estimated LATE is just the Wald ratio between the ITT and the first stage (Diamond and Persson, 2016):

[16] For each bootstrap sample, I estimate the counterfactual distribution by fitting the optimal polynomial for that particular sample. Allowing a new polynomial fit for each bootstrap sample, I make standard errors reflect uncertainty relative to the optimal polynomial specification in the population. The standard errors are the standard deviations of 200 of these bootstrapped estimates. I also derive confidence intervals directly from the distributions of the bootstrap estimates. See Appendix C.4 for details.

[17] See Appendix C for details

$$\hat{\tau}_p := \frac{\hat{\tau}_{Y,p}}{\hat{\tau}_{T,p}}. \tag{11}$$

As for the first stage estimates, I use block bootstrapped standard errors, to reflect clustering at the classroom level. For each iteration of the bootstrap, I estimate the first stage ($\hat{\tau}_{T,p}$), the ITT ($\hat{\tau}_{Y,p}$) and the LATE ($\hat{\tau}_p$) using the same bootstrapped sample.

Testing for Selective Manipulation. In Portugal, graders of national exams do not know the students they grade. This feature of the data provides a natural laboratory to test whether graders—blind to the students they evaluate—still differentially bump up students with particular characteristics. I leverage the fact that these predetermined characteristics are observable in the data, although not being directly observed by the grader.

If manipulation is not selective on student type, I expect the group of manipulated students—the compliers ($i \in C^p$)—to be indistinguishable from the group of students just below the cutoff in a state of the world without manipulation—the group of eligible compliers and never-takers ($i \in C^p \cup N^p$). The null hypothesis of interest can be stated as:

$$\mathrm{H}_0: \quad \underbrace{\mathbb{E}\left[X_i | i \in C^p\right]}_{\text{Char. of Compliers}} = \underbrace{\mathbb{E}\left[X_i | s_i^* < \bar{s}_p\right]}_{\text{Char. of Eligible}}, \quad s_i^* \in \mathcal{P}_P. \tag{12}$$

In this case, X_i is a given characteristic of student i, such as gender. Compliers, as defined in Section IV1, are bumped up across each grade cutoff. The average characteristics of compliers can be inferred from re-weighted sums of observed and counterfactual characteristics of students above ($\mathbb{E}[X_i | s_i \geq \bar{s}_p], \mathbb{E}[X_i | s_i^* \geq \bar{s}_p]$) and below the cutoff ($\mathbb{E}[X_i | s_i < \bar{s}_p], \mathbb{E}[X_i | s_i^* < \bar{s}_p]$) (Diamond and Persson, 2016). Appendix C.4 shows these derivations in detail.

Key quantities of interest will be the imputed counterfactual relationships between each relevant characteristic and test scores, akin to Equation 9 for the case of outcomes. I impute the expected counterfactual characteristics by fitting a polynomial to the data, excluding observations with test scores in the manipulation regions:

$$X_i = \underbrace{\sum_{q=0}^{Q''} \eta_q \cdot s_i^q}_{:= g^X(s_i, \boldsymbol{\eta})} + u_i, \quad s_i \notin \mathcal{P}_p. \tag{13}$$

The set $\{\eta\}_{q=1}^{Q''}$ collects the coefficients of a Q''th order polynomial, determined by an optimal mean squared error criterion, based on a cross-validation method. The error term is u_i.

As for the outcome variables, the estimated counterfactual characteristics in the manipulation region are the fitted values of $g^X(\cdot)$ inside the manipulation regions $\hat{g}^X(s_i, \hat{\boldsymbol{\eta}}), s_i \in \mathcal{P}_p$. The imputed expected

characteristic for individuals below the cutoff in a state of the world without manipulation is:

$$\hat{X}_p^- = \sum_{s_i < \bar{s}_p} \hat{g}^X(s_i, \hat{\boldsymbol{\eta}}) \cdot \frac{\hat{F}_s^*}{\sum_{s < \bar{s}_p} \hat{F}_s^*}, \quad s_i \in \mathcal{P}_p. \tag{14}$$

Analogously, the mean characteristic of the group of always-takers ($i \in A^p$) is estimated as:

$$\hat{X}_p^+ = \sum_{s_i \geq \bar{s}_p} \hat{g}^X(s_i, \hat{\boldsymbol{\eta}}) \cdot \frac{\hat{F}_s^*}{\sum_{s \geq \bar{s}_p} \hat{F}_s^*}, \quad s_i \in \mathcal{P}_p. \tag{15}$$

The estimate of selective manipulation for each individual characteristic X is given by:

$$\hat{\Gamma}_p^X = \hat{X}_p^C - \hat{X}_p^-. \tag{16}$$

In which $\hat{X}_p^C$ is the estimated mean value of characteristic X among the group of compliers[18]. As for outcomes, I estimate the standard errors and confidence intervals of Γ_p^X through block bootstrap. The magnitude and statistical significance of the estimates allow me to test whether graders are, on average, manipulating students with particular characteristics.

V. Grade Manipulation

As a first step in the analysis I show and quantify the extent of grade manipulation in my setting. I find evidence of substantial grade manipulation close to every grade cutoff. Manipulation is larger in the Language exam and somewhat more limited in the case of Math. In particular, I find evidence that graders tend to manipulate the scoring of exam items where what constitutes a fully right answer is more open to individual interpretation. I also find that the type of grade reporting—i.e., whether the test score is also reported to students—does not seem to change the intensity of graders' manipulation. Based on this evidence, I argue that this type of manipulation is consistent with graders being lenient with students at the margin of a higher grade.

V.1. Documenting Grade Manipulation

Figure 5 shows evidence of clear manipulation around grade cutoffs. It documents the existence and extent of grade manipulation in ninth grade Language (top panel) and Math exams (bottom panel) in Portugal, between 2007 and 2012[19]. Scores just below each of the cutoffs are substantially less frequent

[18] The estimate for the group of compliers is given by:

$$\hat{X}_p^C = \frac{1}{2 \cdot \hat{B}_p} \left[\bar{X}_p^+ \cdot \sum_{s \geq \bar{s}_p} F_s - \hat{X}_p^+ \cdot \left(\sum_{s \geq \bar{s}_p} F_s - \hat{B}_p \right) + \left(\hat{X}_p^- - \bar{X}_p^- \right) \cdot \sum_{s < \bar{s}_p} \hat{F}_s^* \right], \tag{17}$$

where $\bar{X}_p^+$ and $\bar{X}_p^-$ are the observed expected characteristics for individuals in the manipulation region with scores below and above the cutoff $\bar{s}_p$, respectively. See Appendix C.4 for details.

[19] To build each plot of test score distributions, I first collapse the data to the score-by-test-by-year level. For each exam year, I identify the set of manipulable scores. I then collapse the data to the score-by-test level, constructing empirical test score

then one would expect given a smooth empirical distribution of test scores. On the other hand, scores just above each of the cutoffs are substantially more frequent. Crucially, the observed empirical distribution is relatively smooth across the range of test scores that exclude the manipulable regions. The dashed line in each of the plots represents the estimated counterfactual distribution $\hat{F}_s$. The shaded areas illustrate the mass of exam scores that are moved from just below to just above each of the cutoffs.

I estimate that 10.34% (standard error [s.e.] = 0.06) of all Language exams and 5.22% (s.e. = 0.04) of all Math exam scores are manipulated to meet given coarse grades. Within the range of scores that have weakly positive probability of being manipulated, I estimate that 57.94% (s.e. = 0.39) of Language exams and 35.33% (s.e. = 0.28) of Math exams are in fact manipulated.

Table 2. Manipulation Estimates by Grade Cutoff (2007-2012)

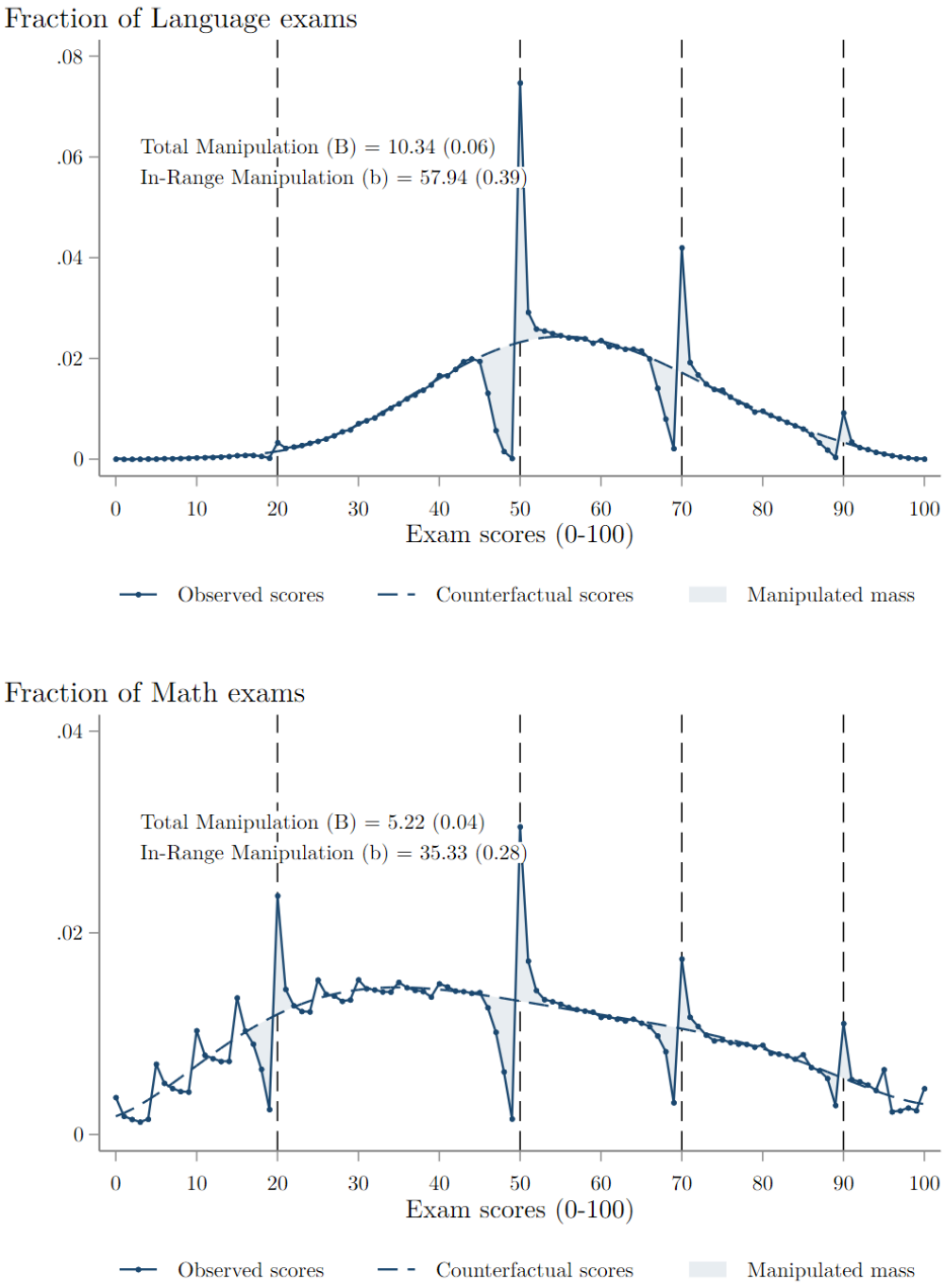
Outcome: **Percent Manipulated**	**Language**		**Math**		**Difference (p.p.)**	
	Total (1)	In-range (2)	Total (3)	In-range (4)	Total (5)	In-range (6)
Grade 1 to 2 ($\bar{s}_2$ = 20)	0.26	76.83	1.54	50.99	-1.28	25.84
	(0.01)	(4.78)	(0.03)	(0.78)	[0.00]	[0.00]
Grade 2 to 3 ($\bar{s}_3$ = 50)	6.65	66.25	2.32	39.08	4.33	27.17
	(0.06)	(0.46)	(0.03)	(0.43)	[0.00]	[0.00]
Grade 3 to 4 ($\bar{s}_4$ = 70)	3.09	48.19	1.09	26.60	2.00	21.59
	(0.03)	(0.52)	(0.03)	(0.62)	[0.00]	[0.00]
Grade 4 to 5 ($\bar{s}_5$ = 90)	0.86	45.89	0.62	25.56	0.23	20.33
	(0.03)	(1.83)	(0.02)	(0.95)	[0.00]	[0.00]
Total	10.34	57.94	5.22	35.33	5.12	22.61
	(0.06)	(0.39)	(0.04)	(0.28)	[0.00]	[0.00]
Total Number of Exams	432,367		434,432			

Notes: The table reports estimates of total and in-range manipulation in the Language (Columns 1 and 2) and Math exams (Columns 3 and 4). It also reports the difference in these estimates between subjects (Columns 5 and 6). It includes data on test scores of all exams in the sample taken between 2007 and 2012. Each row indicates the cutoff to which the estimates refer to. Section 2 describes how the estimates are constructed. Total manipulation estimates measure the percent of all test scores that were manipulated. In-range manipulation measures the percent of all test scores below the cutoff in each manipulation region that are bumped up over the cutoff. Standard errors are presented below the estimates in parentheses, computed through block bootstrap. The p-value for a two-sample z-test testing for the null of no difference between Language and Math estimates is presented in square brackets (Columns 5 and 6). See Appendix C.1 for details.

Graders' manipulation behavior also depends on the raw score. Table 2 presents estimates of manipulation close to each of the grade cutoffs. Columns 1 and 3 report estimates of the percentage of manipulated test scores in Language and Math, respectively. As expected, most manipulation occurs around the pass cutoff from grades 2 to 3. Columns 2 and 4 report in-range manipulation estimates. In the case of the Language exam, 66.25% (s.e. = 0.46) of the eligible students are bumped up across the pass cutoff ($\bar{s}_3$ = 50). On the other hand, only 39.08% (s.e. = 0.43) of Math test takers are bumped across the same cutoff. In-range manipulation is relatively large across all grades, especially for a context of blind grading. As comparison, extant evidence shows comparable total manipulation (9.86%) and lower in-range manipulation (47.3%)

distributions for each subject, pooled across years.

FIGURE 5. TEST SCORE DISTRIBUTIONS, 2007-2012



Notes. The figure shows the distribution of test scores in the Language (top panel) and Math (bottom panel) exams, taken between 2007 and 2012. I only consider the first attempt test, before any re-scoring, for each student in the sample. Each point shows the fraction of students in a given score bin of integers between 0 and 100. The dashed line is a ninth degree (top panel) or a sixth (bottom panel) degree polynomial fitting the empirical distribution, excluding scores that fall into the manipulation regions, as defined in Section 2. The vertical dashed lines represent the relevant grade cutoffs for $p \in \{2, ..., 5\}$. The shaded area illustrates the manipulated mass, either missing, if below the cutoff or in excess, if above the cutoff. Total manipulation is the percentage of exam scores that were manipulated. In-range manipulation is the percentage of test scores that are manipulated, normalized by the average height of the counterfactual distribution to the left of each cutoff. Manipulation estimates are computed on an yearly basis, and then averaged across the years. Standard errors are presented in parentheses and are computed through block bootstrap.

in Language exams in a non-blind grading context in New York City high school exit exams (Dee et al., 2019). In-range manipulation estimates also suggest that graders find it less costly to bump up the scores of students close to the two lower grade cutoffs in both subjects. Moreover, the intensity of manipulation is monotonically decreasing with the grade cutoff.

Differences across Subjects. Table 2 also reports the difference in manipulation estimates between subjects. Total manipulation is 5.12 percentage points (p.p.) larger in the Language exam, a difference which is statistically significant (p-value = 0.00)[20]. Furthermore, in-range manipulation in the Language exam is around 23 p.p. higher (p-value = 0.00) than that of the Math exam. I argue that even if graders of each subject have identical leniency preferences, grade manipulation in the Math exam can be considerably more costly than in the Language exam. In support of the latter hypothesis, I find evidence that Language graders manipulate in the essay item—typically a 30 score points question in the end of the exam, for which scoring is open to a relatively wide margin of interpretation. On the other hand, Math graders have less scope for manipulation[21].

Differences in Grading Regimes. Does the type of grade reported to students change graders' manipulation incentives? Starting in 2013, exam takers receive a different type of feedback on their exam performance. After this reform, officials report student performance through both the students' grade (p) and their granular test score (s). Prior to this reform, students had no access to their underlying granular score (see Section II). Table 3 presents in-range manipulation estimates for the period after this grade reporting reform. Columns 2 and 4 present the difference between the post- and pre-reform period in terms of this measure of grader's manipulation intensity, for Language and Math, respectively. The p-values associated with the null of no difference between these periods are reported in square brackets below. Post-reform there was a statistically significant 4.58 p.p. increase in the intensity of grader's manipulation in the case of the Language and 6.01 p.p. in the case of Math. Except for the lowest grade cutoff, in range-manipulation significantly increased across all other grade cutoffs, between the periods after and before the reform.

The observed increase in the intensity of manipulation suggests that it is less costly, on average, for graders to manipulate if the student observes the granular test score (s). However, I cannot exclude the hypothesis that the change in the intensity of manipulation is not driven by other confounding incentives to graders' behavior. Figure 6 depicts estimated in-range manipulation for each year in the sample for both Language (upper panel) and Math (lower panel) exams, and their corresponding 99% confidence intervals.

[20] To formally test the difference in manipulation estimates across subjects, I construct a z-statistic of the form:

$$z = \frac{\hat{B}_{Lang.} - \hat{B}_{Math}}{\sqrt{\frac{\widehat{s.e.}^2\left(\hat{B}_{Lang.}\right) + \widehat{s.e.}^2\left(\hat{B}_{Math}\right)}{\mathcal{J}}}},$$

where $\mathcal{J} = 200$ corresponds to the number of bootstrap samples used to draw the standard errors of the estimates.

[21] See Appendix D for an an analysis of the different manipulation margins used by graders.

Table 3. In-range Manipulation By Grade Cutoff After the Reform (2013-2015)

Outcome: **In-range manipulation**	**Language**		**Math**	
	Post-reform (1)	Diff. Post-Pre (2)	Post-reform (3)	Diff. Post-Pre (4)
Grade 1 to 2 (s = 20)	66.68	-10.15	38.34	-12.65
	(3.06)	[0.00]	(0.85)	[0.00]
Grade 2 to 3 (s = 50)	70.92	4.67	47.62	8.54
	(1.83)	[0.00]	(0.61)	[0.00]
Grade 3 to 4 (s = 70)	52.05	3.87	40.20	13.60
	(0.61)	[0.00]	(0.83)	[0.00]
Grade 4 to 5 (s = 90)	52.54	6.65	33.93	8.37
	(2.25)	[0.00]	(1.23)	[0.00]
Total	62.52	4.58	41.34	6.01
	(0.58)	[0.00]	(0.41)	[0.00]
Total Number of Exams	215,915		216,421	

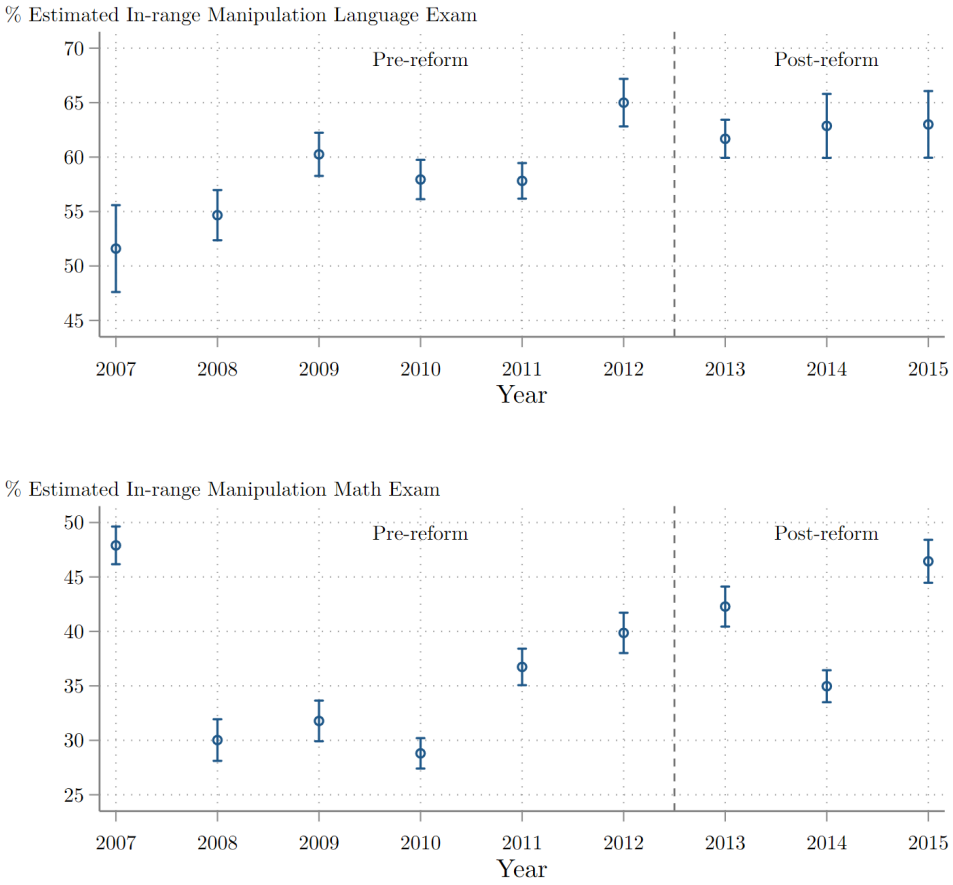
Notes: The table reports in-range manipulation estimates. Columns 1 and 3 present in-range manipulation estimates for the years between 2013 and 2015 (post-reform period), for the Language and Math exam, respectively. In-range manipulation measures the percent of all test scores below the cutoff in each manipulation region that are bumped up over the cutoff. Columns 2 and 4 report the difference in the estimates between the post- and pre-reform periods. Standard errors are presented below the estimates in parentheses, computed through block bootstrap. The p-value for a two-sample z-test testing for the null of no difference between the post-reform and the pre-reform periods is presented in square brackets (Columns 2 and 4). See Section C.1 for details.

The figure shows no noticeable discontinuity close to the the reform year, suggesting other may be driving the upward-sloping trend in in-range manipulation throughout the years.

Potential Concerns. An important concern is that the observed bunching may at least partially reflect student sorting. It could be that bunching above the cutoff is the result of a marginal additional effort from students just below in order to avoid falling in this grade level rather than graders' leniency. This type of manipulation is unlikely to be significant for at least two reasons. First, students would have to be able to predict their overall performance in the exam with a reasonable degree of accuracy. Second, conditional on their accurate prediction, students would be able to precisely manipulate their performance to be just above the cutoff. I argue that, in this case, their latent ability would be given by a raw score $s_i^* \geq \bar{s}_p$ and would therefore be always-takers of a higher grade level. For this to be true, one would have to accept that the underlying counterfactual distribution of scores is not smooth, but rather defined by these discontinuities around grade levels.

Another concern relates to the availability of re-scoring appeals. Manipulation could reflect bunching from appeals of students that are judged more leniently than their initially graded exams. I address this issue by only considering the score originally attributed to the student, before any re-scoring appeal. Furthermore, appeals in the ninth grade exam are just 0.3 and 0.1% in the Language and Math ninth grade first phase exams, respectively. Finally, the grader of a given exam is never responsible for re-scoring it, in case of an appeal. Therefore, the possibility of an appeal is not driving observed bunching or seems to

Figure 6. In-Range Manipulation by Year and Exam Subject



Notes. The figure shows in-range manipulation estimates in the Language (top panel) and Math (bottom panel) exams, taken between 2007 and 2015, and the respective 99% confidence intervals. In-range manipulation is the percentage of test scores that are manipulated, normalized by the average height of the counterfactual distribution to the left of each cutoff. The vertical dashed separates pre-reform from post-reform periods. In the pre-reform period, students are not reported their underlying test score (s), only their coarse grade (p). In the post-reform period, students are reported their both s and p. Standard errors from which the confidence intervals are derived are computed through block bootstrap.

be an important behavioral driver of graders' leniency.

V.2. Manipulation as Leniency

Manipulation Incentives. Why do graders manipulate? Extant research puts forward mainly three hypotheses. First, test-based accountability systems often lead teachers and schools to respond strategically, with some cheating and manipulating students' grades (e.g., Angrist et al., 2017; Neal and Schanzenbach, 2010; Figlio and Getzler, 2006; Jacob, 2005; Jacob and Levitt, 2003). Nonetheless, these behaviors are often localized, and mostly found among those with highest incentives. In a context of decentralized grading in New York City exit exams, widespread manipulation existed before the implementation of 'No Child Left Behind' (NCLB)—United States' high-stakes accountability system introduced in 2001. The reform did not lead to any increase in the extent of manipulation, even among schools with highest accountability pressure (Dee et al., 2019). In my setting, national exams provide some source of public accountability. In particular, school-level means in national exams determine publicly available school rankings. Parents' school and residence choices are influenced by this type of public reporting, with consistently low-performing schools being more likely to close (Nunes et al., 2015). Arguably, however, the stakes national exams for school accountability in Portugal are considerably lower than those associated with NCLB (see, e.g., Liebowitz et al., 2018). Furthermore, because grading is centralized, an incentive for grade manipulation based on school accountability concerns is unlikely[22].

Second, teachers' pay-for-performance is often argued as an important incentive for teacher behavior, performance and manipulation incentives, even when it fails to improve student achievement (e.g., Fryer, 2013). However, previous evidence finds no change in the extent of manipulation from introducing differential pay based on student achievement (Dee et al., 2019). Furthermore, in my setting—where exam grading is centralized—teacher pay and career incentives are uncoupled from student performance (see, e.g., Liebowitz et al., 2018). Therefore, individual teacher incentives are likely not driving manipulation in Portugal.

Third, graders may manipulate students to simply help them have a higher grade. In a decentralized, non-blind grading setting, Diamond and Persson (2016) find that teachers tend to bump up the grades of students who had a below-expected performance in the test. On the other hand, Dee et al. (2019) find manipulation is mostly driven by teachers' 'altruistic' motivation to make students meet given performance levels, rather than compensating for disadvantage or 'bad luck'. However, in both settings, teachers observe the students they grade. In the context of Portugal, manipulation is likely being driven by graders leniency, rather than institutional incentives. Importantly, graders do not have any caps or official minimum requirements on the the number of students that must be assigned to each grade. Furthermore, although students can plead for test re-scoring, the first grader of each exam is never responsible for

[22] School rankings were published throughout the entire period in my sample, since 2001. Unlike Dee et al. (2019), I have no time variation to empirically investigate whether there was a change in observed manipulation.

re-appreciation. Thus, incentives are not driven by complaint avoidance and potential additional work on re-scoring. In light of these facts, it is plausible to assume that at least some—lenient—graders derive utility from awarding a higher grade if given sufficient discretion.

Leniency Framework. Formally, I assume each grader observes the raw, *true* score s_i^* of each student in their randomly assigned batch of exams. In this context, the true score is the one that would prevail if the grader would strictly follow the scoring criteria in judging the student performance. Crucially, I also assume that, in the population, the empirical distribution of s_i^* would be a smooth probability mass function. Based on the utility derived from a higher grade, lenient graders score the exam at $s_i > s_i^*$, so that a student that would have grade p, is graded at level $p + 1$.

However, manipulation is potentially costly. Its cost reflects both a taste for fairness, a factoring in of detection risk, and the actual format of the exam. First, I assume graders have a psychological cost from attributing an amount of points that would decouple too much the final exam score from the perceived student performance. Second, graders may also fear detection, provided strict grading criteria. Third, exam scores may be harder to manipulate by the way they are designed. For instance, there is considerable less leeway for leniency in a fully multiple choice exam than in a long-form essay, where grading is arguably more subjective. This is consistent with the finding that manipulation is larger in the Language exam and that graders manipulate in the essay item.

Graders' manipulate until the marginal cost of manipulation equals its marginal benefit[23]. As each graders only derive utility from higher grades, it is never optimal to attribute less than the necessary points to bump up the students to a new grade ($p + 1$). By the same logic, it is never optimal to attribute more than the necessary points required to reach the new grade. Therefore, each grader defines the ranges of scores around each grade cutoff on which they will manipulate a student's test score (see Proposition 1, Appendix B). The assumptions on graders' behavior underpin a key condition for the validity of bunching methods, that of bounded manipulation (Kleven, 2016). Around each grade cutoff $\bar{s}_p$ there are bounded sets of test scores that are potentially manipulated, or—as previously defined—manipulation regions ($\mathcal{P}_p$)[24].

VI. Effects of Grade Manipulation

In Section V I presented evidence of substantial grade manipulation in the neighborhood of all grade cutoffs, consistent with graders being lenient with students at the margin of a higher grade. I now turn to investigate the impact of manipulation for those students that are manipulated on their short-term educational choices and high school achievement. I find that graders' leniency encourages mostly low-performing students to follow more demanding high school tracks. However, the change induced by a

[23] See Appendix B for a full derivation of the model.

[24] The main limitation in understanding the motivation behind manipulation in this setting is not having student-grader linked data. In particular, I cannot identify the characteristics of graders who are more likely to manipulate and to which extent. Therefore, a full characterization of blind grade manipulation is out of the scope of this paper and left for future research.

higher grade in track choices for these students is likely mostly explained by a mechanical change in grade retention. I find that low-performing students are substantially less likely to being retained in the same grade, effectively forcing these students to postpone this human capital investment decision for at least one year. Finally, I find no evidence that receiving an exogenously awarded higher grade in middle school exams leads to a consistent change in academic achievement throughout high school.

VI.1. Educational Choices

Track Choice. The choice of high school track is a major career decision for students in Portugal. Upon completing middle school, students can select among two types of high school offer: academic or vocational. Within the academic track, the scientific sub-track is the most prestigious, and with most enrolled students. Students graduating from the scientific sub-track typically enrol in highly demanded university STEM degrees or medical school.

Even though high school track choice depends only on family and student preferences, it is highly correlated with test scores. As Table 1, in Section III, shows, the choice of high school track and the scientific sub-track is strongly correlated with students exam score. Appendix E.1, Table E.8 shows further correlational evidence that having a higher exam grade is associated with a higher probability of choosing the academic track. Students with a grade level 3 instead of 2 in the Language and Math exam are, respectively, 26 p.p. (s.e. = 0.14) and 22 p.p. (s.e. = 0.14) more likely to enrol in the academic track. Furthermore, students with better grades in either exam are also significantly more likely to choose the scientific sub-track.

However, are grades in end of middle school a relevant signal of ability for students? Indeed, one can think of each individual has only having an uncertain prior about their ability in each subject, which updates upon receiving new evidence in a Bayesian fashion (as in, e.g., Li and Xia, 2022). The new evidence is the coarse grade in the exam, a noisy signal of underlying ability in the subject. If the subjective perception of academic ability informs students'—and their families'—choices of human capital investment, getting a bumped up grade in the exam may increase the likelihood of choosing the academic track in high school[25].

Table 4 presents the estimated effects on the likelihood of students selecting the high school academic track. I document estimated local average treatment effects (LATE, $\hat{\tau}_p$) and first stage coefficients ($\hat{\tau}_{T,p}$) for the Language and Math exams in Panels A and B, respectively. The reported coefficients capture the effect of being manipulated for those bumped up across each of the grade cutoffs (Columns 1 to 4), among those that completed basic education and are ever found in high school.

Without manipulation, I estimate that about 30.1% and 49.2% of the students located around the lowest grade cutoff ($\bar{s}_2 = 20$) in the Language and Math exam, respectively, would have chosen the academic track

[25] See Appendix B for a formal framework.

Table 4. LATE: Choice of Academic Track

Outcome: **Chose Academic Track (%)**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Language				
Manipulated Exam Grade	4.1	0.5	0.9	-1.4
	(3.6)	(0.6)	(0.8)	(0.9)
	[30.1]	[64.5]	[86.6]	[97.7]
First Stage	0.43	0.40	0.29	0.34
	(0.02)	(0.00)	(0.00)	(0.02)
B. Math				
Manipulated Exam Grade	5.9	2.5	0.6	1.2
	(2.0)*	(1.2)	(1.7)	(1.7)
	[49.2]	[72.5]	[87.9]	[96.9]
First Stage	0.29	0.25	0.15	0.15
	(0.00)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff in the Language (Panel A) and Math (Panel B) exam on the likelihood of choosing to attend the academic track in high school. The coefficients for the effect on the dependent variable are expressed in percentage point terms. The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012 (the period for which the granular exam score is not reported) and are ever found enrolled in high school. Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

(Table 4, Col. 1). According to the first stage estimates, 43% (in Language) and 29% (in Math), of these students have an inflated grade. For those bumped up across this grade cutoff in Language, I find that being manipulated increases the probability of choosing the academic track by 4.1 p.p. (s.e. = 3.6). However, the estimate is not statistically different from zero for a 5% significance level. For the same grade cutoff in Math, I find that manipulation significantly increases the likelihood of choosing the academic track by 5.9 p.p. (s.e. = 2). The magnitude of the effect corresponds to about 12% of the control group mean. For the cutoff between grade levels 2 and 3 ($\bar{s}_2 = 50$), I also find that having a higher Math grade leads to a 2.5 p.p. (s.e. = 1.2) increase in the likelihood of choosing the academic track (Table 4, Col. 2), although the 95% confidence interval of these estimates includes zero. I find substantially smaller and not statistically significant effects for all the other grade cutoffs.

Are there also any impacts on the choice of high school curricula? To study this question, I restrict the population of interest to all those that are ever found in the academic track in high school. Table 5 shows estimates analogous to those in Table 4 but for the choice of the scientific sub-track. Receiving a positive signal from a higher a grade in any of the exams may marginally induce some students to enrol in the more demanding scientific sub-track. Column 1 of Table 5 shows that, in a counterfactual state of the world without manipulation, a relatively high proportion of low achieving students around the first

grade cutoff still enrol in the scientific sub-track; 43% for Language and 28% for Math. However, having a higher grade in either of the exam does not seem to change the incentive for students to enrol in this curricular offer.

Table 5. LATE: Choice of Science Sub-track

Outcome: **Chose Science Sub-track** (%)	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Language				
Manipulated Exam Grade	0.3	0.0	0.6	-4.9
	(6.1)	(0.8)	(1.1)	**(1.7)***
	[43.4]	[53.3]	[65.2]	[78.1]
First Stage	0.43	0.40	0.29	0.34
	(0.01)	(0.00)	(0.00)	(0.01)
B. Math				
Manipulated Exam Grade	2.6	0.3	-0.1	3.5
	(2.5)	(1.6)	(2.6)	(3.1)
	[28.0]	[58.2]	[72.4]	[82.9]
First Stage	0.29	0.24	0.15	0.15
	(0.00)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff in the Language (Panel A) and Math (Panel B) exam on the likelihood of choosing science curriculum sub-track in high school. The coefficients for the effect on the dependent variable are expressed in percentage point terms. The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012 (the period for which the granular exam score is not reported) and are ever found in the academic track in high school. Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

For the highest grade cutoff ($\bar{s}_5 = 90$), the magnitude of the effect is relatively larger. I find that students inflated to the highest grade in the Math exam are 3.5 p.p. (s.e. = 3.1) more likely to choose the scientific sub-track, although the result is not statistically significant at a 5% level (Table 5, Panel B, Col. 4). On the other hand, being bumped up to the best grade level in the Language exam significantly decreases the probability of choosing the scientific sub-track by 4.9 p.p. (s.e. = 1.7), or by about 6% of the control mean of 78%. This evidence suggests that being bumped up to the highest grade in the Language exams diverts some of the students in the academic track to enrol in sub-tracks other than the scientific one.

Grade Retention. A channel through which a higher grade may induce a change in the choice of high school track is the timing at which this choice is made. Specifically, to the extent that grade retention decisions depend on exam grades, students may be forced to postpone their high school track decisions by at least one year due to grade retention.

In my setting, a student typically has to repeat the same grade if she fails both Math and Language, or a total of three subjects (see Section II). Furthermore, Math and Language final grades depend both on school grades (awarded by the students' teachers) and exam grades (awarded by blind graders). Therefore,

particular combinations of school and exam grades lead students to fail the subject.

Student retention is decided after both exam and subject grades are reported. The subject exam grade (p) contributes to the final subject grade (see Section II). Since student retention decisions directly depend upon final grades in both Math and Language, having a marginal, exogenous higher grade can positively impact student progression[26].

Table 6 presents the final grade in the subject by combination of the students' school grade (measured before the exam), and the exam grade. Scoring below 3 implies failing the subject. Relevantly, students with a school grade of 3 can still fail the subject if they have the lowest possible grade in the exam. However, by having a coarse grade of 2 instead of 1 in the exam, they can still pass the subject even if not passing the exam. Likewise, students with a school grade of 2 that have an exam grade of 4 (instead of 3) avoid failing the subject. Finally, for any other school grade, passing from a grade 2 to a 3 or a grade 4 to a 5 in the exam cannot possibly make a student switch its pass or fail status. Table 6 highlights that the largest effects on the retention margin are to expected to be found among students bumped up to grade 2 and to grade 4, who will be less likely to repeat.

Table 6. Final subject grade by school and exam grade combination

School Grade (70%)	Exam Grade (30%)				
	1	2	3	4	5
1	1	1	2	2	2
2	2	2	2	3	3
3	2	3	3	3	4
4	3	3	4	4	4
5	4	4	4	5	5

Notes: Each cell of the table computes the final coarse grade for each combination of school grade, as measured prior to taking the exam, and the exam score. Values in gray indicate cells for which the combination of school and exam grade represents failing the subject.

Table 7 shows the estimated effects of manipulation on student repetition in ninth grade. I document estimated local average treatment effects (LATE, $\hat{\tau}_p$) and first stage coefficients ($\hat{\tau}_{T,p}$) for the Language and Math exams in Panels A and B, respectively. The reported coefficients capture the effect of being manipulated for those bumped up across each of the grade cutoffs (Columns 1 to 4), according to the estimation method detailed in Section 2. Standard errors are block bootstrapped and presented in parentheses. In square brackets, I report the estimated dependent variable mean for the imputed control group, i.e., students outcomes in a world without manipulation.

I estimate that, in a world without manipulation, about 33.7% of the students located around the lowest

[26] Appendix E.1, Table E.7, Panel A, shows the associations between having a higher grade and grade repetition. As expected, having a higher grade in either of the exams predicts a lower likelihood of repetition. For instance, students that have a grade 2 in the Language test are 28 p.p. (s.e. = 0.71) less likely to repeat ninth grade, compared to those which just got a grade 1. Also as expected, the magnitude of these coefficients is considerably smaller on the top of the distribution, where the incidence of repetition is residual. Crucially, the OLS coefficients do not reflect the causal effect of having a higher grade in these exams.

Table 7. LATE: Ninth Grade Repetition

Outcome: **Has to Repeat Grade 9 (%)**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Language				
Manipulated Exam Grade	-12.7	-0.3	-1.0	-0.2
	(3.6)*	(0.5)	**(0.5)***	(0.9)
	[33.7]	[10.4]	[2.5]	[0.2]
First Stage	0.44	0.40	0.28	0.34
	(0.01)	(0.00)	(0.00)	(0.09)
B. Math				
Manipulated Exam Grade	-6.3	0.0	-1.2	0.6
	(1.5)*	(0.7)	(0.9)	(0.9)
	[20.7]	[6.0]	[1.3]	[0.0]
First Stage	0.29	0.24	0.15	0.15
	(0.00)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff in the Language (Panel A) and Math (Panel B) exam on the likelihood of being retained and having to repeat school grade nine. The coefficients for the effect on the dependent variable are expressed in percentage point terms. The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012 (the period for which the granular exam score is not reported). Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

Language grade cutoff ($\bar{s}_2 = 20$) would have to repeat the grade (Table 7, Panel A, Col. 1). I also estimate that about 44% of these students are bumped up across the cutoff due to grader manipulation. For those bumped up across this grade cutoff in Language I find that being manipulated significantly decreases the probability of repeating ninth grade by 12.7 p.p. (s.e. = 3.6). The magnitude of the effect corresponds to about 38% of the control group mean. For the case of Math (Table 7, Panel B, Col. 1), I find that manipulation significantly decreases the probability of repeating ninth grade by 6.3 p.p. (s.e. = 1.5), or by about 30% of the control group mean of 21%. The effects refer to the 29% of those in the manipulation region who are effectively manipulated, as indicated by the first stage estimate.

The economically significant effects I find for the lower grade cutoffs are generally not to be found for other grade cutoffs. These are either zero (e.g., Panel B, Col. 3), for the passing cutoff or statistically indistinguishable from zero. The exception is for those at the cutoff for the second to best grade. Students at the margin of a grade 4 in the Language exam (Table 7, Panel A, Col. 3) repeat the grade at relatively low levels (2.5%). I find that for the 28% of compliers in this manipulation region, being bumped up significantly decreases the probability of repeating the grade by 1 p.p. (s.e. = 0.5), equivalent to 40% of the control mean.

I also run the same analysis for the outcome of graduation from basic education in the three years

following the exam and find qualitatively similar results (Table E.11, Appendix E.3). Students bumped up above the lowest Language grade cutoff are 15.5 p.p. (s.e. = 3.4) more likely to complete basic education, compared to a control mean of 63%. Likewise, students bumped up above the lowest Math grade cutoff are 6 p.p. (s.e. = 1.4) more likely to complete basic education.

These results show that low-performing students in the exam—i.e., those in the vicinity of the lowest grade cutoff—are less likely to postpone their high school track decisions. On average, students that are encouraged to choose the academic track are also those less likely to being retained. Therefore, it is likely that the effect on academic track choice is partly explained by the change in the likelihood of being retained.

Differences in Grading Regimes. I now investigate whether these findings change with the type of signal received. Crucially, the existence and magnitude of the effect of manipulation is potentially dependent on the information available to the individual. Consider two possible scenarios: complete and incomplete information. In a scenario of incomplete information—the one of my main analysis—test takers are just informed of their coarse grade p. In a scenario of complete information, test takers are informed of both their coarse grade in the exam p, and the granular score s. Under any of the two grading regimes, students bumped up across the cutoff $\bar{s}_p$ directly benefit from a higher grade level p. However, under a complete information scenario, students observe their 'luck', or how far they were from getting a higher or lower coarse grade. In this case, one should expect the update on the marginal students' belief to be weaker than under incomplete information, where the uncertainty about underlying ability is smaller[27]. I run the same empirical analysis for the sample years in which the score is also reported to individuals, to study whether there is a differential response to a more granular signal of performance. Under complete information, students may still respond to a higher performance label, even if aware of their 'luck' (e.g., Murphy and Weinhardt, 2020; Papay et al., 2016; Main and Ost, 2014). However, I expect the magnitude of the effects to be smaller for the years where students are fully informed about their score in the exam.

I find that, in limited instances, by obtaining a positive, noisy signal of ability, students are induced to change their choice. Is there still a change in behavior in the case of a cleaner signal of performance? I run the same analysis as above for the sample period in which grade reporting includes both the exam grade level (p) and the exact exam score (s). Table E.15 in Appendix E.4 shows estimates analogous to those in Tables 4 and 5 for the period between 2013 and 2015. I find that students respond differently, depending on the grade cutoff. In the years after the reform, coefficients relative to the choice of the scientific sub-track are generally non-significant. For those at the margin of the highest grade in Language, being bumped up and reported a score close to the grade cutoff does not induce any change in the students' choice (0.1 p.p., with s.e. = 2.2). The finding contrasts with the one in Table 5, for the case in which the exam score

[27] See Appendix B.

is not reported. On the other hand, being bumped up across the lowest Math grade cutoff still leads to a higher probability of choosing the academic track among those that are ever found in high school.

I find that, for the case of grade retention, the effect is somewhat muted. Table E.13 in Appendix E.4 shows results analogous to the estimates presented in Table 7, but for the period between 2013 and 2015. For the case of Language, being bumped up across the lowest grade cutoff decreases the probability of repeating the grade by 9 p.p. (s.e. = 4) from a control mean of 50%. For the case of Math, a manipulated grade among low performing students reduces the probability of repetition by about 4 p.p. (s.e. = 2.3) from an estimated baseline of 25%, a result that is nevertheless not statistically different from zero. Results for basic education completion in the post-reform period are similarly attenuated (Appendix E.4, Table E.14).

VI.2. Academic Achievement in High School

Through which causal mechanism may an inflated grade in a ninth grade exam affect student performance in high school? A first hypothesis is that a change in the subjective perception of ability makes students adjust their learning effort. Theoretically, the effect of a positive signal of ability on subsequent effort and performance is ambiguous. On the one hand, students may display excess self-confidence in their ability (Bénabou and Tirole, 2002, 2003). A bumped up grade increases the perceived return to learning effort, which—depending on the cost of effort function—may lead individuals to exert less learning effort. Since the subjective perception of ability increases due to manipulation, while the underlying ability does not change, less learning effort can cause lower academic performance in high school. On the other hand, if cost of effort is convex, a positive signal of ability always lead individuals to increase their effort and subsequent performance (as in, e.g., Li and Xia, 2022; Delavande et al., 2022). Therefore, empirically, I have no prior on the direction of the effect[28]. Indeed, previous evidence finds mixed results on achievement (e.g. Li and Xia, 2022; Delavande et al., 2022; Azmat et al., 2019; Fischer and Wagner, 2018; Diamond and Persson, 2016; Bandiera et al., 2015).

A second potential channel is through teachers' expectations. A higher grade in the ninth grade exam could also work as a biased signal of ability to others' within the education system. Relevantly, a higher grade in the exam may change the subjective perception of ability that future high school teachers have about a student with a manipulated grade. Although I cannot reliably estimate the importance of this channel, it is unlikely that it would drive the main effects. First, the end of ninth grade represents the end of an educational stage. Enrolling in high school, students typically change school environment and take classes with an entirely new set of teachers. Second, previous exam grades and GPA are not salient in the new school environment, as these are not part of admission criteria in public high schools. Therefore, it is unlikely that high school teachers take ninth grade academic performance as a relevant prior in building potentially biased beliefs about a student's academic ability in the subject. In order to curtail this concern

[28] See Appendix B for a formal framework.

I not only look at academic performance in the school subject, but also performance in high school exit exams, which are—as the ninth grade exams—blindly evaluated by anonymous graders. However, I cannot confidently exclude any complex interplay between student effort and teacher expectations.

As expected, exam grades in the end of middle school strongly predict both school subject grades in tenth grade, as well as performance of high school exit exams (Appendix E.1, Table E.9). But does having a bumped up grade lead to any change in the future academic achievement of students? To analyze this question, I restrict the analysis to individuals that enrol in the academic track in high school and attend classes in Portuguese Language or Math.

Table 8 shows the estimated effects on performance of students in each school grade subject, measured in a integer scale between 0 and 20. I present LATE estimates ($\hat{\tau}_p$) and first stage coefficients ($\hat{\tau}_{T,p}$) for performance in the same subject (Panel A), as well as cross-effects on the other subject (Panel B). The reported coefficients capture the effect of being manipulated for those bumped up across each of the grade cutoffs (Columns 1 to 4). Across any of the grade cutoffs, I find null results. On average, having an inflated grade in the ninth grade exam does not change performance of the students during the next year, in either of the subjects. Likewise, I do not find significant results for the outcome of students' performance in high school exit exams (Table E.12, Appendix E.3).

Table 8. LATE: School Subject Performance in Tenth Grade

Outcome: **Subject Grade Next Year (0-20)**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Effect on the Same Subject				
A.1. Language				
Manipulated Exam Grade	-0.11	0.03	-0.03	-0.15
	(0.24)	(0.04)	(0.05)	(0.10)
	[10.2]	[11.4]	[13.4]	[16.2]
First Stage	0.43	0.40	0.28	0.33
	(0.04)	(0.00)	(0.00)	(0.01)
A.2. Math				
Manipulated Exam Grade	-0.36	-0.01	0.38	0.33
	(0.20)	(0.10)	(0.16)	(0.22)
	[9.8]	[10.3]	[12.0]	[15.4]
First Stage	0.29	0.25	0.15	0.15
	(0.01)	(0.00)	(0.00)	(0.29)
B. Effect on the Other Subject				
B.1. Language				
Manipulated Exam Grade	0.10	0.00	0.07	-0.14
	(0.43)	(0.06)	(0.08)	(0.17)
	[10.2]	[10.5]	[12.4]	[15.7]
First Stage	0.47	0.39	0.28	0.34
	(0.04)	(0.00)	(0.00)	(0.01)
B.2. Math				
Manipulated Exam Grade	-0.05	0.09	-0.05	-0.08
	(0.12)	(0.07)	(0.13)	(0.18)
	[11.1]	[12.0]	[13.1]	[15.2]
First Stage	0.30	0.24	0.15	0.15
	(0.01)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff in the Language (Panels A.1 and B.1) and Math (Panel A.2 and B.2) exam on the grade in the same subject (Panel A) or the other subject (Panel B) in school grade 10. The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012 (the period for which the granular exam score is not reported) and are ever found taking Language or Math courses in high school. Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

VII. Who is Manipulated?

So far I have shown the extent of manipulation, as well as its effects—or lack thereof—on students' choices and future academic performance. Importantly, the estimated effects are local to students who are effectively bumped up across grade cutoffs. But who are these students? Should one expect to see a specific type of student being more likely to have an inflated grade even in a context where both graders and students are blind to each others' identities? My empirical setting provides a natural laboratory to study these questions. Whereas graders do not directly observe students, I have information on several of their characteristics. Therefore, I am able to provide a characterization of compliers, and how they differ from the average eligible student.

There are a number of hypotheses for why one could observe this type of grader bias. First, graders may infer student characteristics from the exam. For instance, a grader may be able to identify a student as female by her handwriting or gender-specific pronouns used in open-response items, deciding to bump up her score making use of this information. Second, abler students may convey their ability in skills that are not to be scored. For instance, graders may be less willing to be lenient with those that are marginally worse in structuring and organizing their answers, even if this should not marginally affect their test score.

To investigate whether selective manipulation exists in this context, I use the methods described in Section 2. I separately estimate the observable characteristics of students who are bumped up (the compliers) and those that could have be chosen for being bumped up (compliers and never-takers). The difference between these two quantities provides the measure of selective manipulation on that particular characteristic.

Table 9 presents selective manipulation estimates on the Language exam for a series of student characteristics measured in the beginning of ninth grade, between 2007 and 2012. Each row of the table indicates the characteristic to which the estimates refer to. Each column identifies the relevant manipulation region. In square brackets I report the estimated mean among all students who are eligible for manipulation, which comprises both the group of compliers and never-takers. As hypothesized, I find that manipulated students are more likely to be female. Around the lowest grade cutoff ($\bar{s}_2 = 20$), about 30% of eligile students are female. I estimate that those students that have an inflated grade are 1.3 percentage points (p.p.) more likely to be female. However, the estimate is imprecise and I cannot reject the hypothesis that the effect is null for a 5% significance level. Around the next two cutoffs, $\bar{s}_3 = 50$ and $\bar{s}_4 = 70$, I find that girls are 1.06 p.p. (s.e. = 0.26) and 1.37 p.p. (s.e. = 0.41) more likely to be bumped up. Interestingly, these findings contrast with evidence from non-blind grading in Sweden, in which teachers do not selectively manipulate based on attributes such as gender or race (Diamond and Persson, 2016).

Positive selection based on gender seems to be restricted to the case of the Language exam. Table E.10

in Appendix E.4 shows analogous results to the ones in Table 9 for the case of he Math exam. I find that, in the case of the Math exam, coefficients are generally small and statistically insignificant.

Are students selectively manipulated on other demographics? I find that immigrants are 0.64 p.p. (s.e. = 0.17) significantly less likely to be bumped up across the Language cutoff from exam grade 2 to 3 ($\bar{s}_4 = 70$). But I find no significant effects at any of the other cutoff in the case of immigrant students. Similarly, I find no evidence of selective manipulation of immigrant students in the case of the Math exam (Table E.10, Appendix E.4).

Table 9. Selective Manipulation by Grade Cutoff in the Language Exam

Outcome: **Selective Manipulation**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
Female	1.30	1.06	1.37	-0.19
	(1.49)	**(0.26)***	**(0.41)***	(0.91)
	[30.4]	[47.9]	[59.4]	[68.9]
Immigrant	-0.96	-0.14	-0.64	0.26
	(1.05)	(0.14)	**(0.17)***	(0.42)
	[12.8]	[6.4]	[4.9]	[3.4]
Free or Reduced Price Lunch	2.40	0.87	0.59	-0.19
	(6.41)	**(0.30)***	(0.39)	(0.65)
	[56.9]	[39.0]	[25.4]	[13.7]
Higher Education at Home	0.17	-0.31	-0.20	0.40
	(0.86)	**(0.17)***	(0.38)	(1.26)
	[4.8]	[9.8]	[20.3]	[40.7]
Unemployment in Household	-0.77	0.06	-0.14	-0.06
	(1.17)	(0.20)	(0.28)	(0.49)
	[16.7]	[14.1]	[11.3]	[7.5]
School Grade (Language)	0.01	0.02	0.05	0.04
	(0.02)	**(0.00)***	**(0.02)***	**(0.02)***
	[2.5]	[2.9]	[3.5]	[4.4]
School Grade (Math)	-0.04	0.01	0.05	0.07
	(0.02)	**(0.01)***	**(0.01)***	**(0.02)***
	[2.5]	[2.7]	[3.4]	[4.4]

Notes: The table reports selective manipulation estimates in the Language exam as the difference between the estimated characteristics of the group of manipulated students and the estimated characteristics of the group of students that could have been manipulated (eligible), for years 2007 through 2012. Each column indicates the manipulation region around the threshold to which the estimates refer to. Each row indicates the characteristic to which the estimate refers to. Coefficients associated with dummy variables are expressed in percentage points terms. In parentheses, the table presents block bootstrap standard errors. In square brackets, the table presents the estimated mean of the characteristic of students eligible for being manipulated close to each grade cutoff. For instance, I estimate that 30.4 percent of students that could have been manipulated close to the lowest proficiency threshold are female. The group of students that is bumped up across this threshold is estimated to have 1.3 percentage points more female students than among the eligible. I use the scores of students outside of the manipulation regions to estimate the characteristic, at any test score bin inside the manipulation regions. Section 2 describes how the estimates are constructed. See Appendix C.4 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the 95% confidence intervals of the estimate do not include zero.

For students falling around the pass grade cutoff ($\bar{s}_3 = 50$), I find that those who benefit from free or

reduced price lunch are somewhat positively selected. For the higher cutoff, manipulated students are 0.87 p.p. more likely to benefit from social support, compared to an estimated average of 39% among the control group. Furthermore, students from households with some level of higher education are 0.31 p.p. (s.e. = 0.17) less likely to be bumped up across the same Language cutoff (Table 9, Column 2). Although the magnitude of these effects is small, the results seem to indicate that Language graders are somehow able to identify and bump up students from relatively disadvantaged socioeconomic status. On the other hand, I find small, statistically non-significant effects in the case of the Math exam (Table E.10, Appendix E.2), failing to reject the hypothesis of identical characteristics between manipulated students and those eligible for being manipulated.

Finally, I find some evidence supporting the idea that better students—as measured by their grades in school prior to the exam—are being positively selected, although marginally. For instance, across the pass cutoff ($\bar{s}_3 = 50$), bumped up students have, on average, 0.02 (s.e. = 0.00) more grade points in Language than the 2.9 mean of eligible students, a difference of only 0.68% relative to the control (Table 9, Column 2). Likewise, I find statistically significant effects of similar magnitude for the case of Math grades, as well as for other cutoffs. In the case of bumped up students in the Math exam, I find results of similar magnitude (Table E.10, Appendix E.2). This evidence is consistent with the hypothesis that abler students are able to convey their ability in skills that are not to be scored. However, the magnitude of the effect is relatively small, as expected. In particular, the magnitude of this type of selection contrast with effects found in a context of non-blind grading, where students awarded inflated grades were much more likely to have higher previous achievement (Diamond and Persson, 2016).

VIII. Conclusion

Extant literature has found ambiguous results on the effect of getting a higher grade in high stakes standardized tests on students' future outcomes. I contribute to this literature by studying the effect of receiving an exogenous positive signal of ability, keeping the level of human capital unchanged. Exploiting blind test score manipulation, I find only limited evidence that receiving a higher grade in Math or Language have medium- to long-term consequences on students' schooling outcomes.

I start by documenting large and significant test score manipulation close to relevant grade cutoffs. These findings shed light on the motivations for teacher discretion in grading. All extant evidence of large scale manipulation originates from contexts of decentralized grading, where graders observe students. The extent of this manipulation in a context of blind grading suggests that grader leniency may also rest on a motivation to not let students—even if anonymous—fail to attain a higher grade. The results demonstrate that decentralizing the grading of high-stakes exams from teachers in schools to graders blind to who the students are is no sufficient condition to make the grading system impermeable to manipulation.

I find that low performing manipulated students in the Math exam are encouraged to follow the aca-

demic track in high school. However, the students that are bumped up from the lowest grade level in either the Math or Language exam have a significantly lower likelihood of repeating the last grade of middle school. On the other hand, high performing students who are bumped up to the highest grade in the Language exam are discouraged from choosing a science curriculum in high school. On the other hand, I find that having an inflated grade produces no consistent changes in students' future achievement.

Although graders do not observe test taker characteristics, I find that female are significantly more likely to being manipulated in the Language exam. On the other hand, I do not find evidence that graders are more lenient with female students in the case of the Math exam. Finally, I find some evidence supporting the hypothesis that better students—as measured by their grades in school prior to the exam—are being positively selected, although marginally.

References

Anelli, M. (2020): "The Returns to Elite University Education: a Quasi-Experimental Analysis," *Journal of the European Economic Association*, 18, 2824–2868.

Angrist, J. D., E. Battistin, and D. Vuri (2017): "In a small moment: Class size and moral hazard in the Italian Mezzogiorno," *American Economic Journal: Applied Economics*, 9, 216–249.

Apperson, J., C. Bueno, and T. R. Sass (2016): "Do the Cheated Ever Prosper? The Long-Run Effects of Test-Score Manipulation by Teachers on Student Outcomes," *CALDER Working Paper Series Nr. 155.*

Avery, C., O. Gurantz, M. Hurwitz, J. Smith, J. Howell, J. Buckley, T. Packer, and M. Reyes (2018): "Shifting College Majors in Response to Advanced Placement Exam Scores," *Journal of Human Resources*, 53, 918–956.

Azmat, G., M. Bagues, A. Cabrales, and N. Iriberri (2019): "What you don't know. . . can't hurt you? A natural field experiment on relative performance feedback in higher education," *Management Science*, 65, 3714–3736.

Azmat, G. and N. Iriberri (2010): "The importance of relative performance feedback information: Evidence from a natural experiment using high school students," *Journal of Public Economics*, 94, 435–452.

Bandiera, O., V. Larcinese, and I. Rasul (2015): "Blissful ignorance? A natural experiment on the effect of feedback on students' performance," *Labour Economics*, 34, 13–25.

Bénabou, R. and J. Tirole (2002): "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 117, 871–915.

——— (2003): "Intrinsic and extrinsic motivation," *Review of Economic Studies*, 70, 489–520.

Bobba, M. and V. Frisancho (2022): "Self-perceptions about academic achievement: Evidence from Mexico City," *Journal of Econometrics*, 231, 58–73.

Borcan, O., M. Lindahl, and A. Mitrut (2017): "Fighting corruption in education: What works and who benefits?" *American Economic Journal: Economic Policy*, 9, 180–209.

Brade, R., O. Himmler, and R. Jäckle (2022): "Relative performance feedback and the effects of being above average — field experiment and replication," *Economics of Education Review*, 89, 102268.

Canaan, S. and P. Mouganie (2018): "Returns to education quality for low-skilled students: Evidence from a discontinuity," *Journal of Labor Economics*, 36, 395–436.

Chen, Z., Z. Liu, J. C. S. Serrato, and D. Y. Xu (2021): "Notching R&D investment with corporate income tax cuts in China," *American Economic Review*, 111, 2065–2100.

Chetty, R., J. Friedman, T. Olsen, L. Pistaferri, and J. Friedman (2010): "Adjustment Costs, Firm Responses, and Labor Supply Elasticities: Evidence from Danish Tax Records," *Quarterly Journal of Economics*, 126, 749–804.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star," *The Quarterly Journal of Economics*, 126, 1593–1660.

Coviello, D., A. Guglielmo, C. Lotti, and G. Spagnolo (2022): "Procurement with Manipulation," *CEPR Discussion Paper No. DP17063*.

Dee, T. S., W. Dobbie, B. A. Jacob, and J. Rockoff (2019): "The causes and consequences of test score manipulation: Evidence from the New York regents examinations," *American Economic Journal: Applied Economics*, 11, 382–423.

Delavande, A., E. Del Bono, A. Holford, S. Sen, and V. Lesic (2022): "Expectations about the Productivity of Effort and Academic Outcomes: Evidence from a Randomized Information Intervention," *Working Paper*, 2015.

Diamond, R. and P. Persson (2016): "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests," *NBER Working Paper No. w22207*.

Dobrescu, L. I., M. Faravelli, R. Megalokonomou, and A. Motta (2021): "Relative Performance Feedback in Education: Evidence from a Randomised Controlled Trial," *The Economic Journal*, 131, 3145–3181.

Ertac, S. (2005): "Social Comparisons and Optimal Information Revelation: Theory and Experiments," *Working Paper*.

Figlio, D. N. and L. S. Getzler (2006): "Accountability, Ability and Disability: Gaming the System?" in *Improving School Accountability (Advances in Applied Microeconomics, Volume 14)*, ed. by T. J. Gronberg and D. W. Jansen, Emerald Group Publishing Limited, 35–49.

Fischer, M. and V. Wagner (2018): "Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment," *IZA Discussion Paper No. 11970*.

Fryer, R. G. (2013): "Teacher incentives and student achievement: Evidence from New York City public schools," *Journal of Labor Economics*, 31, 373–407.

Gerard, F., M. Rokkanen, and C. Rothe (2020): "Bounds on treatment effects in regression discontinuity designs with a manipulated running variable," *Quantitative Economics*, 11, 839–870.

Goodman, J., M. Hurwitz, and J. Smith (2017): "Access to 4-year public colleges and degree completion," *Journal of Labor Economics*, 35, 829–867.

Goulas, S. and R. Megalokonomou (2021): "Knowing who you actually are: The effect of feedback on short- and longer-term outcomes," *Journal of Economic Behavior and Organization*, 183, 589–615.

Hanushek, E. A., G. Schwerdt, S. Wiederhold, and L. Woessmann (2015): "Returns to skills around the world: Evidence from PIAAC," *European Economic Review*, 73, 103–130.

Heckman, J. J., J. E. Humphries, and G. Veramendi (2018): "Returns to education: The causal effects of education on earnings, health, and smoking," *Journal of Political Economy*, 126, S197–S246.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Imbens, G. W. and D. B. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, 2011, Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47, 5–86.

Ishihara, T. and M. Sawada (2022): "Manipulation-Robust Regression Discontinuity Designs," *arXiv:2009.07551v3*.

Jacob, B. A. (2005): "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools," *Journal of Public Economics*, 89, 761–796.

Jacob, B. A. and S. D. Levitt (2003): "Rotten apples: An investigation of the prevalence and predictors of teacher cheating," *Quarterly Journal of Economics*, 118, 843–877.

Kajitani, S., K. Morimoto, and S. Suzuki (2020): "Information feedback in relative grading: Evidence from a field experiment," *PLoS ONE*, 15.

Kinne, L. (2022): "Good or Bad News First? The Effect of Feedback Order on Motivation and Performance," *Job Market Paper*.

Kleven, H. J. (2016): "Bunching," *Annual Review of Economics*, 8, 435–464.

Kleven, H. J. and M. Waseem (2013): "Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from Pakistan," *Quarterly Journal of Economics*, 128, 669–723.

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

Li, H. and X. Xia (2022): "Grades as Signals of Comparative Advantage: How Letter Grades Affect Major Choices," *SSRN Electronic Journal*.

Liebowitz, D., P. González, E. Hooge, and G. Lima (2018): *OECD Reviews of School Resources: Portugal 2018*, OECD Reviews of School Resources, Paris: OECD Publishing.

Main, J. B. and B. Ost (2014): "The impact of letter grades on student effort, course selection, and major choice: A regression-discontinuity analysis," *Journal of Economic Education*, 45, 1–10.

Manski, C. F. (1989): "Schooling as experimentation: a reappraisal of the postsecondary dropout phenomenon," *Economics of Education Review*, 8, 305–312.

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142, 698–714.

McEwan, P. J., S. Rogers, and A. Weerapana (2021): "Grade Sensitivity and the Economics Major at a Women's College," *AEA Papers and Proceedings*, 111, 102–106.

Murphy, R. and F. Weinhardt (2020): "Top of the Class: The Importance of Ordinal Rank," *Review of Economic Studies*, 87, 2777–2826.

Neal, D. and D. W. Schanzenbach (2010): "Left Behind By Design : Proficiency Counts and Test-Based Accountability," *The Review of Economics and Statistics*, 92, 263–283.

Nunes, L. C., A. B. Reis, and C. Seabra (2015): "The publication of school rankings: A step toward increased accountability?" *Economics of Education Review*, 49, 15–23.

Oreopoulos, P. and K. G. Salvanes (2011): "Priceless: The nonpecuniary benefits of schooling," *Journal of Economic Perspectives*, 25, 159–184.

Papay, J. P., R. J. Murnane, and J. B. Willett (2016): "The Impact of Test Score Labels on Human-Capital Investment Decisions," *Journal of Human Resources*, 51, 357–388.

Saez, E. (2010): "Do taxpayers bunch at kink points?" *American Economic Journal: Economic Policy*, 2, 180–212.

Smith, J., M. Hurwitz, and C. Avery (2017): "Giving college credit where it is due: Advanced placement exam scores and college outcomes," *Journal of Labor Economics*, 35, 67–147.

Stange, K. M. (2012): "An Empirical Investigation of the Option Value of College Enrollment," *American Economic Journal: Applied Economics*, 4, 49–84.

Stinebrickner, T. and R. Stinebrickner (2012): "Learning about academic ability and the college dropout decision," *Journal of Labor Economics*, 30, 707–748.

Tan, B. J. (2022): "The Consequences of Letter Grades on Labor Market Outcomes and Student Behavior," *Journal of Labor Economics*, [Online].

Tran, A. and R. Zeckhauser (2012): "Rank as an inherent incentive: Evidence from a field experiment," *Journal of Public Economics*, 96, 645–650.

Zax, J. S. and D. I. Rees (2002): "IQ, academic performance, environment, and earnings," *Review of Economics and Statistics*, 84, 600–616.

APPENDICES

A. The Data

The data for this paper relies on a combination of multiple administrative data files. Below I describe all relevant information about the cleaning and coding of the variables used in the analysis.

A.1. Data Sources

Test Scores: The test score data is organized at the student-by-test-by-phase level. Students can sit the exam in a first or second evaluation phase, which are different exams at different dates. Each observation includes a unique student identifier and information on the score points awarded, whether the student asked for a re-scoring of the test, and what was the new re-score level. The data are available for all students (enrolled in either public or private schools), in mainland Portugal, that took a national exam between the academic years of 2006-2007 and 2017-2018. The file contains data for all available subjects and grades. It includes fourth, sixth, ninth, eleventh and twelfth grade high- and low-stakes national exams taking place during this time period. The data is provided by the *Júri Nacional de Exames*, a public platform, under the indirect purview of the Ministry of Education, responsible for administering, registering, monitoring and recording national exam grades. Anonymized data on exam grades is made publicly available on the platform's website. I use a restricted-use version of this dataset, that allows me to merge it with other data files.

Enrolment Data: The enrolment data is organized at the student-by-year level. Each record contains a unique student identifier and information on students' birth date, gender, country of origin, address, parent's education and employment status, free and reduced-price lunch eligibility, access to computer and Internet at home, school, class, curriculum, grade and in-class final grade per subject. The student identifier allows to track individuals throughout classes, grades and schools across years, which enables me to collect additional information about their educational career such as grade retention, graduation and curricular track choices. Attrition in the data may occur for a few reasons: If the student moves abroad, drops from the education system altogether, dies, or the matching algorithm is unable to correctly assign the unique identifier to new instance of the same student in the system. However, in following students, attrition rates for different cohorts are relatively limited (below 10% for follow-up periods of ten years). Enrolment data comes from MISI-PUB, MISI-PRIV and INQ-PRIV, administrative datasets with information on every student enrolled in public, publicly-funded private and private schools in mainland Portugal. Information on socio-economic characteristics of students enrolled in private schools, however, is substantially more limited. The data is available between the academic years of 2006-2007 and 2017-2018.

Data Access: Access to the data is restricted. All data used in this paper is hosted in a server at a safe center in Nova School of Business and Economics, located in Lisbon, Portugal. As of now, accessing this information requires the researcher to be physically present in the safe center. Access requests can

be sent to alice.caetano@novasbe.pt.

A.2. Sample Restrictions

I make the following sample restrictions to the final dataset:

1. Only include students enrolled in regular curricular options in public schools. This excludes students in second-chance adult programs and alternative curricular pathways. I choose to only include students enrolled in public schools at the time of the exam as it is for these students that there is more available information in their background characteristics.

2. Only include students that sat at least one of the two following exams: Ninth grade Portuguese Language exam and ninth grade Math exam. Not all students take these exams (about 92% do take the first phase exams). Students with special education needs may take other exams adapted to their learning requirements, which I do not consider. Other students take special Language exams, if Portuguese is not their native language. Finally, some students may just skip that exam phase (or any exam phase).

3. In order to be able to track students throughout high school (tenth to twelfth grade), I further restrict the sample to students that took their exams between the academic years of 2006-2007 and 2014-2015.

4. I only take into account exams that were sat during the first phase to avoid bunching around grade cutoffs due to exam re-taking. As the test score I take the score originally attributed by the first grader that evaluated the exam to avoid bunching around grade cutoffs due to re-scoring. In total, I consider information from a total of $652,829$ individuals, and $1,299,081$ exams—respectively, an average of over 72.5 thousand individuals, and 144 thousand exams per year.

A.3. Data Cleaning

From the restricted sample of students I take the following series of steps until arriving to the dataset with which I run the analysis:

1. In the enrolment dataset, I identify the students that took at least one of the ninth grade Language or Math exams in each year. Identify each individual student outcomes at the end of ninth grade and throughout high school. I also identify each available student characteristic measured at the beginning of the ninth grade. Merge this data with the test score data.

2. Collapse the data to the score-by-year-by-test level. For instance, for each score bin of the Language exam in 2007, I have the frequency of exams in each score, as well as the mean value of each outcome and student characteristic in that score bin, such as percentage of students that choose the academic track by the end of ninth grade or the percentage of female students in each score-by-year cell.

3. For the pooled estimates, I collapse the data to the score-by-test level, using frequency counts and the mean of each variable in each cell.

4. For imputation of counterfactual outcomes and student characteristics I analyze the data at the individual-by-test level.

A.4. Additional Descriptive Statistics

Table A.1 shows descriptive statistics of students' socio-economic characteristics by two groups: Exam takers and exam non-takers. About 91% of the students in the sample take at least one of the ninth grade exams. Column 5 shows the difference in that particular characteristic between students that sat at least one of the ninth grade exams and those that did not sit neither. Columns 6 and 7 present the variance for each of the characteristics among exam takers and exam non-takers, respectively. Column 8 presents the absolute standardized difference across groups for each variable. As expected, students that do not take the exams are typically boys, immigrants, older and from a more disadvantaged socio-economic background as measured by parental education and unemployment status as well as eligibility to social support, through reduced price or free lunch.

Table A.1. Socio-economic Characteristics by Exam Taking Status

	Exam Takers		Exam Non-Takers		Diff.	Var. ET	Var. ENT	Std. Diff.
	Obs.	Mean / %	Obs.	Mean / %				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Socio-economic Characteristics								
Female	652,829	52.3	60,499	45.5	6.9	24.9	24.8	0.97
Immigrant	652,829	6.0	60,497	21.5	-15.5	5.7	16.9	3.26
Age	652,829	14.5	60,499	15.7	-1.1	0.5	1.3	0.84
Free or Reduced Price Lunch	652,829	34.6	60,499	50.0	-15.5	22.6	25.0	2.24
Unemployment in the Household	639,187	14.4	55,582	19.6	-5.2	12.3	15.8	0.99
Higher Education in the Household	627,292	17.3	52,156	10.3	7.0	14.3	9.2	1.45
Baseline School Performance								
School Grade (Language) [1-5]	551,012	3.2	16,257	2.6	0.6	0.5	0.6	0.56
School Grade (Math) [1-5]	555,965	3.0	16,244	2.5	0.6	0.8	0.7	0.46
Outcomes								
Repeated Ninth Grade	652,829	8.9	60,499	13.5	-4.6	8.9	11.7	1.01
Graduated from Ninth Grade	649,757	90.5	54,788	68.9	21.6	90.5	21.4	2.04
Chose Academic Track	626,828	72.0	32,552	28.6	43.4	72.0	20.4	4.51
Chose Science Sub-track	451,019	58.2	9,301	43.4	14.9	58.2	24.6	1.63
School Performance in Tenth Grade (Language) [1-20]	469,693	12.5	10,196	11.4	1.1	12.5	6.0	0.26
School Performance in Tenth Grade (Math) [1-20]	327,495	11.8	6,969	11.3	0.6	11.8	9.4	0.12
Exam Score Percentile Twelfth Grade (Language)	394,905	49.6	3,017	33.1	16.5	49.6	877.2	0.54
Exam Score Percentile Twelfth Grade (Math)	228,550	50.2	1,009	43.2	7.1	50.2	944.2	0.22

Notes: The table reports the socio-economic characteristics and outcomes of exam takers and exam non-takers, as well as their differences. Columns 1 through 4 present the number of non-missing observations and mean of each characteristic among exam takers and exam non-takers, respectively. Column 5 documents the difference between Column 2 and Column 4. Columns 6 and 7 present the variance for each of the characteristics among exam takers (ET) and exam non-takers (ENT), respectively. Column 8 presents the absolute standardized difference across groups for each variable. The absolute standardized difference is computed as $\frac{X_{ET}-X_{ENT}}{\left(\sqrt{\sigma^2_{ET}+\sigma^2_{ENT}}\right)}$, where σ^2 represents the sample variance in the characteristic X in each group, as recommended in Imbens and Wooldridge (2009). Absolute standardized differences below the rule-of-thumb value of 0.25 typically indicate that the characteristic is balanced across groups (Imbens and Rubin, 2015).

Table A.2 shows the same descriptive statistics for the difference between the full sample of students and only exam takers. In this case, the two groups are relatively similar. The absolute standardized differences between the full sample of students and only exam takers are all below the rule of thumb threshold of 0.25 (Imbens and Rubin, 2015), except for immigration, for which there is a small difference in magnitude. Exam takers are much likelier to be a selected portion of the population in this characteristic. This is not surprising as many immigrant children are not eligible to sit these exams, namely those for which Portuguese is not their mother tongue.

Table A.2. Socio-economic Characteristics by Full Sample and only Exam Takers

	Full Sample		Exam Takers		Diff.	Var. FS	Var. ET	Std. Diff.
	Obs.	Mean / %	Obs.	Mean / %				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Socio-economic Characteristics								
Female	713,318	51.8	652,829	52.3	-0.6	25.0	24.9	0.08
Immigrant	713,316	7.3	652,829	6.0	1.3	6.8	5.7	0.37
Age	713,318	14.6	652,829	14.5	0.1	0.7	0.5	0.09
Free or Reduced Price Lunch	713,318	35.9	652,829	34.6	1.3	23.0	22.6	0.19
Unemployment in the Household	694,759	14.8	639,187	14.4	0.4	12.6	12.3	0.08
Higher Education in the Household	679,438	16.8	627,292	17.3	-0.5	14.0	14.3	0.10
							0.0	
Baseline School Performance							0.0	
School Grade (Language) [1-5]	567,260	3.2	551,012	3.2	0.0	0.5	0.5	0.02
School Grade (Math) [1-5]	572,199	3.0	555,965	3.0	0.0	0.8	0.8	0.01
							0.0	
Outcomes							0.0	
Repeated Ninth Grade	704,535	9.3	652,829	8.9	0.4	9.9	8.9	0.09
Graduated from Ninth Grade	713,318	88.9	649,757	90.5	-1.7	8.4	90.5	0.17
Chose Academic Track	659,380	69.8	626,828	72.0	-2.1	21.1	72.0	0.22
Chose Science Sub-track	460,320	57.9	451,019	58.2	-0.3	24.4	58.2	0.03
School Performance in Tenth Grade (Language) [1-20]	479,889	12.5	469,693	12.5	0.0	6.9	12.5	0.01
School Performance in Tenth Grade (Math) [1-20]	334,464	11.8	327,495	11.8	0.0	12.9	11.8	0.00
Exam Score Percentile Twelfth Grade (Language)	397,922	49.5	394,905	49.6	-0.1	829.2	49.6	0.00
Exam Score Percentile Twelfth Grade (Math)	229,559	50.2	228,550	50.2	0.0	849.2	50.2	0.00

Notes: The table reports the socio-economic characteristics and outcomes of all students in the full sample before restrictions and exam takers, as well as their differences. Columns 1 through 4 present the number of non-missing observations and mean of each characteristic among full sample and exam takers, respectively. Column 5 documents the difference between Column 2 and Column 4. Columns 6 and 7 present the variance for each of the characteristics among all students (FS) and only exam takers (ET), respectively. Column 8 presents the absolute standardized difference across groups for each variable. The absolute standardized difference is computed as $\frac{X_{ET}-X_{ENT}}{\left(\sqrt{\sigma^2_{FS}+\sigma^2_{ET}}\right)}$, where σ^2 represents the sample variance in the characteristic X in each group, as recommended in Imbens and Wooldridge (2009). Absolute standardized differences below the rule-of-thumb value of 0.25 typically indicate that the characteristic is balanced across groups (Imbens and Rubin, 2015).

Table A.3 reports the intervals of discrete scores corresponding to each of the manipulation regions. The table provides details on the range of potentially manipulated scores by year and test subject. Within each test subject and manipulation region, there is some variation across the years. The variation in the length of the manipulation regions reflects idiosyncrasies of each exam, even for a fixed level of graders' leniency. With a varying cost of manipulation across the years, one should expect the length of the manipulation regions to vary.

Table A.4 presents the number of observations in the sample of exam takers, by test subject and year in which the exam was taken.

Table A.3. Manipulation Regions by Year and Test Subject

	Language				Math			
Manipulation Region	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Year								
2007	-	[46,51]	[67,71]	[87,91]	[18,20]	[45,52]	[67,71]	[88,90]
2008	-	[46,52]	[67,72]	[86,91]	[18,20]	[46,53]	[66,72]	[86,90]
2009	[18,20]	[45,53]	[67,72]	[86,91]	[19,20]	[46,51]	[67,73]	[86,90]
2010	[19,20]	[46,51]	[66,72]	[86,91]	[18,20]	[45,52]	[65,71]	[86,91]
2011	[18,20]	[45,52]	[66,72]	[87,90]	[17,22]	[46,51]	[66,72]	[86,93]
2012	[18,20]	[46,52]	[66,72]	[86,91]	[16,22]	[46,52]	[67,72]	[88,92]
2013	[17,20]	[45,52]	[66,72]	[86,91]	[17,22]	[46,52]	[66,72]	[88,91]
2014	[17,20]	[46,52]	[66,71]	[86,91]	[16,20]	[46,51]	[67,71]	[86,91]
2015	-	[46,52]	[66,71]	[86,91]	[17,21]	[46,51]	[67,71]	[86,92]

Notes: The table reports the intervals of discrete scores corresponding to each manipulation region, by year and test subject in the analysis sample. The intervals of potentially manipulated scores are determined by visual inspection. Empty cells indicate that there was no manipulation detected for that particular year-region-test.

Table A.4. Number of Observations by Year and Test Subject

Number of Exams	**Language**	**Math**
	(1)	(2)
Year		
2007	80,130	80,075
2008	72,723	73,052
2009	70,288	70,547
2010	68,978	69,471
2011	69,449	70,005
2012	70,799	71,282
2013	73,894	73,906
2014	71,105	71,314
2015	70,884	71,179
Total Observations	648,250	650,831

Notes: The table reports the number of observations by year, in each of the test subject samples.

B. Conceptual Framework

B.1. Setup

Each student i sits an exam in subject j. I assume that before the exam in subject j is taken, students have incomplete information about their level of ability, only observing its underlying distribution, which is individual- and subject-specific:

$$a_{ij} \sim N\left(\alpha_{ij}, \sigma^2_{ij,a}\right) \tag{18}$$

In the absence of any grading manipulation or bias, the score of a given student in an exam is given by:

$$s^*_{ij} = a_{ij} + \varepsilon_{ij} \tag{19}$$

Where $\varepsilon_{ij} \sim N\left(0, \sigma^2_{j,\varepsilon}\right)$ is an idiosyncratic error term that depends on factors other than ability that independently affect performance in the exam. I assume that these factors are uncorrelated with academic ability in the subject $\left(a_{ij} \perp\!\!\!\perp \varepsilon_{ij}\right)$. Therefore, each student believes that $s^*_{ij} \sim N\left(\alpha_{ij}, \sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}\right)$.

For now, and without loss of generality, I assume that students can get only one of two coarse grades: high ($p = H$) and low ($p = L$), which depends upon their score (s) in the exam.

After receiving the grade, student i has to choose whether to enrol in an academically-oriented track in high school. I assume that the decision to enroll depends on a latent variable capturing a preference for academic studies in high school (A^*). Student i expected academic preference upon graduation from middle school is given by :

$$\mathbb{E}\left[A^*_i | p_j\right] = \sum_e \gamma_j \cdot \mathbb{E}\left[a_{ij} | p_j\right] + \delta_i \tag{20}$$

Such that when $\mathbb{E}\left[A^*_i | p_j\right] \geq \bar{A}$ students enrol in the academic track. Otherwise they enrol in the vocational track. In this case, δ_i represents an idiosyncratic preference parameter for the academic track, independent of ability, and γ_j capture the gradients associated with the subjective perception of ability in each exam subject j.

Among the population of students that opt for the academic track, the expected utility per subject is given by:

$$\mathbb{E}\left[W_{ij}\left(a_{ij}, e_{ij}\right)\right] = \mathbb{E}\left[a_{ij} \cdot e_{ij} + \tilde{\varepsilon}_{ij} - c(e_{ij})\right] \tag{21}$$

Where $\left(a_{ij} \cdot e_{ij} + \tilde{\varepsilon}_{ij}\right)$ defines the subject-specific exam score in in high school, which depends on learning effort e_{ij}, subject-specific ability a_{ij}, and a mean zero idiosyncratic parameter $\tilde{\varepsilon}_{ij}$. The expectations

operator reflects the uncertain, subjective belief about own ability. Learning effort is costly, as reflected by the cost function $c(e_{ij})$. I assume disutility of effort is increasing in effort, but make no assumptions about the concavity or convexity of $c(\cdot)$. The level of effort is chosen by the student. Conditional on the grade observed in the national exam (p_j) expected utility can be re-written as:

$$\mathbb{E}\left[W_{ij}\left(a_{ij}, e_{ij}\right)|p_j\right] = \mathbb{E}\left[a_{ij}|p_j\right] \cdot e_{ij} - c(e_{ij}) \tag{22}$$

B.2. Bayesian Updating

Students are Bayesian agents. Upon receiving the coarse exam grade in subject j, each student i updates beliefs about their subject-specific ability. Receiving a high grade ($p = H$), implies the following posterior[29]:

$$\mathbb{E}\left[a_{ij}|p_{ij} = H\right] = \mathbb{E}\left[a_{ij}|s^*_{ij} \geq \bar{s}\right] = \alpha_{ij} + \frac{\sigma^2_{ij,a}}{\sqrt{\sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}}} \cdot \lambda(z) \tag{23}$$

Where $\lambda(z) = \frac{\phi(z)}{1-\Phi(z)}$ is the inverse Mills ratio, with $\phi(\cdot)$ and $\Phi(\cdot)$ being the probability density function and cumulative function of the standard normal, respectively. Finally, $z = \frac{\bar{s}-\alpha_{ij}}{\sqrt{\sigma^2_{ij,a}+\sigma^2_{j,\varepsilon}}}$.

Analogously, in case of a low grade in the exam, the subjective perception of ability is given by:

$$\mathbb{E}\left[a_{ij}|p_{ij} = L\right] = \mathbb{E}\left[a_{ij}|s^*_{ij} < \bar{s}\right] = \alpha_{ij} - \frac{\sigma^2_{ij,a}}{\sqrt{\sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}}} \cdot \lambda(-z) \tag{24}$$

The change in perceived ability from scoring above the grade cutoff relative to below is thus given by:

$$\Delta_p \mathbb{E}\left[a_{ij}\right] := \mathbb{E}\left[a_{ij}|p_{ij} = H\right] - \mathbb{E}\left[a_{ij}|p_{ij} = L\right] = \frac{\sigma^2_{ij,a}}{\sqrt{\sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}}} \cdot \left[\lambda(z) + \lambda(-z)\right] \tag{25}$$

[29] Full derivation:

$$\begin{aligned}
\mathbb{E}\left[a_{ij}|p_{ij} = 1\right] &= \mathbb{E}\left[a_{ij}|s^*_{ij} \geq \bar{s}\right] = \int_{\bar{s}}^{+\infty} \mathbb{E}\left[a_{ij}|s^*_{ij} = s\right] \cdot \frac{f_s(s)}{Pr(s \geq \bar{s})} ds = \\
&= \int_{\bar{s}}^{+\infty} \left[\alpha_{ij} + \frac{\sigma^2_{ij,a}}{\sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}}(s - \alpha_{ij})\right] \cdot \frac{f_s(s)}{Pr(s \geq \bar{s})} ds = \alpha_{ij} + \frac{\sigma^2_{ij,a}}{\sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}} \left[\mathbb{E}\left[s^*_{ij}|s^*_{ij} \geq \bar{s}\right] - \alpha_{ij}\right] = \\
&= \alpha_{ij} + \frac{\sigma^2_{ij,a}}{\sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}} \left[\alpha_{ij} + \sqrt{\sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}} \cdot \lambda(z) - \alpha_{ij}\right] = \alpha_{ij} + \frac{\sigma^2_{ij,a}}{\sqrt{\sigma^2_{ij,a} + \sigma^2_{j,\varepsilon}}} \cdot \lambda(z)
\end{aligned}$$

B.3. Student's Problem

Given grade p, and after belief updating and enrolling in the high school academic track, each student chooses how much learning effort to exert according to the following problem:

$$e_{ij}^{*} := \arg\max_{e} \mathbb{E}\left[a_{ij}|p\right] \cdot e - c(e) \tag{26}$$

The optimal learning effort is thus implicitly given by $\mathbb{E}\left[a_{ij}|p\right] = c'(e)$. As such, the way effort changes with grading signal p depends on the second derivative of the cost function $c(\cdot)$. By the implicit function theorem, one can show that:

$$\frac{\partial e^{*}}{\partial p} = \frac{\Delta_p \mathbb{E}\left[a_{ij}\right]}{c''(e)} \tag{27}$$

With a concave cost of effort ($c''(e) < 0$), optimal effort in high school will unambiguously decrease in response to a higher exam grade in the end of middle school. With a convex cost function, on the other hand, students exert higher effort (as assumed in Li and Xia, 2022; Delavande et al., 2022).

B.4. Effect of Manipulation on Future Achievement

Suppose that student i, scoring at $s^{*} = \bar{s} - 1$, or $p = L$ has a manipulated grade such that $s = \bar{s}$ and her new signal is $p = H$. According to Equation 27 the student responds on the effort margin depending on the second derivative of the cost function. Also suppose the *true* ability of student i in the subject is $\tilde{\alpha}$, such that the future score observed in the subject, conditional on the grade obtained is: $Y_i(p) = \tilde{\alpha} \cdot e^{*}(p) + \tilde{\varepsilon}_i$.

Student's learning effort changes differentially, depending on getting a high or low grade p, given the change in subjective perception of ability. However, because true ability $\tilde{\alpha}$ does not change with a manipulated grade:

$$\frac{\partial Y_i(p)}{\partial p} = \tilde{\alpha} \cdot \frac{\Delta_p \mathbb{E}\left[a_{ij}\right]}{c''(e)} \tag{28}$$

The effect of a higher grade in the exam on future academic achievement is theoretically ambiguous. With a strictly convex cost function, having a higher coarse grade leads to higher academic achievement through an increase in exerted effort. On the other hand, the effect of a higher grade is weakly negative given a concave cost function.

B.5. Grader's Behavior

The performance of each student i is judged by a grader h which observes ex-post publicly available correction and scoring criteria. There are a total of H_j graders in each exam subject. Each student is randomly matched with a grader. Graders do not observe any student characteristics except those that can be inferred from the exam itself.

I assume that, in the population, ability to perform a given exam is distributed according to a smooth density function $f_{a_j}(\cdot)$. I assume each grader can observe the raw score of her assigned student i, according to the specific evaluation criteria. Having observed the raw score, each grader h chooses how many points to add to a given exam, ϕ_{ieh}. Thus, each student's observed exam score is given by:

$$s_{ij} = s^*_{ij} + \phi_{ijh}, \quad s \in \{0, 1, ..., 100\} \tag{29}$$

Observed exam scores have a one-to-one mapping to a coarser scale of grade levels, p, with $p \in \{1, ...5\}$. The mapping of observed exam scores to grade levels is defined by:

$$p_{ij} = \sum_{p=1}^{5} p \cdot \mathbb{1}\{\bar{s}_p \leq s_{ij} < \bar{s}_{p+1}\} \tag{30}$$

With $\bar{s}_1 = 0, \bar{s}_2 = 20, \bar{s}_3 = 50, \bar{s}_4 = 70, \bar{s}_5 = 90, \bar{s}_6 = 101$. The mapping of the granular to the coarse grading does not depend on the proportion of students at each implied coarse grade, i.e., grading is explicitly absolute, not relative. Grades are salient as important reference levels for both students and teachers. Since final GPA is measured in the coarser scale p, exam scores are first converted into grade levels before being used to compute the final grade at a given subject.

Each grader h is matched with multiple students i, having the following per-student utility function:

$$u_{ijh}(\phi_{ijh}) = \beta_h \cdot \left[p_{ij}\left(\phi_{ieh}, s^*_{ij}\right) - 1\right] - c_j(\phi_{ijh}, s^*_{ij}) \tag{31}$$

Where $\beta_h \geq 0$ is a leniency parameter that reflects graders' myopic other-regarding preferences.In these case, the preferences are myopic because the grader only cares about the present utility of the student—given by her coarse grade—rather than future period utility W_{ij}. Crucially, β_h may be heterogenous across graders. A more lenient teacher has higher levels of β_h. For the same grader, it may also be heterogeneous across different types of students depending on characteristics that can be indirectly inferred from the exam. Under this interpretation, the leniency parameter is the marginal utility gain from attributing an exam score corresponding to a higher grade. Subtracting 1 from the grade level scores ensures that the grader derives no utility from attributing the lowest grade. The crucial behavioral assumption is that lenient graders derive utility from attributing a higher grade. Thus, each grader chooses ϕ_{ijh}, which has a potentially non-linear relationship with the grade level p, depending on the raw score s^*_{ij}.

The disutility function $c_j(\cdot, \cdot)$ reflects how costly it is for grader h to correct exam subject j. I assume the cost reflects both a taste for fairness, a factoring in of detection risk, and the format of the exam. First, graders have a psychological cost from attributing an amount of points that would decouple the final exam score from the perceived student performance. Second, graders may also fear detection from con-

siderable manipulation, provided strict grading criteria. Third, exam scores may be harder to manipulate, conditional on a given raw score, by the way exams are constructed. For instance, there is considerable less leeway for manipulation in a fully multiple choice exam than in a long-form essay, where grading is arguably more subjective. Thus, I assume that, for each level of s^*_{ij}, cost is increasing in the extent of manipulation $c_{e,\phi} > 0$, in almost every region of its domain. By construction, $c_j(0, s^*_{ij}) = 0$. I make no assumptions about higher order derivatives or the sign of c_{j,s^*}. Importantly, the cost function can change according to the test subject j.

As in Diamond and Persson (2016), for each student's exam, each grader solves the following problem:

$$\max_{\phi} \beta_h \left[\sum_{p=1}^{5} p \cdot \mathbb{1}\{\bar{s}_p \leq s^*_{ij} + \phi < \bar{s}_{p+1}\} - 1 \right] - c_j(\phi, s^*_{ij}) \tag{32}$$

In the special case where $\beta_{ih} = 0$ the optimal strategy will be to not manipulate the grade, independently of the raw score, $\phi^*(s^*_{ij}) = 0$. For the case where $\beta_{ih} > 0$, then $\phi^*(s^*_{ij}) \in \{0, \bar{s}_{p+1} - s^*_{ij}\}$. Given the manipulation cost and parameter β_{ih}, the grader will decide whether to attribute the points necessary to make the student pass to a new grade level $p + 1$. As graders only derive utility from the coarser grade levels, it is never optimal to attribute less than the necessary points to bump up the students to a new grade level $(\bar{s}_{p+1} - s^*_{ij})$ as this would still be costly and would not provide any utility gain. By the same logic, it is never optimal to attribute more than the necessary points required to reach the closest grade level.

Importantly, due to the discrete nature of the problem and the lumpiness of points in given exam questions, it may still be optimal to attribute some marginal points $\eta_{ij} \to 0$ above the cutoff scores $\bar{s}_p$ if and only if $c_j(\bar{s}_{p+1} - s^*_{ij}, s^*_{ij}) > c_j(\bar{s}_{p+1} + \eta_{ij} - s^*_{ij}, s^*_{ij})$. Parameter η_{ij} captures the idea that, in some cases, may be easier to justify attributing residual points than those just sufficient to make the student pass to the next grade level. Whereas this feature is important to explain what is observed in the data, I abstain from adding this additional layer of complexity to the optimal decision problem.

Depending on leniency and manipulation cost, each grader manipulates as long as the marginal benefit of manipulation is weakly larger than the marginal cost. I assume that the manipulation cost is always sufficiently high for it to never be optimal to move a student more than one grade level relative to the one predicted by its raw score, i.e., $c_j\left(\bar{s}_{p+1+j} - s^*_{ij}, s^*_{ij}\right) > \beta_{ih} \cdot (p + j), \forall j \in \{1, ..., 5 - p\}$ and $p > 1$.

Since, for each grader h, β_h is constant, and $c_j(\phi, s^*_{ij})$ is increasing in ϕ, then:

Proposition 1 (Bounded manipulation) *For each grade cutoff in set $\mathcal{T} := \{\bar{s}_p\}_{p=2}^5$, and for each grader h, there will be lower bound scores $\{\underline{s}_{peh}\}_{p=2}^5$, implicitly defined by:*

$$\beta_h \cdot (p - 1) = c_j\left(\bar{s}_{p+1} - \underline{s}_{pjh}, \underline{s}_{pjh}\right), \quad \forall p \in \{2, ..., 5\} \tag{33}$$

The exam score of student i is manipulated by grader h if and only if her raw score s^*_{ij} is such that $s^*_{ij} \in \mathcal{M}_{jh} := \bigcup_{p=1}^{5}[\underline{s}_{pjh}, \bar{s}_{p+1})$. Independently of the assigned grader h, each student i has a weakly positive probability of having her exam score in subject e manipulated if $s^*_{ij} \in \mathcal{M}_j := \bigcup_{h=1}^{H} \mathcal{M}_{jh}$. Otherwise, the probability of the student having a manipulated score will be zero.

Proposition 1 implies that each grader h defines her own set of manipulation regions, $\mathcal{M}_{jh}$ (Diamond and Persson, 2016). If the exam of a given student i is assigned to grader h and her raw score falls within this region, then she will have her grade manipulated up until it reaches the new grade level, $p+1$. More lenient graders—i.e., those with higher β_h—have wider manipulation regions. The union of the manipulation sets of all graders identifies all the raw scores that are potentially manipulated. Therefore, students whose raw scores fall in the set of manipulation regions $\mathcal{M}_j$ have a weakly positive probability of having their scores manipulated.

Close to each grade level p there will thus be intervals of observed scores s defined by:

$$\mathcal{P}_{pj} := \left[\underline{s}_{pj}, \bar{s}_{p+1} + \eta_{pj}\right] \tag{34}$$

Where $\underline{s}_{pj}$ is the lowest lower bound score, $\underline{s}_{pjh}$ among all graders, close to each grade cutoff $\bar{s}_{p+1}$, and η_{pj} will be the scores above the cutoff for which there will be manipulated students due to lumpiness. The regions $\mathcal{P}_{pj}$ will be key to identify the latent groups in an hypothetical experiment, as described in Section IV of the main text.

C. Estimation

C.1. Estimating Manipulation

To estimate manipulation measures, I use data collapsed at the score-by-test-by-year level.

The algorithm for the estimation of manipulation estimates is as follows:

1. For each year and subject, define the manipulation regions $\mathcal{P}_p$ around each of the grade cutoffs, according to what comes described in Section 2 of the main text;

2. Define a vector of potential polynomial degrees $\mathcal{Q} = (1, ..., 12)'$;

3. Divide the data into $K = 10$ randomly defined folds;

4. For each polynomial degree $Q \in \mathcal{Q}$, exclude from the dataset one fold k at a time and run the regression specified in Equation 5, using the data from the remaining $K - 1$ folds;

5. For each iteration, compute the mean squared error out of sample ($MSE^{Q}_{k,oos}$), by fitting the estimated model in the excluded fold k. Sum the mean squared errors across all the folds for each polynomial order Q and keep the measure $MSE^{Q}_{oos} = \sum_k MSE^{Q}_{k,oos}$;

6. Select the polynomial order for which the mean squared error statistic (MSE^{Q}_{oos}) computed in Step (5) is the lowest, defining Q^* as the optimal polynomial order:

$$Q^* := \min_{Q} \{MSE^{Q}_{oos}\};$$

7. Run the regression specified in Equation 5, in the main text, with Q^* as the optimal polynomial degree. Derive the counterfactual distribution as described in Section 2 of the main text;

8. (*Block Bootstrap*) Construct a new sample from the original dataset, drawing blocks of observations at random with replacement. The blocks are defined by the school class of the student. Run Steps (3) through (7), $\mathcal{J} = 200$ times. Collect the resulting counterfactual distributions for each bootstrap sample, as in Step (7);

9. For each year, construct manipulation estimates $\{\hat{B}_p\}_{p=2}^{5}$, in-range manipulation estimates $\{\hat{b}_p\}_{p=2}^{5}$ and first stage estimates $\{\hat{\tau}_{T,p}\}_{p=2}^{5}$, according to what is described in Section 2 in the main text, making use of the counterfactual distributions estimated in Steps (7). Compute the same measures for all the bootstrap counterfactual computed in Step (8). The standard deviation of these counterfactual estimates is the estimated standard error. Also construct 95% confidence intervals, where the lower bound is the 2.5th percentile of the distribution of bootstrapped counterfactual estimates and the upper bound is the 97.5th percentile of the distribution of bootstrapped counterfactual estimates.

10. (*Average Across Years.*) Compute the number of students in each bootstrap sample, manipulation region and year combination. Compute the total number of students in each bootstrap sample and manipulation region, across years. Construct yearly weights for each bootstrap sample, as the proportion of students in a given year, among all years in the sample. Average all measures computed in Step (9) across years, using yearly weights. The final estimates are weighted averages across exam years in the sample.

C.2. Estimating ITT and LATE

To estimate intention-to-treat and local average treatment effects I use data at the student-by-test level. The algorithm for the estimation of counterfactual outcomes is as follows, and is separately run for each considered outcome variable:

1. For each year, take the manipulation regions $\mathcal{P}_p$ around each of the grade cutoffs, according to what comes described in Section 2 as given;

2. Define a vector of potential polynomial degrees $\mathcal{Q}' = (1, ..., 10)'$;

3. Use an optimal mean squared error criterion, as specified in Subsection C.1 of this Appendix to select the optimal polynomial degree Q^*;

4. Run the regression specified in Equation 9, in the main text, with Q^* as the optimal polynomial degree. Derive the counterfactual distribution of outcomes as described in Section 2 of the main text;

5. (*Block Bootstrap*) Construct a new sample from the original dataset, drawing blocks of observations at random with replacement. The blocks are defined by the school class of the student. The bootstrap samples are the same as the ones in Step (8) in Subsection C.1 of this Appendix. Run Steps (2) through (4), $\mathcal{J} = 200$ times. Collect the resulting counterfactual distributions for each bootstrap sample, as in Step (4);

6. Construct mean counterfactual outcomes $\{\hat{Y}_p^*\}_{p=2}^5$, intention-to-treat effects $\{\hat{\tau}_{Y,p}\}_{p=2}^5$ and local average treatment effect estimates $\{\hat{\tau}_p\}_{p=2}^5$, according to what is described in Section 2 in the main text, using the counterfactual outcomes estimated in Step (4) and the first stage estimates in Step (9) of Subsection C.1. Compute the same measures for all the bootstrap counterfactuals collected in Step (5). Construct standard errors and 95% confidence intervals.

7. (*Average Across Years.*) Compute the number of students in each bootstrap sample, manipulation region and year combination. Compute the total number of students in each bootstrap sample and manipulation region, across years. Construct yearly weights for each bootstrap sample, as

the proportion of students in a given year, among all years in the sample. Average all measures computed in Step (6) across years, using yearly weights. The final estimates are weighted averages across exam years in the sample.

C.3. Estimating Selective Manipulation

To estimate selective manipulation I use data at the student-by-test level. The algorithm for the estimation of counterfactual outcomes is as follows:

1. For each year, take the manipulation regions $\mathcal{P}_p$ around each of the grade cutoffs, according to what comes described in Section 2 as given;

2. Define a vector of potential polynomial degrees $\mathcal{Q}^{''} = (1, ..., 4)'$;

3. Use an optimal mean squared error criterion, as specified in Subsection C.1 of this Appendix to select the optimal polynomial degree Q^*;

4. Run the regression specified in Equation 13, in the main text, with Q^* as the optimal polynomial degree. Derive the counterfactual distribution of student characteristics as described in Section 2 of the main text;

5. (*Block Bootstrap*) Construct a new sample from the original dataset, drawing blocks of observations at random with replacement. The blocks are defined by the school class of the student. The bootstrap samples are the same as the ones in Step (8) in Subsection C.1 of this Appendix. Run Steps (2) through (4), $\mathcal{J} = 200$ times. Collect the resulting counterfactual distributions for each bootstrap sample, as in Step (4);

6. Construct selective manipulation estimates $\{\hat{\Gamma}_p^X\}_{p=2}^5$, according to what is described in Section 2 in the main text, making use of the counterfactual characteristics estimated in Step (4) and the manipulation estimates in Step (9) of Subsection C.1. Compute the same measures for all the bootstrap counterfactual estimates collected in Step (5). Construct standard errors and 95% confidence intervals.

7. (*Average Across Years.*) Compute the number of students in each bootstrap sample, manipulation region and year combination. Compute the total number of students in each bootstrap sample and manipulation region, across years. Construct yearly weights for each bootstrap sample, as the proportion of students in a given year, among all years in the sample. Average all measures computed in Step (6) across years, using yearly weights. The final estimates are weighted averages across exam years in the sample.

C.4. Derivation of Selective Manipulation

To estimate selective manipulation, I follow Diamond and Persson (2016).

As detailed in Section 2 of the main text, the null hypotheses of interest for individuals close to each grade cutoff $\{\bar{s}_p\}_{p=2}^{5}$ is:

$$H_0: \quad \Gamma_p^X := \mathbb{E}\left[X_i | i \in C^p\right] - \mathbb{E}\left[X_i | s_i^* < \bar{s}_p\right] = 0, \quad s_i^* \in \mathcal{P}_P \tag{35}$$

Where $\mathbb{E}\left[X_i | i \in C^p\right]$ is the mean characteristic of each group of compliers C^p and $\mathbb{E}\left[X_i | s_i^* < \bar{s}_p\right]$ is the mean characteristic of the group of those eligible to manipulation.

To derive the characteristics of compliers, one can start by the observation of the fact that the observed mean characteristics of those individuals above the cutoff and in the manipulation is a weighted average of the characteristics of compliers and always-takers:

$$\mathbb{E}\left[X_i | s_i \geq \bar{s}_p\right] = \mathbb{E}\left[X_i | i \in C^p\right] \cdot \frac{B_p}{\sum_{s \geq \bar{s}_p} F_s} + \mathbb{E}\left[X_i | i \in A^p\right] \cdot \frac{\sum_{s \geq \bar{s}_p} F_s - B_p}{\sum_{s \geq \bar{s}_p} F_s}, \quad s \in \mathcal{P}_p \tag{36}$$

Where B_p is the proportion of compliers and $\sum_{s \geq \bar{s}_p} F_s$ is the proportion of those above the cutoff.

Likewise, the observed mean characteristics of those individuals below the cutoff (never-takers) is given by:

$$\mathbb{E}\left[X_i | s_i < \bar{s}_p\right] = \mathbb{E}\left[X_i | s_i^* < \bar{s}_p\right] - \mathbb{E}\left[X_i | i \in C^p\right] \cdot \frac{B_p}{\sum_{s < \bar{s}_p} F_s^*}, \quad s \in \mathcal{P}_p \tag{37}$$

Where $\sum_{s < \bar{s}_p} F_s^*$ is the counterfactual proportion of those individuals that would be found below the cutoff in a state of the world without manipulation.

From the equations defined above, one can derive two definitions of mean characteristics of compliers:

$$\mathbb{E}\left[X_i | i \in C^p\right]^+ = \frac{1}{B_p} \cdot \left[\mathbb{E}\left[X_i | s_i \geq \bar{s}_p\right] \cdot \sum_{s \geq \bar{s}_p} F_s - \mathbb{E}\left[X_i | s_i^* \geq \bar{s}_p\right] \cdot \left(\sum_{s \geq \bar{s}_p} F_s - B_p\right)\right] \tag{38}$$

$$\mathbb{E}\left[X_i | i \in C^p\right]^- = \frac{\sum_{s < \bar{s}_p} F_s^*}{B_p} \cdot \left(\mathbb{E}\left[X_i | s_i^* < \bar{s}_p\right] - \mathbb{E}\left[X_i | s_i < \bar{s}_p\right]\right) \tag{39}$$

The estimate each of the above measures by inputting the corresponding estimated values:

$$\hat{X}_p^{C,+} = \frac{1}{\hat{B}_p} \cdot \left[\bar{X}_p^{+} \cdot \sum_{s \geq \bar{s}_p} F_s - \hat{X}_p^{+} \cdot \left(\sum_{s \geq \bar{s}_p} F_s - \hat{B}_p \right) \right] \tag{40}$$

$$\hat{X}_p^{C,-} = \frac{\sum_{s < \bar{s}_p} \hat{F}_s^*}{\hat{B}_p} \cdot \left(\hat{X}_p^{-} - \bar{X}_p^{-} \right) \tag{41}$$

Where $\hat{X}_p^{+}$, $\hat{X}_p^{-}$ are the estimated counterfactual measures as defined in Section 2 of the main text, and $\bar{X}_p^{+}$, $\bar{X}_p^{-}$ are the observed mean characteristics above and below the grade cutoff, respectively.

To increase power, I estimate the mean characteristic of the complier groups as the simple average of the objects above:

$$\hat{X}_p^{C} = \frac{1}{2} \cdot \left[\hat{X}_p^{C,+} + \hat{X}_p^{C,-} \right] \tag{42}$$

The estimate of selective manipulation will thus be $\hat{\Gamma}_p^{X} = \hat{X}_p^{C} - \hat{X}_p^{-}$.

D. Manipulation in Specific Exam Domains

In Section V, in the main text, I show that there is substantial manipulation of test scores. Furthermore, I show that there are significant differences in manipulation across subjects. But how are graders manipulating?

To answer this question I use detailed data on the number of points awarded to each item in the exams, for years in which such data is available (2012 through 2015). I construct a dataset containing information on the number of attained points, but also attainable points for each item in each exam. Furthermore, I collect information on the curricular domain and type of each exam item (multiple choice, open-response, etc.).

Each exam item falls within a specific curricular domain, set by the Ministry of Education curricular guidelines. In the Language exam the domains are: Reading, Grammar, Literature, and Essay. Within each curricular domain there may be different types of questions, except for the essay item, which consists in writing a long-form text about a specific topic according to a prompt provided in the exam. In the case of the Math exam the domains are: Geometry, Statistics, Functions, Numbers and Operations, and Algebra.

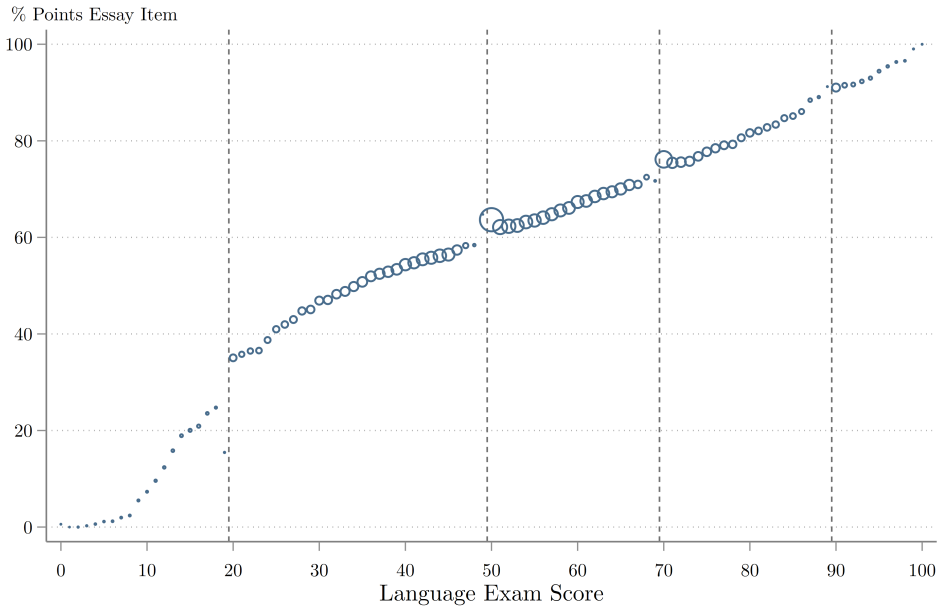
My main hypothesis is that, in the Language exam, graders mostly manipulate in the essay item, for which what constitutes a full or partially right answer is open to considerable subjective interpretation. If the hypothesis is true, I expect to observe discontinuities in the underlying relationship between Language exam scores and the percentage of points attained in the essay item.

Figure D.1 offers initial evidence that graders do manipulate in the exam item, at least in some of the grade cutoffs. The figure shows the average percentage of points attained in the essay item for each Language exam score bin. The relative size of each dot represents the number of observations in each bin. As it is apparent, there are discontinuous jumps in the relationship between these two variables across these grade cutoffs. Since students above each grade cutoff include those who had their exam scores inflated, a positive discontinuity in the cutoff suggests that manipulation is occurring in that item. On the other hand, I expect to observe a negative, mechanical, discontinuity on all other items that are not being manipulated. The negative effect stems from the fact that if students below and across the cutoffs are to be identical in their total exam score, then if those above are being manipulated upwards in one item, they should have, on average, fewer points on other items.

To formally test this hypothesis I employ a regression discontinuity design, normalizing and pooling across cutoffs and exam years. First, I normalize each test score to take a value of zero at each cutoff, and make it vary between −10 and 10 according to the distance in test scores to the cutoffs, providing a normalized running variable ($\dot{s}$). I then define a post-cutoff indicator, $\mathbb{1}\{\dot{s}_i \geq 0\}$, as my main regressor of interest.

I run OLS specifications of the form:

Figure D.1. Percentage of Score Points Attained in the Exam Item by Language Exam Score



Notes: Each point depicts the average percentage of points attained in the essay item for each Language exam score bin. The relative size of each dot represents the number of observations in each bin, so that points associated with bins with lower number of observations have a relatively smaller size in the graph. Vertical dash lines indicate the grade cutoffs.

$$\text{Domain}^{d}_{ipt} = \alpha_0 + \alpha_1 \mathbb{1}\{\dot{s}_i \geq 0\} + \alpha_2 \dot{s}_i + \alpha_3 \dot{s}_i \times \mathbb{1}\{\dot{s}_i \geq 0\} + \iota_{pt} + \nu_{ipt} \tag{43}$$

Where Domain^{d}_{ipt} represents the percentage of points attained in a given curricular domain, varying by individual i, grade cutoff p and exam year t. To capture the slope of the relationship of interest close to the cutoffs I include in the specification test scores (s_i) and test scores interacted with the post-cutoff dummy. I also control for cutoff-by-year fixed effects (ι_{pt}), to exploit the variation around each grade cutoff for each different exam year. I weigh observations using a triangular kernel, with a bandwidth of 10 before and after the cutoff. I also run identical specifications, with fixed effects by year, for each different grade cutoff separately. For inference, I cluster the standard errors at the class level. The coefficient of interest is α_1, which captures the discontinuous change at the cutoff. It should not be interpreted causally, but as suggestive evidence of curricular domains in which graders are being more lenient.

Table D.5 presents the main results for the case of the Language exam. As hypothesized, I find that in the Language exam, students right above grade cutoffs have on average 3.29 p.p. (s.e. = 0.19) of possible attained points in the exam, from a control mean of 64.4% (Column 1). Furthermore, manipulation is more prevalent in lower grade cutoffs, while I do not find evidence that graders are bumping up the essay grades of students at the margin the highest grade. On the other hand, as expected, I find negative coefficients for exam items in each of the other curricular domains (Columns 2-4).

Table D.6 presents results analogous to the ones in Table D.5 for the case of the Math exam. I find that

Table D.5. RDD Estimates on Language Domains' Scores

Outcome: % Score Points Attained in Item Domain	Essay	Reading	Grammar	Literature
	(1)	(2)	(3)	(4)
Normalized Pooled	3.29***	-2.91***	-0.86***	-0.95***
	(0.19)	(0.29)	(0.26)	(0.20)
	64.44	65.27	44.95	53.71
By Cutoff				
Grade 1 to 2 (s = 20)	7.28***	-6.71***	-3.74***	-1.22**
	(1.26)	(1.46)	(0.91)	(0.62)
	20.30	21.53	14.21	9.51
Grade 2 to 3 (s = 50)	3.35***	-4.49***	-1.71***	0.34
	(0.32)	(0.52)	(0.39)	(0.34)
	56.44	51.61	31.80	38.05
Grade 3 to 4 (s = 70)	2.44***	-1.68***	-0.27	-1.15***
	(0.27)	(0.41)	(0.42)	(0.30)
	70.01	74.63	51.11	66.49
Grade 4 to 5 (s = 90)	0.50	0.88	0.72	-1.24***
	(0.40)	(0.55)	(0.64)	(0.44)
	85.10	91.64	77.95	84.26

Notes: The table shows estimates of the effect of the association between being located just above the grade cutoffs and the percentage of points attained in exam items of each Language curricular domain. The normalized pooled estimates pool every grade cutoff and normalize exam scores to be 0 at the cutoff value. The table also presents the same estimates by cutoff. Standard errors clustered at the class level are presented in parentheses. Below standard errors are presented the mean of the dependent variable for those immediately below the cutoff. ***, **, *: Statistical significance at the 1, 5 and 10% levels, respectively.

graders seem to be more lenient with students in algebra questions. I find that students right above grade cutoffs have on average 1.61 p.p. (s.e. = 0.21) of possible attained points in the exam, from a control mean of 54.41% (Column 1).

Table D.6. RDD Estimates on Math Domains' Scores

Outcome: % Score Points Attained in Item Domain	Algebra	Geometry	Statistics	Functions	Operations
	(1)	(2)	(3)	(4)	(5)
Normalized Pooled	1.61***	-0.03	-1.34***	-0.97**	-0.16
	(0.21)	(0.10)	(0.28)	(0.38)	(0.24)
	54.41	43.15	63.20	52.30	53.36
By Cutoff					
Grade 1 to 2 (s = 20)	3.19***	0.11	-3.73***	-2.04***	-0.51
	(0.34)	(0.18)	(0.58)	(0.69)	(0.45)
	11.20	12.50	30.20	19.22	19.82
Grade 2 to 3 (s = 50)	1.45***	0.42*	-1.74***	-0.11	-0.77
	(0.54)	(0.24)	(0.60)	(0.94)	(0.54)
	54.34	36.19	62.00	46.41	51.86
Grade 3 to 4 (s = 70)	0.07	-0.00	-0.07	-0.36	-0.03
	(0.40)	(0.22)	(0.51)	(0.77)	(0.47)
	76.02	57.85	79.76	69.67	68.03
Grade 4 to 5 (s = 90)	0.26	-0.21	0.19	-0.01	0.24
	(0.27)	(0.18)	(0.37)	(0.54)	(0.38)
	91.22	81.08	93.07	88.54	87.27

Notes: The table shows estimates of the effect of the association between being located just above the grade cutoffs and the percentage of points attained in exam items of each Math curricular domain. The normalized pooled estimates pool every grade cutoff and normalize exam scores to be 0 at the cutoff value. The table also presents the same estimates by cutoff. Standard errors clustered at the class level are presented in parentheses. Below standard errors are presented the mean of the dependent variable for those immediately below the cutoff. ***, **, *: Statistical significance at the 1, 5 and 10% levels, respectively.

E. Other Results

E.1. OLS Estimates

In this section I present evidence on the association between having each exam grade and the outcomes of interest.

Table E.7 shows the association between exam grades and short-term educational attainment outcomes. As expected, students with higher grades in either exam are both significantly less likely to repeat ninth grade and more likely to complete basic education.

Table E.8 shows the association between exam grades and students choices regarding high school curricula. As expected, students with higher grades in either exam are both significantly less likely to repeat ninth grade and more likely to complete basic education.

Table E.9 shows the association between exam grades in ninth grade and academic achievement in high school, for those that enrolled in the academic track. As expected, students with higher grades in either exam are both significantly better students during high school, both in terms of their teacher grades in tenth grade and in their performance in twelfth grade exam in the same subject.

Table E.7. OLS: Short-term Attainment

Outcomes: **Educational Attainment**	**Grade Cutoffs**			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Has to Repeat Grade 9 (%)				
A.1. Language	-28	-11.27	-5.9	-0.89
	(0.71)	(0.10)	(0.05)	(0.04)
Observations	196,418	494,477	438,320	148,861
R-squared	0.0256	0.0355	0.0192	0.0054
A.2. Math	-14.46	-9.76	-2.83	-0.37
	(0.17)	(0.08)	(0.05)	(0.02
Observations	340,877	423,710	275,593	147,016
R-squared	0.0436	0.0387	0.0139	0.0022
B. Graduated from Grade 9 (%)				
B.1. Language	32.52	12.2	6.28	0.91
	(0.71)	(0.10)	(0.05)	(0.03)
Observations	195,448	492,127	436,308	148,213
R-squared	0.0283	0.0388	0.0219	0.0057
B.2. Math	16.09	10.31	2.89	0.38
	(0.17)	(0.08)	(0.05)	(0.02)
Observations	339,089	421,697	274,456	146,466
R-squared	0.0463	0.0405	0.0143	0.0023

Notes: The table reports OLS estimates of receiving a higher coarse grade in the Language and Math exam on the likelihood of having to repeat ninth grade (Panel A) and completing basic education (Panel B). The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012. Each coefficient estimates the difference in the outcome between those that have an exam grade at or above the indicated cutoff, but below the following cutoff, and those that have an exam grade below the cutoff. Every regression controls for year fixed effects. For instance, students who have a coarse grade of 2 instead of 1 in the Language exam are, on average, 28 percentage points less likely to repeat ninth grade (Panel A.1, Col. 1). Robust standard errors are presented in parentheses.

Table E.8. OLS: High School Choices

Outcomes: **High School Choices**	**Grade Cutoffs**			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Chose Academic Track in High School (%)				
A.1. Language	30.37	25.8	18.5	5.9
	(0.68)	(0.14)	(0.11)	(0.14)
Observations	182,829	472,006	426,810	147,106
R-squared	0.0139	0.0711	0.0611	0.0195
A.2. Math	18.69	21.47	12.78	4.75
	(0.21)	(0.14)	(0.12)	(0.11)
Observations	319,827	407,751	271,387	145,589
R-squared	0.0263	0.0570	0.0425	0.0136
B. Chose Science Sub-track (%)				
B.1. Language	4.82	11.22	13.66	10.19
	(1.65)	(0.20)	(0.17)	(0.38)
Observations	93,294	311,727	342,544	136,409
R-squared	0.0067	0.0147	0.0212	0.0084
B.2. Math	24.48	22.96	13.39	9
	(0.26)	(0.18)	(0.19	(0.23)
Observations	182,609	280,750	234,339	137,780
R-squared	0.0527	0.0618	0.0251	0.0112

Notes: The table reports OLS estimates of receiving a higher coarse grade in the Language and Math exam on the likelihood of choosing the academic track (Panel A) and the likelihood of choosing the science sub-track in the academic track (Panel B), for Language (Panels A.1 and B.1) and Math (Panels A.2 and B.2). The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012. Each coefficient estimates the difference in the outcome between those that have an exam grade at or above the indicated cutoff, but below the following cutoff, and those that have an exam grade below the cutoff. Every regression controls for year fixed effects. For instance, students who have a coarse grade of 2 instead of 1 in the Language exam are, on average, 30 percentage points more likely to choose the academic track (Panel A.1, Col. 1). Robust standard errors are presented in parentheses.

Table E.9. OLS: High School Achievement

Outcomes: **High School Achievement**	**Grade Cutoffs**			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Subject Grade Next Year (0-20)				
A.1. Language	0.58	1.45	2.33	2.43
	(0.06)	(0.00)	(0.01)	(0.02)
Observations	111,077	336,653	344,271	129,471
R-squared	0.0093	0.0924	0.2097	0.1251
A.2. Math	-0.21	1.35	2.74	3.19
	(0.03)	(0.01)	(0.01)	(0.02)
Observations	109,433	193,666	187,986	117,884
R-squared	0.0093	0.0628	0.2122	0.2316
B. Exam Score Percentile in Grade 12				
B.1. Language	12.77	18.83	23.43	19.65
	(0.85)	(0.1)	(0.09)	(0.16)
Observations	70,030	262,924	310,887	130,327
R-squared	0.0169	0.1095	0.1841	0.0911
B.2. Math	13.69	15.52	20.76	21.06
	(0.40)	(0.15)	(0.12)	(0.14)
Observations	39,278	111,967	158,466	114,713
R-squared	0.0331	0.0984	0.1592	0.1480

Notes: The table reports OLS estimates of receiving a higher coarse grade in the Language and Math exam on the same subject grade in tenth grade (Panel A) and the same subject exam score percentile in grade 12 (Panel B), for Language (Panels A.1 and B.1) and Math (Panels A.2 and B.2). The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012. Each coefficient estimates the difference in the outcome between those that have an exam grade at or above the indicated cutoff, but below the following cutoff, and those that have an exam grade below the cutoff. Every regression controls for year fixed effects. For instance, students who have a coarse grade of 2 instead of 1 in the Language exam score, on average, 13 percentiles above in the Language exam in grade 12 (Panel B.1, Col. 1). Robust standard errors are presented in parentheses.

Table E.10. Selective Manipulation by Grade Cutoff in the Math Exam

Outcome: **Selective Manipulation**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
Female	0.39	0.48	0.05	1.46
	(0.79)	(0.55)	(1.03)	(3.99)
	[54.7]	[50.9]	[51.3]	[53.6]
Immigrant	0.05	0.35	-0.34	-0.25
	(0.44)	(0.28)	(0.50)	(1.07)
	[8.8]	[6.2]	[5.2]	[4.3]
Free or Reduced Price Lunch	1.22	0.80	0.81	2.52
	(0.74)	(0.48)	(1.01)	(11.05)
	[44.3]	[33.9]	[27.5]	[18.4]
Higher Education at Home	0.13	0.05	-0.69	-1.89
	(0.37)	(0.37)	(0.95)	(11.13)
	[5.7]	[12.0]	[21.4]	[38.4]
Unemployment in Household	0.56	1.01	-0.39	0.37
	(0.61)	**(0.32)***	(0.72)	(2.03)
	[16.7]	[14.1]	[11.3]	[7.5]
School Grade (Language)	0.00	0.04	0.02	0.03
	(0.01)	**(0.02)***	(0.02)	(0.04)
	[2.8]	[3.0]	[3.4]	[4.1]
School Grade (Math)	0.00	0.04	0.01	0.04
	(0.04)	**(0.01)***	(0.02)	(0.04)
	[2.2]	[2.9]	[3.5]	[4.3]

Notes: The table reports selective manipulation estimates in the Math exam as the difference between the estimated characteristics of the group of manipulated students and the estimated characteristics of the group of students that could have been manipulated (eligible), for years 2007 through 2012. Each column indicates the manipulation region around the threshold to which the estimates refer to. Each row indicates the characteristic to which the estimate refers to. Coefficients associated with dummy variables are expressed in percentage points terms. In parentheses, the table presents block bootstrap standard errors. In square brackets, the table presents the estimated mean of the characteristic of students eligible for being manipulated close to each grade cutoff. For instance, I estimate that 30.4 percent of students that could have been manipulated close to the lowest proficiency threshold are female. The group of students that is bumped up across this threshold is estimated to have 1.3 percentage points more female students than among the eligible. I use the scores of students outside of the manipulation regions to estimate the characteristic, at any test score bin inside the manipulation regions. Section 2 describes how the estimates are constructed. See Appendix C.4 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the 95% confidence intervals of the estimate do not include zero.

E.2. Selective Manipulation in the Math Exam

E.3. LATE for Other Outcomes

To further explore the consequences of the short term impact of being retained in the same grade, I investigate the effect of manipulation on the likelihood of completing basic education in the following three years. Table E.11 presents estimates analogous to those presented in Table 7, in the main text, but for the outcome of graduation from basic schooling. In line with the previous results, I find that around the lowest cutoff ($\bar{s}_2 = 20$), exposure to manipulation in the Language exam increases the likelihood of graduation from basic education by 14 p.p. (s.e. = 2.6) for those with bumped up grades. For the case of the Math exam, the likelihood of graduation increases by 7.6 p.p. (s.e. = 1.6).

Table E.11. LATE: Basic Education Completion

Outcome: **Graduated from Grade 9 (%)**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Language				
Manipulated Exam Grade	15.4	0.3	1.0	0.2
	(3.4)*	(0.5)	**(0.4)***	(0.6)
	[62.7]	[88.9]	[97.4]	[99.7]
First Stage	0.44	0.40	0.29	0.34
	(0.01)	(0.00)	(0.00)	(0.01)
B. Math				
Manipulated Exam Grade	6.0	0.3	1.0	0.3
	(1.4)*	(0.7)	(0.9)	(0.8)
	[77.8]	[93.8]	[98.6]	[99.9]
First Stage	0.29	0.24	0.15	0.15
	(0.00)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff in the Language (Panel A) and Math (Panel B) exam on the likelihood of completing school grade nine in the following three years. The coefficients for the effect on the dependent variable are expressed in percentage point terms. The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012 (the period for which the granular exam score is not reported). Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

Table E.12 shows estimates analogous to the ones in Table 8, in the main text, but having percentiles of twelfth grade national exams' performance as the outcome of interest.

E.4. Results After Grade Reform

Table E.12. LATE: Percentile Exam Performance in Twelfth Grade

Outcome: **Exam Score Percentile in Grade 12**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Effect on the Same Subject				
A.1. Language				
Manipulated Exam Grade	-2.1	0.8	-0.1	-1.6
	(3.4)	(0.5)	(0.6)	(2.0)
	[15.4]	[35.1]	[57.4]	[82.2]
First Stage	0.50	0.39	0.29	0.34
	(0.12)	(0.01)	(0.00)	(0.03)
A.2. Math				
Manipulated Exam Grade	-4.1	-0.7	1.6	1.9
	(5.6)	(1.0)	(1.7)	(1.8)
	[17.7]	[33.9]	[48.0]	[71.9]
First Stage	0.25	0.25	0.16	0.14
	(0.70)	(0.00)	(0.00)	(0.01)
B. Effect on the Other Subject				
B.1. Language				
Manipulated Exam Grade	0.5	-0.6	1.7	-1.8
	(136.9)	(0.9)	(0.9)	(1.1)
	[20.2]	[35.9]	[52.4]	[74.6]
First Stage	0.30	0.38	0.28	0.35
	(0.94)	(0.01)	(0.00)	(0.09)
B.2. Math				
Manipulated Exam Grade	3.0	0.9	-1.6	2.3
	(1.4)	(0.8)	(1.4)	(1.7)
	[31.0]	[42.9]	[54.4]	[71.3]
First Stage	0.29	0.25	0.15	0.15
	(0.01)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff in the Language (Panels A.1 and B.1) and Math (Panel A.2 and B.2) exam on the percentile of exam score in the same subject (Panel A) or the other subject (Panel B) in school grade 12. The sample includes all ninth grade exam takers that attend ninth grade between 2007 and 2012 (the period for which the granular exam score is not reported) and are ever found taking Language or Math courses in high school. Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

Table E.13. LATE: Ninth Grade Repetition, Post Reporting Reform (2013-2015)

Outcome: **Has to Repeat Grade 9 (%)**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Language				
Manipulated Exam Grade	-9.0	-0.3	-0.8	-1.1
	(4.0)*	(0.8)	(0.8)	(2.49)
	[50.5]	[9.1]	[1.44]	[0.51]
First Stage	0.48	0.42	0.37	0.43
	(0.01)	(0.01)	(0.00)	(0.01)
B. Math				
Manipulated Exam Grade	-3.8	0.9	0.5	-0.1
	(2.3)	(0.6)	(0.6)	(0.6)
	[25.4]	[4.5]	[0.5]	[0.1]
First Stage	0.22	0.31	0.24	0.21
	(0.00)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff in the Language (Panel A) and Math (Panel B) exam on the likelihood of being retained and having to repeat school grade nine. The coefficients for the effect on the dependent variable are expressed in percentage point terms. The sample includes all ninth grade exam takers that attend ninth grade between 2013 and 2015 (the period for which the granular exam score is also reported, alongside the exam grade). Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

Table E.14. LATE: Basic Education Completion, Post Reporting Reform (2013-2015)

Outcome: **Graduated from Grade 9 (%)**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Language				
Manipulated Exam Grade	11.9	0.5	0.5	0.8
	(3.9)*	(0.6)	(0.4)	(0.5)
	[46.1]	[90.8]	[98.7]	[99.6]
First Stage	0.48	0.42	0.37	0.43
	(0.01)	(0.01)	(0.00)	(0.01)
B. Math				
Manipulated Exam Grade	4.9	-1.1	0.6	-0.6
	(2.4)	(0.7)	(0.7)	(0.6)
	[73.6]	[95.7]	[99.3]	[100.1]
First Stage	0.22	0.31	0.24	0.21
	(0.00)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff in the Language (Panel A) and Math (Panel B) exam on the likelihood of completing school grade nine in the following three years. The coefficients for the effect on the dependent variable are expressed in percentage point terms. The sample includes all ninth grade exam takers that attend ninth grade between 2013 and 2015 (the period for which the granular exam score is also reported, alongside the exam grade). Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

Table E.15. LATE: High School Choices, Post Reporting Reform (2013-2015)

Outcomes: **High School Choices**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Chose Academic Track in High School (%)				
A.1. Language				
Manipulated Exam Grade	-0.7	0.2	0.2	-0.6
	(3.6)	(0.7)	(0.8)	(0.7)
	[29.1]	[69.7]	[90.8]	[98.5]
First Stage	0.48	0.42	0.36	0.42
	(0.01)	(0.01)	(0.00)	(0.01)
A.2. Math				
Manipulated Exam Grade	4.6	-1.1	-0.8	-2.5
	(2.5)*	(1.2)	(1.2)	**(1.3)***
	[48.2]	[76.9]	[92.0]	[98.1]
First Stage	0.22	0.31	0.24	0.22
	(0.00)	(0.00)	(0.00)	(0.01)
B. Chose Science Sub-track (%)				
B.1. Language				
Manipulated Exam Grade	-7.1	0.8	0.3	0.1
	(6.0)	(0.9)	(1.4)	(2.2)
	[36.7]	[46.8]	[62.8]	[79.0]
First Stage	0.48	0.42	0.37	0.43
	(0.01)	(0.01)	(0.00)	(0.01)
B.2. Math				
Manipulated Exam Grade	2.1	-1.2	2.9	3.3
	(2.7)	(1.6)	(2.3)	(2.8)
	[20.2]	[53.2]	[70.4]	[81.4]
First Stage	0.22	0.31	0.24	0.21
	(0.00)	(0.00)	(0.00)	(0.01)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff on the likelihood of choosing the academic track (Panel A) and the scientific sub-track (Panel B), in both the Language (Panels A.1 and B.1) and Math (Panels A.2 and B.2) exams. The coefficients for the effect on the dependent variable are expressed in percentage point terms. The sample includes all ninth grade exam takers that attend ninth grade between 2013 and 2015 (the period for which the granular exam score is also reported, alongside the exam grade). Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.

Table E.16. LATE: High School Achievement, Post Reporting Reform (2013-2015)

Outcome: **High School Academic Achievement in the Same Subject**	Grade Cutoffs			
	$\bar{s}_2 = 20$	$\bar{s}_3 = 50$	$\bar{s}_4 = 70$	$\bar{s}_5 = 90$
	(1)	(2)	(3)	(4)
A. Grade in High School (0-20)				
A.1. Language				
Manipulated Exam Grade	-0.2	0.0	0.0	0.1
	(0.2)	(0.0)	(0.1)	(0.1)
	[10.3]	[11.4]	[13.6]	[16.4]
First Stage	0.50	0.40	0.37	0.41
	(0.02)	(0.01)	(0.00)	(0.01)
A.2. Math				
Manipulated Exam Grade	-0.8	0.0	-0.1	-0.3
	(0.3)	(0.1)	(0.1)	(0.2)
	[9.6]	[10.1]	[12.5]	[16.0]
First Stage	0.24	0.31	0.24	0.21
	(0.01)	(0.01)	(0.01)	(0.01)
B. Exam Score Percentile in Grade 12				
B.1. Language				
Manipulated Exam Grade	-0.6	0.8	0.3	2.5
	(1089.9)	(0.6)	(0.6)	(1.0)
	[16.5]	[38.6]	[61.0]	[84.0]
First Stage	0.45	0.40	0.36	0.40
	(0.08)	(0.04)	(0.04)	(0.21)
B.2. Math				
Manipulated Exam Grade	12.4	2.2	-1.4	2.5
	(21.3)	(1.1)	(1.3)	(1.5)
	[14.2]	[32.3]	[51.1]	[74.0]
First Stage	0.22	0.31	0.24	0.21
	(0.04)	(0.03)	(0.03)	(0.02)

Notes: The table reports the estimated effect of being bumped up across each grade cutoff on school subject grade in Grade 10 (Panel A) and the exam score percentile in Grade 12 (Panel B), in both the Language (Panels A.1 and B.1) and Math (Panels A.2 and B.2) exams. The sample includes all ninth grade exam takers that attend ninth grade between 2013 and 2015 (the period for which the granular exam score is also reported, alongside the exam grade) which can be found in high school. Block bootstrapped standard errors are presented in parentheses. In square brackets, the table presents the estimated dependent variable mean for individuals in the manipulation region in a state of the world without manipulation. First stage is the estimated proportion of individuals in each manipulation region that are bumped up across the cutoff. Section 2 describes how the estimates are constructed. See Appendix C.2 for details. Coefficients with the asterisk (*) close to the corresponding standard error indicate that the bootstrapped 95% confidence interval does not include zero.