The Psychological Effect of a Math Signal*

Pedro Albarran[a], Marianna Battaglia[a], and Marcello Sartarelli[†b]

[a]Fundamentos del Análisis Económico (FAE), Universidad de Alicante
[b]Análisis Económico y Economía Cuantitativa, Universidad Complutense de Madrid

September 13, 2021

Abstract

This paper tests whether barely obtaining a pass score in at least one of two midterm tests has an effect on subsequent achievement in a Math course. To estimate the effect, we created a novel dataset by linking administrative and survey data on students at a medium size Spanish University and used a regression discontinuity design in which the cutoff is 5, i.e. the pass score in the national grading system. Although obtaining a score just equal to or barely greater than 5 in midterm tests has no immediate consequence for students, it may have a psychological effect by acting as a (de)motivating signal to study and pass the course, with the sign of the effect being unclear ex-ante. We find that obtaining a pass score in at least one midterm has a positive effect on the final exam score. The result seems to be explained by students' study strategy, i.e. the ability to obtain a value in the final exam score that when averaged along with the midterm tests scores leads to an overall passing score in the course. Overall, our results suggest that partly unexplored psychological mechanisms may help us deepen our understanding of the determinants of achievement in higher education.

JEL Classification: I23, D91
Keywords: achievement, math, regression discontinuity

*The usual disclaimers apply. We thank for useful feedback to Ramses Abul Naga, Alex Armand, David Jiménez-Gomez, Paolo Masella, Dan Rees, Núria Rodríguez-Planas, Francesco Sobbrio and Alessandro Tondini. Financial support from the Spanish Ministerio de Economía y Competitividad (ECO2015-65820-P and ECO2016-77200-P), Generalitat Valenciana (Research Projects Grupos 3/086) and Instituto Valenciano de Investigaciones Económicas (IVIE) is gratefully acknowledged.

†Corresponding author. E-mail: marcellosartarelli@gmail.com

1 Introduction

Innate ability is perhaps one of the most important factors influencing educational choices and success in the labour market. However, recent research has shown that even among students with very similar predetermined characteristics and, hence, similar ability, those obtaining a score just above the pass cutoff in at least one high school final test has a higher probability of attending college than those just below in the UK (Machin et al., 2020) and in the US (Papay et al., 2010). Less is known, instead, about the mechanisms driving students' behaviour and how they react to a positive achievement signal. Although it has been hypothesized that psychological factors may play a relevant role in these settings (see for a review Koch et al., 2015; Lavecchia et al., 2016), quantifying them and estimating the clean effect of passing a test on subsequent achievement requires a context in which, first, one or more tests are taken over time and, second, at least one of them is low stake.

In this paper, we take advantage of the assessment structure of the first Math course in Business Economics and in Economics degrees at the University of Alicante in Spain, with two low stakes midterm tests followed by a final exam, to quantify the empirical relevance of a psychological effect in achievement. Midterm tests and exams are graded on a continuous scale from 0 to 10, and 5 is the pass score, as well as in all courses in all Spanish universities and traditionally in compulsory education. However, scoring 5.1 rather than 4.9 in the midterm tests has a quantitative irrelevant effect on the probability of passing the Math course since one can take the final exam, regardless of having passed the midterm tests. To test whether barely obtaining a pass score in at least one of two midterm tests leads to an increase in the final exam score, we created a novel dataset by linking administrative and survey data on three cohorts of first-year students in the academic years 2016 to 2019. We estimate the effect by way of a regression discontinuity design (RDD) in which the running variable is the greatest score in the midterm tests and the pass score, i.e. 5, is the cutoff. Importantly, thanks to the surveys we conducted during the course we obtained detailed information on students' socio-demographics and found that they are similar on both sides of the cutoff, i.e. balanced, thus offering support to the validity of our design.

The most important reason for studying Math achievement in a higher education setting is the widely documented positive relationship between Math achievement, success in a university degree, and in the labour market (Levine and Zimmerman, 1995; Rose and Betts, 2004; Joensen and Nielsen, 2009). Since a Math gender gap has been documented in several studies (see for a review Niederle and Vesterlund, 2010), we have also tested whether gender differences in behavioural responses to signals in the tests could shed light on the reasons underlying such gap.

The behavioural mechanism behind the effect of barely obtaining a pass score in one or more midterm tests is interpreted as being of a mainly psychological nature for two reasons. Firstly, there is no direct consequence for obtaining a score barely greater than or equal to 5 rather than barely smaller than 5, since all students take the same final exam independently of the midterm score. Secondly, although a slightly greater score in the midterm has a negligible effect on the probability of passing the course, those students obtaining a score barely greater than or equal to 5 in the midterm are more likely, in theory and based on anecdotal evidence on students' reactions, to give a positive interpretation to their score and think that they will do better than others in subsequent assessment in the course. The empirical evidence of such a left-digit effect has been documented in related studies in the education literature, finding that a score to the left of a cutoff in an exam increases performance in the re-take exam (Olsen, 2013) and in surveys over school quality a value of a proxy for quality to the left of a cutoff is more frequently associated with lower perceived quality (Goodman et al., 2018). However, the net effect is ambiguous, as it may increase for some students and decrease for others. Kahneman et al. (2021) provides a general review of evidence that a little change, partially due to a mistake, can influence decision-making.

In our setting, we find that the effect of barely scoring at or above 5 in at least one of the midterms on students' score in the final exam is positive and significant, with the magnitude being approximately 0.25-0.30 standard deviations of the final exam grade. The size of this effect is in line with the magnitude found in the literature for other educational interventions, in particular those related to the effect of positive feedbacks. Conversely, we find no gender difference: boys and girls react to the signal in the same way. Additional results show that the psychological effect is most likely due to a better study strategy proxied by the difference in the score obtained in the final exam and the student-specific minimum score needed to pass the course given the midterms scores. Students seem to use information on the score in the midterms to “calibrate” their study strategy in the final exam in order to maximise the probability to pass the course. When we assess the validity of our research design, we find that the distribution of the greatest midterm score and of predetermined characteristics, such as gender and socio-demographics are balanced at the 5 cutoff after dropping observations of students scoring in a small neighbourhood of the cutoff. Our results are robust to adding controls, varying the RDD bandwidth or using quadratic specifications.

Our paper contributes to the literature with a case study in education in which a positive signal in midterm tests plays a predominantly psychological role by broadening our knowledge on the behavioural underpinnings of students’ choices. It complements related studies on the effect of modifying the information disclosed on achievement or the assessment rules in higher education by way of field experiments (Azmat et al., 2019; Chevalier et al., 2018), and contributes to the literature on left-digit effect in the context of education. In addition, our paper speaks to studies on the Math gender gap by suggesting that gender differences in processing information on Math achievement are not relevant mechanisms to explain part of this gap.

The remainder of the paper is structured as follows. Section 2 reviews the related literature. Section 3 presents the institutions and data. Section 4 describes the research design and assesses its validity. Section 5 describes our results and, finally, section 6 discusses them and concludes. Additional results are reported in the Appendix.

2 Literature review

Our paper is related to studies which test the effect of barely passing end-of-secondary school tests, i.e., high-stakes tests, on students’ decisions to attend post-compulsory education as they have in common exploiting cutoff scores in the tests with a regression discontinuity design. Papay et al. (2010) study whether barely passing high school final tests in English or Math has an impact on the probability of high school graduation for students in Massachusetts. They find that barely passing the Math test increases the probability of graduation, but only for poor students while barely passing the English test has no effect. In related studies in US the same authors find that barely passing one or more high school final tests increases the probability of attending college (Papay et al., 2011, 2014, 2015). A similar result is found for the UK. Students whose score in the high school final test in English is just above a relevant cutoff are more likely to attend the two years long post-compulsory education cycle which precedes university, called A-level, to enrol later on in a university degree and to obtain higher labour market returns (Machin et al., 2020). Overall, this evidence suggests that cutoff scores in high stakes exams may introduce a subtle but potentially important source of inequality, worthy to investigate further.

Azmat et al. (2019) is perhaps the only study which, alongside ours, tests the role of information on performance in tests on subsequent achievement in a university degree setting. While we use information routinely disclosed to students after the midterms and look at the mainly psychological effect of scoring at least 5 in the low stakes setting of a midterm test, they study the effect of giving additional information only to some students on their relative performance on their grade point average (GPA) in the future. Thanks to a field experiment carried out at the Carlos III University in Madrid over three years, they find that in the first two years information GPA and the number of exams passed are lower for treated students. Additional analysis shows that these results are driven by students underestimating their real position in the grade distribution. However, catch up is observed in year three when differences between treated and controls are small and no longer significant.

In a related work on college students, Chevalier et al. (2018) study the effect of incentivising online quizzes on study effort and on performance in the final exam of a first year introductory course taught in degrees in Social Science at a large college of the University of London. They find that making online quizzes count in the course assessment increases students' effort in the quizzes and also their performance in the exam. Differently, in our setting midterm tests are compulsory and count for the final score in the course, which is advantageous because we are not forced to alter a course assessment rules over time to study our effect of interest. In addition, in our setting we focus our attention on students whose first midterm score is arbitrarily close to 5, with the advantage relative to Chevalier et al. (2018) of being able to assume away differences in ability, i.e., high internal validity, and the disadvantage that our results only apply to those students whose first midterm score is close to 5, i.e., low external validity. Our results are in line with Chevalier et al. (2018)'s since in both an intermediate assessment in a course has a positive impact on subsequent assessment in the course, although the two settings are not fully comparable.

In addition, our paper is related to a number of studies in social science ranging from Marketing to Economics that have documented the empirical relevance of the left-digit effect, i.e. a predominantly psychological effect inducing different decisions in people exposed to numbers whose integer part differs although only for a small difference in the decimal part. We believe our paper is closely related to two of them since they report evidence of a left-digit effect in the context of education. Goodman et al. (2018) study the effect of re-taking college entrance exams. Differently from our manuscript that focuses on first year college students, the authors look at students aspiring to go to college. They estimate the effect by exploiting the empirical regularity of a higher percentage of re-takers among those whose score in the first attempt is barely smaller than multiples of 100 (relative to those whose first attempt score is barely greater than 100) and find that retaking leads to a higher score and to improved college enrollment outcomes. Olsen (2013) is a related although different study whose objective is testing whether subjective views over the performance of a school, measured as a grade average of pupils' achievement, differ for schools scoring on a continuous scale barely to the left rather than to the right of an integer value. The main result is a more positive view of schools whose performance level is just to the right of an integer relative to just to the left. Although in different contexts, this study and our manuscript have in common a psychological effect of scoring barely to the right rather than to the left of a relevant integer number due to a few decimals. This left-digit effect belongs to the more general class of behavioral biases discussed by Kahneman et al. (2021). These authors provide compelling examples showing that many choices are altered by the presence of randomness.

Finally, while no study to the best of our knowledge has looked at gender differences in the effect of barely passing a test, several looked more broadly at gender differences in achievement in compulsory and post-compulsory education. A candidate mechanism put forward to explain higher Math performance for males has been competition. Niederle and Vesterlund (2010) survey the studies testing for the effect of competition on Math scores and find that gender differences tend to be explained by lower competitiveness for girls. In partial contrast, Bandiera et al. (2018) find no gender differences in a meta-study of experimental research on the effect of performance pay. We contribute to these studies by looking at gender-specific responses to positive signal in the form of a tenuous informational shock about achievement as it is not clear-cut whether the (less) more competitive gender, i.e. (fe)males, responds more strongly to it.

3 Institutions and data

We estimate the effect of barely passing at least one midterm test thanks to a novel dataset which we created by linking administrative data on achievement and survey data with additional information on socio-economic characteristics for students in the first Math course in the degrees of Business Economics and in Economics at the University of Alicante over three academic years: 2016-17, 2017-18 and 2018-19. Students in these degrees account for about 90% of all students in degrees offered by the School of Economics.Math is a compulsory course that students take in the first term which starts in September and ends in December, in the first year of the degree. The course content is calculus using one-variable functions. The course syllabus is divided into the following four blocks: introduction to one-variable functions, differentiation, integration and numeric sequences and, finally, elements of linear algebra and systems of linear equations. This is very similar to the syllabus of an introductory Math course for social science degrees held in universities all over the world. The course consists of four contact hours per week over 15 weeks, two of which are devoted to theory and two to tutorials consisting in solving exercises. Office hours are available for students on a weekly basis.

We believe that the Math course is well suited to test the effect of a positive signal in a midterm. First, the first midterms in the course are not just another midterm for the students, as they are one of the very first tests taken by students in the degree. Hence, students tend to pay considerable attention to achievement in these tests to update their beliefs over own ability and the necessary study effort to pass the Math course, as well as other courses. Second, a positive relationship between Math achievement, for example, at high school, and success in a university degree and in the labour market has been found in a number of empirical studies (Levine and Zimmerman, 1995; Rose and Betts, 2004; Joensen and Nielsen, 2009).

The course assessment consists of two midterm tests held during the course, T_1 and T_2 hereafter, and the final exam (F) that is held in January every year. The overall score (O) in the course is a weighted average of scores in the tests and in the final exam given by the following formula: $O = 0.2T_1 + 0.3T_2 + 0.5F$. T_1 is held in week 5 or 6, i.e. typically in mid-October, while T_2 is held in week 11 or 12, i.e. typically in the second half of November. The content of T_1 is material from the first block in the syllabus, while that of T_2 is material from the first three blocks. Finally, in the final exam, knowledge of material in all four blocks is assessed. Scores in all three assessments are reals in the interval 0-10. The grading system at the University of Alicante sets the pass cutoff at 5 and is the same as in all other universities in Spain. In addition, it attributes specific values to the following scores: those in the interval 7-8.99 are considered high pass and from 9 onwards very high pass. An important feature of the first Math course, that we chose as proof of concept to test the role of a positive signal in a midterm, is that the performance in the midterms has no consequence on subsequent assessment as all students take the same final exam.

Students are divided every year into teaching groups, with approximately 40 students per group and the division in groups being the same for all first-year courses. Math instructors typically teach one or more groups, with the maximum in our dataset being 3 in the period we consider. Instructors set midterm tests following the same guidelines on the test material from the course coordinator although, since tests for different groups are held on different dates, they may differ in an attempt to avoid spillover effects from those groups who already took it to those who have not. The final exam is, instead, identical for all students as it is held simultaneously for all students. Instructors mark both the midterm tests and the final exam. Since our treatment of interest is whether students achieve a score greater than or equal to 5 in at least one of the two tests, i.e. T_1 and T_2, and they are marked by students' own instructors, we will carefully discuss in the next sections how the practice by instructors to grade with a 5 a high enough number of tests close to it, which leads to a “jump” at the cutoff in the test score distribution, may affect our results.

We obtained administrative data on students and kept only observations of students who took the two tests a well as the final exam. This simplifies the analysis as otherwise we would have to face the tricky choice of whether to assign a zero score to those absent in a test or in the final exam. We also conducted surveys to obtain information on students' socio-demographics. Students gave consent to link their administrative data on achievement with survey data when the survey questionnaires were administered. They could opt out of the survey, and if they did, we did not include data from these students in our dataset. Students opting out are about 10% and almost all of them did so because they were absent either in more than a midterm or in the final exam. Table 1 reports summary statistics of our main outcomes of interest and of students' predetermined characteristics. This information is shown for the entire sample in column (1), and separately for those students not achieving the pass score 5 in at least one of the two tests, i.e. with the maximum score between T_1 and T_2 being smaller than 5 in column (2), and for those achieving it or exceeding it in column (3). Column (4) reports the p-value of the null hypothesis of no difference in the mean values in a given variable for the two sub-samples. Similarly, we report in column (5) the mean values for students not achieving the pass score in at least one test by at most one point score, i.e. whose maximum score between T_1 and T_2 is between 4 and 5, and in column (6) the mean value for those students achieving it by at most one point score, i.e. whose maximum score between T_1 and T_2 is between 5 and 6. The p-value of the null hypothesis of no difference in the mean of these two sub-samples is reported in column (7).[1]

The top panel in Table 1 shows in column (1) that out of all in our dataset, i.e. those who sit both tests and the final exam, 51% obtain a score greater than or equal to 5 in T_1, 40% in T_2 and 33% in the final exam (F). Only 39% of students pass the course as their overall score (0) is greater than or equal to 5. These shares as well as scores in midterm tests and in the final exam are higher for students with

[1] Table A.1 in the Appendix reports cross-tabulations of the number of students by whether their score in $T_1 \geq 5$ and $T_2 \geq 5$.

Table 1: Summary statistics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	All	$\bar{T} = max(T_1, T_2)$		p-value	$4 \leq \bar{T} < 5$	$5 \leq \bar{T} \leq 6$	p-value
		< 5	≥ 5	(2)=(3)			(7)=(8)
Scores in tests, final exam and overall score							
Test 1	4.70	2.53	6.15	0.00	3.83	4.90	0.00
S.d.	2.39	1.40	1.72		0.93	0.75	
$T_1 \geq 5$	0.51	0.00	0.85	0.00	0.00	0.75	0.00
Test 2	4.16	2.22	5.45	0.00	3.36	4.28	0.00
S.d.	2.38	1.36	2.02		1.17	0.37	
$T_2 \geq 5$	0.40	0.00	0.67	0.00	0.00	0.52	0.00
Final (F) score	3.79	2.23	4.83	0.00	3.11	3.99	0.00
S.d.	2.40	1.73	2.22		1.88	0.10	
F $\geq$ 5	0.33	0.07	0.51	0.00	0.18	0.37	0.00
Overall score (O)	4.24	2.63	5.32	0.00	3.38	4.31	0.00
S.d.	2.12	1.64	1.67		1.13	0.33	
O $\geq$ 5	0.39	0.09	0.59	0.00	0.09	0.36	0.00
Control variables							
Female	0.47	0.48	0.47	0.78	0.48	0.45	0.39
Foreigner	0.10	0.12	0.09	0.09	0.09	0.10	0.92
Bus. Adm. (BA) deg.	0.60	0.60	0.60	0.90	0.64	0.61	0.53
BA+Law deg.	0.09	0.14	0.07	0.00	0.06	0.07	0.40
BA+Tourism deg.	0.13	0.12	0.13	0.58	0.15	0.14	0.77
Economics degree	0.18	0.14	0.20	0.00	0.16	0.18	0.57
Dummy for Year 2018	0.35	0.33	0.36	0.25	0.27	0.33	0.14
Dummy for Year 2019	0.29	0.35	0.24	0.00	0.34	0.29	0.18
Repeater	0.14	0.15	0.13	0.18	0.12	0.18	0.07
N	1,873	750	1,123		233	444	

at least one test score greater than or equal to 5, as shown by means column (3) and by small p-values in column (4). We find similar differences when we look at students whose greatest score between T_1 and T_2 is within one point score on either side of the cutoff in columns (5)-(7), except for smaller magnitudes.

The bottom panel of Table 1 shows the control variables that we obtained thanks to the surveys we conducted and we use in our empirical analysis. When we look at students' socio-demographics in column (1), we find that about 47% are females and 10% are foreigners. When we look at students' degrees, we find that 60% of them are in the Business administration (BA) degree, while the remaining students are somewhat evenly split between double degrees, i.e. BA and Law and BA and Tourism, and the degree in Economics. In the full sample students in the BA and Law and those who took the Math course in 2019 are under-represented in the group with at least one of the two scores greater than or equal to a pass while Economics students are over-represented, as suggested by low p-values in column (4) of the difference between the means between columns (2) and (3). However, when we look at differences for the subgroup of students whose greatest score in the tests is within one point score from the cutoff in columns (5)-(7), we find reassuringly that no difference in predetermined characteristic is significant.

In the next section we assess whether the significant association between the final exam score and the probability that the score in at least one of the two tests is greater than 5 can also be interpreted causally. This is particularly important, given the significant differences we observe in some predetermined characteristics in our full data sample.

4 Research design

In this section we, first, describe the details of our research design and, second, we discuss its validity and offer evidence in support of it.

4.1 Regression discontinuity design

We let T_1 and T_2 denote students' scores in the first and second test that were held during the Math course, with its support being reals in the interval 0-10. We also define $T_1^* = T_1 - 5$ and $T_2^* = T_2 - 5$ by subtracting 5 from T_1 and T_2 to rescale them in such a way that negative values of T_j^* indicate that a student scored less than 5 in test j and vice versa for positive T_j^* values. Also let $P_j = I(T_j^* \geq 0)$ be a dummy equal to 1 if the score in the test j is greater than or equal to 5. We denote it with the letter P_j to highlight the psychological component associated with scoring at least 5 in test j. Since students take two tests, we define the variable $\bar{T}^* = max(T_1^*, T_2^*)$ to measure the greatest score in the two tests and $\bar{P} = (\bar{T}^* \geq 0)$ a dummy variable equal to 1 if the greatest score over the two tests, i.e. in at least one of the tests, is greater than or equal to 5. Finally, Y denotes subsequent achievement, i.e. the final exam.

Equation (1) shows that in a linear regression of Y, the dependent variable, on $\bar{P}$, our parameter of interest, α_1, measures the difference in subsequent achievement between those students obtaining a score greater than or equal to 5 and those obtaining a score smaller than 5 in the first test.

$$Y = \alpha_0 + \alpha_1 \bar{P} + U_1 \tag{1}$$

However, estimates of α_1 in equation (1) may be biased since $\bar{P}$ is likely to be correlated with unobservable variables we cannot control for. A typical example is unobserved ability, which is likely to differ for students whose T_1 and/or T_2 scores are substantially greater or smaller than 5. By using, instead, a regression discontinuity design with as running variable the maximum score in the tests $\bar{T}$ we identify the psychological effect by comparing subsequent achievement only for students whose score in at least one of the tests is equal to or barely greater than 5 and for those whose maximum score in the tests is barely smaller than 5. The reason is that by considering only students with at least one score arbitrarily close to the cutoff 5, being on either side of the cutoff is roughly due to chance and hence students on either side have very similar observed and unobserved characteristics, e.g., ability.

We estimate the psychological effect, which is captured by parameter β_1 in equation (2) by using a flexible polynomial in $\bar{T}^*$, the maximum score in the tests after subtracting 5 from it, i.e., the running variable, and allowing the polynomial to be different to the right and to the left of the cutoff. We use both linear and quadratic approximations of the polynomial in T^*. We also add as controls in the regressions students' predetermined characteristics, whose full list can be found in Table 1.

$$Y = \beta_0 + \beta_1 \bar{P} + \beta_2 \bar{T}^* + \beta_3 \bar{T}^* \times \bar{P} + \beta_4 \bar{T}^{*2} + \beta_5 \bar{T}^{*2} \times \bar{P} + U_2 \tag{2}$$

In addition, we use different values of the midterm score bandwidth, i.e. how far away is a student's $\bar{T}^*$ score from 0, to only consider those students whose scores are very close to the cutoff score 5. The bandwidth choice implies a trade-off. With a large value, many observations are included, but students with scores much higher and much lower than the cutoff cannot be regarded as being similar. In contrast, with a small bandwidth value, only students whose $\bar{T}^*$ score is very close to the cutoff are considered, thus leading to a smaller number of observations and a lower test power. We consider several choices for the bandwidth, although our starting point is the optimal bandwidth obtained with the method proposed by Imbens and

Kalyanaraman (2012).[2]

4.2 Research design validity

In this subsection we assess the validity of the research design by quantifying whether the distribution of students' greatest score in the tests, i.e. the running variable $\bar{T}^*$ in our RDD, and their predetermined characteristics are balanced at the cutoff 5. Figure 1 shows in the top panel histograms of the running variable. The histogram on the left-hand side shows a jump in the frequency of students whose greatest score in the test is exactly 5 while the one on the right hand-side shows that, once we have removed observations only for these students from the dataset, the histogram is continuous at the cutoff. The histograms also show that frequencies tend to be higher at integer values other than 5, e.g. 6 and 7, which may suggest that some instructors may round up decimal scores to the closest integer.

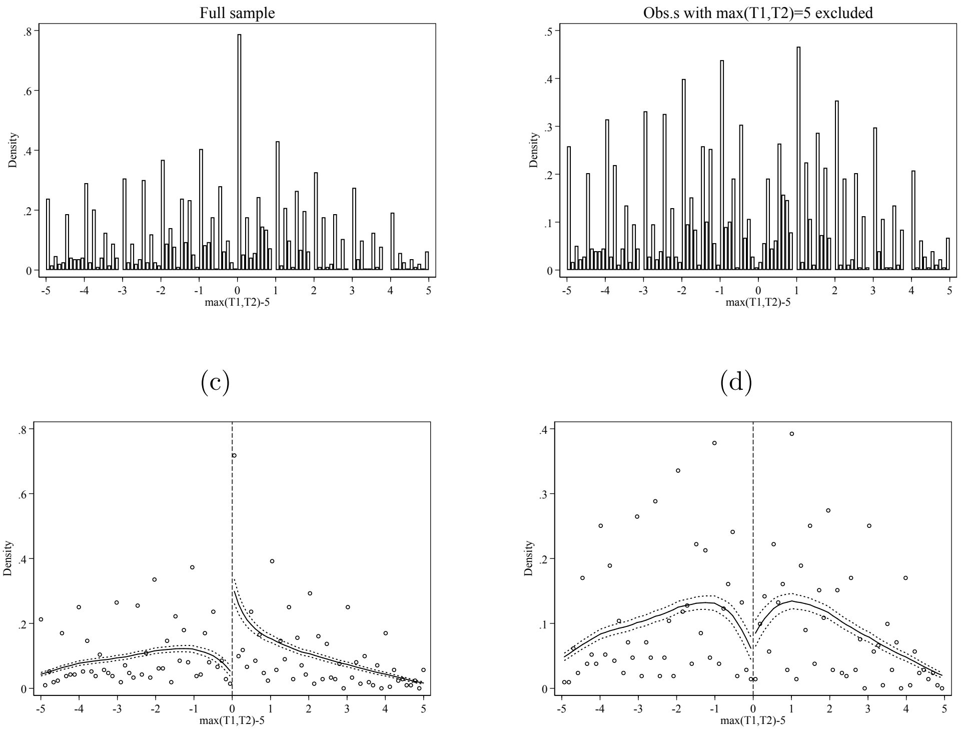
The bottom panel in Figure 1 shows a kernel density plot of $\bar{T}$, along with 95% confidence intervals reported as dashed lines (McCrary, 2008). The results show that once observations of students achieving 5 in at least one midterm are removed, a jump in the density is no longer observed in line with the histograms shown in the top panel. This suggests that the target of potential manipulation of scores by teachers is limited to students scoring slightly below or above 5, with the former seemingly being more frequently subject to rounding up to the closest integer.[3]

In order to deal with this evidence of manipulation around the cutoff 5, our estimation results in the next section are based on the so-called "donut" RDD approach suggested by Barreca et al. (2011, 2016). This procedure consists in dropping not only observations just to the right of the cutoff but also on the left of it by creating a "donut hole" at the cutoff. This ensures that the observations which are dropped are not only those potentially manipulated, which are typically to the right of the cutoff in our setting, but also those that are just on the left. Those may have not

[2] Most of the existing methods start by noting that the choice of the bandwidth is analogous to choosing the polynomial order to approximate the polynomial in T^*. The optimal bandwidth value is then derived by minimizing the mean square error when using local polynomial regression. In particular, Imbens and Kalyanaraman (2012) provide an algorithm to consistently estimate the unknown functionals on which the (infeasible) optimal bandwidth relies. This complex procedure is popular among practitioners since it is implemented in Stata and other statistical software.

[3] A histogram of $\bar{T}$ reporting the number of observations per histogram bin can be found in Figure A.1 in the Appendix.

Figure 1: Test score histograms



(a)

(b)

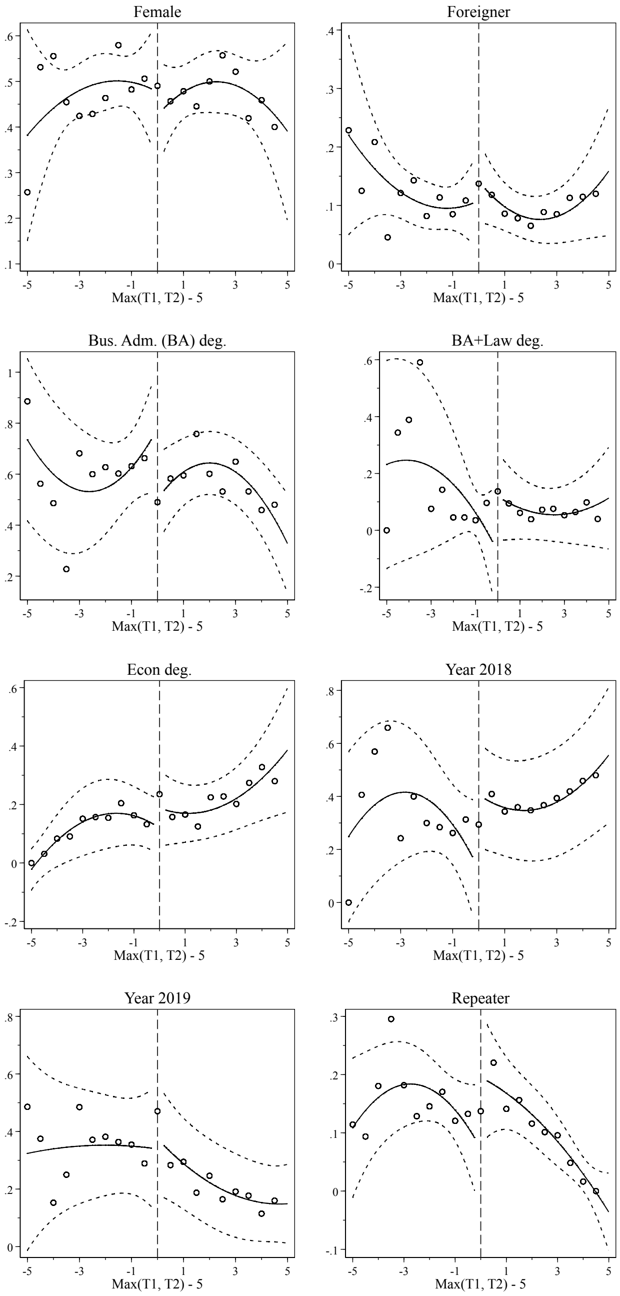
(c)

(d)

been manipulated partly because of observable characteristics and partly because of unobservable ones, and this may contribute to bias RDD estimates.

We therefore assess whether students' baseline characteristics are balanced at the cutoff 5 in the running variable, i.e. the maximum value between scores in T_1 and T_2. Figure 2 shows plots of second order polynomials in $\bar{T}$ of students' baseline characteristics. The polynomials have been fitted separately to the left and to the right of the cutoff and after dropping observations for students whose greatest score is within a radius of 0.1 around the cutoff, what has been termed a "donut". This ensures that students arbitrarily close to the cutoff on either side are excluded as potential manipulation shifts students in a small "donut" from one side of the cutoff to the other, thus leading to biased estimates. We do not observe in Figure 2 substantial jumps in any baseline characteristic or significant ones at the cutoff, as shown by overlapping confidence intervals reported as dashed lines. This offers evidence in support of the research design validity. Estimates of differences in individual baseline characteristics are reported in Table A.2 in the Appendix. They tend to be in line with the results in Figure 2, although they are significant for some

Figure 2: Baseline characteristics balance at $max(T_1, T_2)$ cutoff



bandwidth values. Reassuringly, when we test whether differences in all baseline characteristics are jointly zero, by way of a seemingly unrelated model with as many outcomes as the number of baseline characteristics, large p-values of the joint test in Table A.3 in the Appendix show that we do not reject the null hypothesis.[4]

[4] We have also carried out similar balancing tests using richer survey data but available only for 2017/18 and 2018/19. These results, available upon request, show again no statistically significant difference in students' characteristics, which now include High School majors and the ratio people per room; only gender and having a major in Arts or Humanities are among the few exception to this general result.

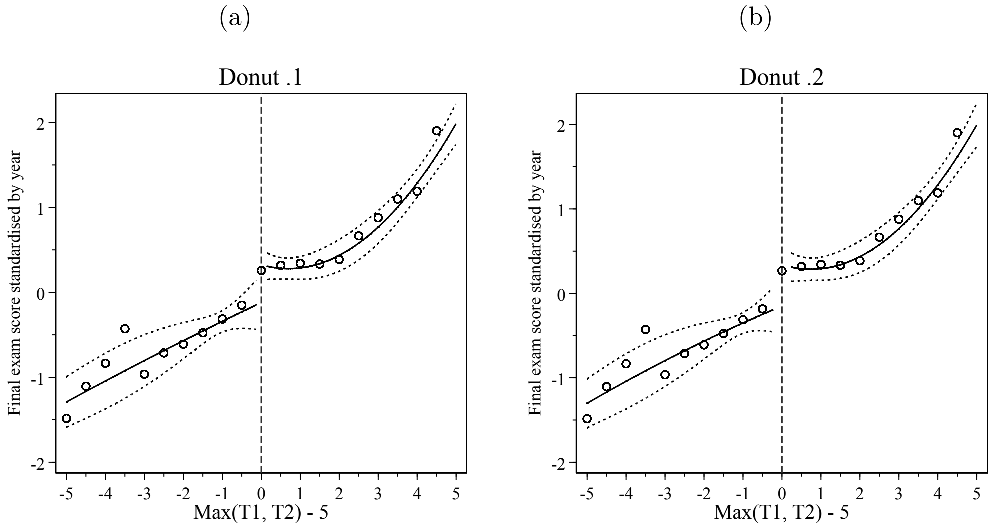
5 Results

This section starts by reporting donut RDD estimates of the effect of barely passing at least one midterm exam (section 5.1). Subsequently, we report results from a sensitivity analysis showing the robustness of our main results to varying the bandwidth (section 5.2).

5.1 Main results

We start reporting our main results by giving a graphical overview in Figure 3, that reports RDD plots with on the vertical the final exam score standardized by academic year, and on the horizontal axis the maximum midterm score in tests centered at 5. We standardized the final exam score to make our estimates comparable to similar studies and discuss their relative size. Panel (a) shows a discontinuous increase in the final exam score at the cutoff 5 for the maximum score, with the psychological effect estimate obtained after dropping observations in a donut with a 0.1 radius. Similar results are reported in panel (b), obtained after dropping observations in a donut with a 0.2 radius.

Figure 3: RDD plots of the psychological effect on the final exam score standardized by academic year



The estimates reported in this section were obtained by using a donut RDD which is the most "conservative" in dealing with a higher frequency of students just to the left of the cutoff as they exclude from the dataset students in a small neighbourhood around it. Our preferred specification excludes from the dataset observations for students whose scores are in an interval of 0.2 both above and below such data heap, i.e. between 4.8 and 5.8. This ensures the continuity of the running variable and of baseline characteristics at the cutoff 5, as previously discussed in section 4.2. In addition, estimates using alternative values of the donut radius to assess the sensitivity of our main results show similar results and can be found in Table A.4 and A.5 in the Appendix.

Table 2 reports estimates of the effect of a positive signal in at least one midterm on the final exam score standardised by year. We report estimates obtained from a local linear regression in $\bar{T}^* = max(T_1^*, T_2^*)$ using the optimal bandwidth value (columns 1 and 2), which was obtained following the procedure in Imbens and Kalyanaraman (2012) previously discussed, as well as using a greater value, equal to 1.5 times the optimal one (columns 3 and 4), and also from a second order polynomial using all observations, i.e., bandwidth 5 (columns 5 and 6). All estimates are reported from regressions without and with controls, whose coefficients are not reported. Controls used in the regression include dummies to account for the following socio-demographics: gender, foreign nationality, whether a student repeats the course and degree type, as reported in Table 1. Bandwidth values and information about whether a regression was run with controls are reported at the bottom of the table.

In Table 2, the psychological effect of barely passing at least one midterm is captured by the coefficient associated with the dummy $\bar{P} = (\bar{T}^* \geq 0)$. This estimated coefficient is positive across all the different specifications, and it implies a greater final exam score by 0.25 to 0.30 standard deviations around the 5 cutoff. The size of this effect is in line (although slightly larger) with the magnitude found in the literature for other educational interventions, in particular those related to the effect of positive feedbacks. [5] The corresponding 95% confidence intervals reassuringly

[5] We do not report estimates of the effect of scoring at least 5 in one of the tests on the probability of obtaining a score greater than or equal to 5 in the final exam as they are positive but small and not significant. In addition, we do not report estimates obtained using the score in the first test as running variable, using the following outcomes: the score in the second test, the final exam score or on the probability of obtaining a score greater than or equal to 5 in the final exam, as they are positive although small and not significant. Similarly, we do not report estimates from a larger dataset including students not sitting the final exam, as the psychological effect on the probability of sitting the final exam is not significant, neither when using as running variable the first test score nor the maximum score over the two tests. However, they are available upon request.

Table 2: Effect of a positive signal in at least one midterm test on the final exam score standardized by year (donut with a 0.2 radius)

	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{P}$	0.263**	0.240**	0.214*	0.165*	0.475**	0.326**
	(0.118)	(0.117)	(0.112)	(0.097)	(0.182)	(0.126)
$\bar{T}^*$	0.343***	0.351***	0.263***	0.290***	0.193	0.303***
	(0.051)	(0.047)	(0.072)	(0.055)	(0.170)	(0.095)
$\bar{T}^* \times \bar{P}$	-0.277***	-0.253***	-0.075	-0.078	-0.315	-0.349**
	(0.097)	(0.085)	(0.084)	(0.072)	(0.199)	(0.142)
$\bar{T}^{*2}$					-0.007	0.010
					(0.024)	(0.015)
$\bar{T}^{*2} \times \bar{P}$					0.097***	0.065**
					(0.035)	(0.026)
Constant	-0.011	0.173	-0.092	0.113	-0.150	0.106
	(0.102)	(0.129)	(0.114)	(0.148)	(0.168)	(0.151)
N	998	998	1,356	1,356	1,693	1,693
Controls	No	Yes	No	Yes	No	Yes
Bandwidth	2.22	2.22	3.33	3.33	5.00	5.00

Note: OLS estimates of the model specified in equation (2) with: final exam score standardised by year; 0.2 donut radius (i.e., excluding students whose value of $max(T_1, T_2)$ is within 0.2 point scores from the cutoff 5); clustered standard errors at the class group and year level in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

contain only positive values for the optimal bandwidth value, the smallest value in columns (1) and (2) in Table 2 as for estimates obtained using the full bandwidth of 5 and a quadratic specification in columns (5) and (6). The zero is only included in confidence intervals for intermediate bandwidth values while in columns (3) and (4).Also reassuringly, we obtain similar results in Tables A.4 and A.5 in the Appendix when we exclude, first, observations for students with a least one test score equal to 5 and, second, when we exclude, following Barreca et al. (2011, 2016), observations in a neighbourhood (or "donut radius") of 0.1 point scores around the 5 cutoff. Overall, our estimates suggest that our hypothesis of a psychological effect on subsequent achievement is confirmed by the data, and it is robust to different specifications.

Subsequently, we move to assess whether the overall psychological effect exhibits heterogeneity by gender. Table 3 reports estimates using the specifications as in

Table 3: Effect of a positive signal in at least one midterm, by gender (donut with a 0.2 radius)

	Final Exam Score standardised by year					
	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{P}$	0.184	0.148	0.174	0.078	0.421*	0.199
	(0.162)	(0.148)	(0.153)	(0.120)	(0.224)	(0.152)
$\bar{T}^*$	0.436***	0.431***	0.245***	0.289***	0.217	0.359***
	(0.075)	(0.068)	(0.083)	(0.060)	(0.189)	(0.110)
$\bar{T}^* \times \bar{P}$	-0.434***	-0.381***	-0.043	-0.055	-0.359	-0.399**
	(0.137)	(0.120)	(0.102)	(0.085)	(0.243)	(0.187)
$\bar{T}^{*2}$					-0.003	0.019
					(0.028)	(0.018)
$\bar{T}^{*2} \times \bar{P}$					0.101**	0.058*
					(0.041)	(0.029)
Female (F)	-0.149	-0.169	0.079	-0.027	0.003	-0.139
	(0.178)	(0.143)	(0.140)	(0.118)	(0.181)	(0.157)
$F \times \bar{P}$	0.171	0.199	0.098	0.190	0.113	0.266
	(0.238)	(0.197)	(0.210)	(0.181)	(0.266)	(0.219)
$F \times \bar{T}^*$	-0.195	-0.165	0.026	-0.004	-0.043	-0.118
	(0.122)	(0.102)	(0.063)	(0.057)	(0.149)	(0.132)
$F \times \bar{T}^* \times \bar{P}$	0.324*	0.257*	-0.060	-0.042	0.080	0.101
	(0.170)	(0.143)	(0.094)	(0.088)	(0.224)	(0.206)
$F \times \bar{T}^{*2}$					-0.006	-0.019
					(0.028)	(0.025)
$F \times \bar{T}^{*2} \times \bar{P}$					-0.009	0.015
					(0.048)	(0.039)
Constant	0.061	0.238	-0.135	0.113	-0.150	0.163
	(0.134)	(0.155)	(0.143)	(0.164)	(0.194)	(0.170)
N	997	997	1,353	1,353	1,687	1,687
Controls	0	1	0	0	0	1
Bandwidth	2.22	2.22	3.33	3.33	5.00	5.00

Note: Clustered standard errors at the class group and year level in parentheses.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2, but the effect is here allowed to vary by gender. This is done by adding a dummy equal to 1 for females and by interacting all terms of the RDD polynomial with the female dummy. Hence, the psychological effect for boys is captured by the coefficient associated to the dummy $\bar{P}$ while the difference for girls relative to boys is captured by the one associated to the interaction between these two dummies, $Female \times \bar{P}$. The table shows that point estimates of the psychological effect for boys are similar (or slightly lower) in magnitude to those obtained in the previous Table 2, although less precisely estimated. When we look at gender differences, we find that the difference in the psychological effect between girls and boys is positive, but not statistically significant. In other words, girls seem to react to a positive signal more positively than boys, but it cannot be excluded that they are both affected in the same manner.

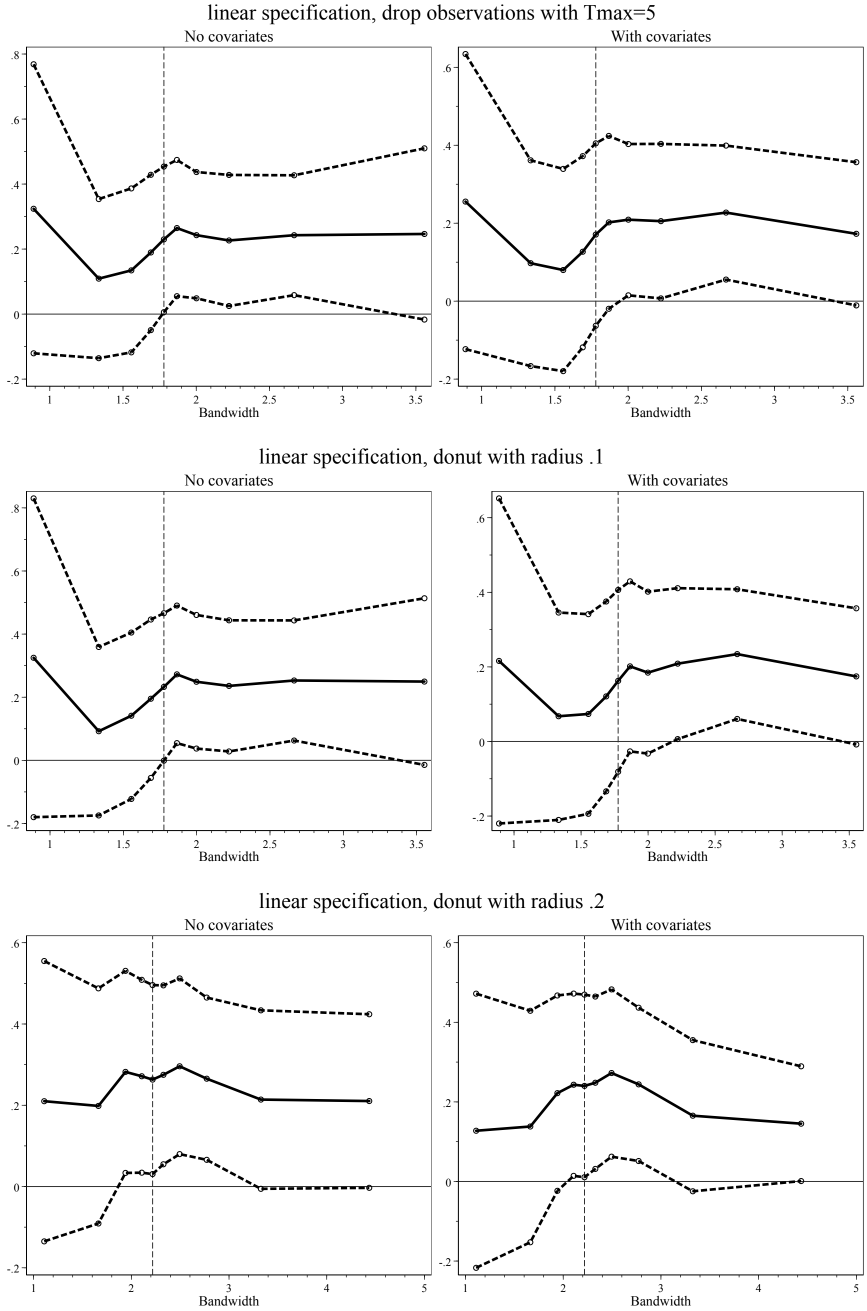
5.2 Sensitivity analysis

In this section we report evidence from a sensitivity analysis whose objective is assessing the robustness of our main results of Table 2. We begin by assessing the sensitivity to using a large variety of bandwidth values different from the ones used in our preferred specification in Table 2, i.e. the optimal one, 1.5 times the optimal one and the full bandwidth of 5. Figure 4 shows how estimates of the effect of a positive signal in at least one midterm, measured along the vertical axis, vary when we vary the bandwidth, measured along the horizontal axis.

Figure 4 is divided into two columns and three rows, giving us six plots in total. Each one of these plots reports estimates from a regression of the positive signal effect, i.e., estimates of the parameter associated to the dummy $\bar{P}$ in Table 2 (thick continuous line) along with 95% confidence intervals (thick dashed lines). Estimates come from a linear polynomial specification in the running variable $\bar{P}$ for a bandwidth up to 4; the optimal bandwidth is drawn using a vertical dashed line. In each case, we consider specifications without covariates and with covariates. Finally, we report separate specifications in different panels, depending on the observations dropped in the donut RDD specification: observations with a score of just 5 (top panel), with scores 0.1 points around 5 (central panel), and with scores 0.2 points around 5 (bottom panel).

Figure 4 shows that the positive and significant effect observed in our main results is robust to using different bandwidths and both without and with controls. Only

Figure 4: Sensitivity to bandwidth values used to estimate psychological effect on final exam score standardised by year (linear specification)



for small values of the bandwidth, confidence intervals are rather ample, so they are not significant at conventional levels. We believe that this is due to the low power to estimate our effect of interest because, first, for bandwidth values smaller than the optimal, we estimate an effect with few observations and, second, the loss of significance is mainly due to greater standard errors rather than to a decrease in the point estimate. As a matter of fact, the point estimates remain relatively unchanged across all the plots, and the confidence intervals mostly lie in the range of positive values. Estimates obtained using quadratic polynomial specifications are reported in Figure A.2 in the Appendix and show that a quadratic precisely estimates our effect of interest for relatively large bandwidths.[6]

Overall, the obtained effect size of 0.29 standard deviations is at the upper end of estimates on student's performance of teacher experience (De Paola, 2009) or of feedback (Bandiera et al., 2015), but it is in line with the effect of most of the literature on randomized evaluation of education policies (Banerjee et al., 2007), of class sizes (Angrist and Lavy, 1999) or of teacher quality (Rockoff, 2004). In particular, it is similar, although slightly higher, than the effects observed on positive feedbacks by Brade et al. (2018) and Azmat et al. (2019), and it confirms the findings of Azmat and Iriberri (2010) that showed that providing (relative) feedback improved high school students' grades by 5%.

5.3 Mechanisms

In this section, we explore potential channels that can explain our previous result, i.e., a positive psychological effect of obtaining a pass score in at least one midterm on the final exam score. We propose study strategies as one of the possible behavioural mechanism at play. We can expect that study effort increases or decreases as a result of the psychological effect of barely passing at least one midterm test. We compute the difference between actual score in the final exam and the score need to pass the course; each student can easily compute this with information about her previous performance. A positive difference can be regarded as an "objective" measure of effort exerted by the student, since she is doing more that what it is required to just pass the course.

Since students take two midterms, they may jointly use information on the score in the midterms to "calibrate" their study strategy in the final exam in order to maximise the probability to pass the course. Since the overall score in the course is

[6] We have also replicated our previous estimates using the subsample from 2017 to 2019. These results, available upon request, are rather similar to those obtained in the larger sample. However, the 90% and 95% confidence interval are wider, which is not surprising since the sample size is about one third lower. As a consequence, we cannot reject here that the psychological effect is not statistically significant. Since the overall picture is similar, we are confident that this subsample can be used for further analysis exploiting the additional information that it contains.

a weighted average of scores in two tests and in the final exam, i.e., $O = 0.2T_1 + 0.3T_2 + 0.5F$, once students know their scores in the midterm tests, they can compute the final exam score value F^* that grants them a pass in the course. This is obtained by setting the formula to compute the overall score in the course equal to 5, the pass score, and substituting out for F^*, i.e., $F^* = (5 - (0.2T_1 + 0.3T_2))2$. We then create as a proxy for student's study strategy the difference between a student's actual final exam score F and the individual-specific pass cutoff F^*, which is positive if a student obtains a final exam score greater than the pass cutoff, zero if it is equal to the pass cutoff value and negative if it is smaller.

Table 4: Effect of a positive signal in at least one test on targeting the final exam pass score (donut with a 0.1 radius)

	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{P}$	0.566*	0.547*	0.512	0.410	1.100**	0.831**
	(0.310)	(0.293)	(0.323)	(0.264)	(0.445)	(0.336)
$\bar{T}^*$	1.773***	1.740***	1.504***	1.570***	1.532***	1.697***
	(0.161)	(0.134)	(0.206)	(0.156)	(0.397)	(0.259)
$\bar{T}^* \times \bar{P}$	-0.763***	-0.585***	-0.182	-0.163	-1.152**	-1.083***
	(0.252)	(0.211)	(0.242)	(0.203)	(0.492)	(0.387)
$(\bar{T}^*)^2$					0.021	0.044
					(0.057)	(0.043)
$(\bar{T}^*)^2 \times \bar{P}$					0.245***	0.178**
					(0.087)	(0.072)
Constant	-2.077***	-1.716***	-2.353***	-1.735***	-2.335***	-1.700***
	(0.325)	(0.523)	(0.357)	(0.577)	(0.439)	(0.569)
N	1,062	1,062	1,393	1,393	1,717	1,717
Controls	No	Yes	No	Yes	No	Yes
Bandwidth	2.29	2.29	3.43	3.43	5.00	5.00

Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Once we have built a measure of students' ability to "calibrate" their study effort in the final exam F, we estimate the effect of a positive signal in at least one test on $F - F^*$. Table 4 reports these estimates and shows that the effect on $F - F^*$, captured by the coefficient associated with the dummy $\bar{P}$ is positive and significant at conventional levels for most specifications. In addition, the confidence intervals again contain basically positive values in all the cases. In our preferred specification in column (2), with covariates and optimal bandwidth, the effect is significant at 10% and the 95% confidence interval is $(-0.027, 1.121)$; and in column (4) with a wider bandwidth $(-0.107, 0.927)$. The upper confidence limit shows that the psychological effect induces students to obtain scores in the final exam up to one point score higher than the minimum final exam score necessary to pass the course while the lower one is negative although it is very small, about a ninth of the upper limit in absolute value. These results show that students affected by the positive signal seem to use more precisely the information obtained in the form of test scores in the preparation for the final exam and suggest study strategies as the most relevant mechanism to interpret our main results of the psychological effect on subsequent achievement. We have also checked whether this behavioural mechanism operates differently according to gender. Results available upon requests show that this mechanism is the same for boys and girls.

6 Conclusion

In this paper we tested, by using administrative and survey linked data on the first Math course in all degrees offered by the School of Economics at the University of Alicante, whether obtaining a score barely greater than or equal to 5 in at least one of the two midterm tests has an impact on students' subsequent achievement in the course. We use 5 as the relevant cutoff since it is a pass score in the nationwide grading system. We identify the effect using a regression discontinuity design, because obtaining a score barely greater or smaller than this cutoff is at least partially due to randomness, and so students with scores on either side of the cutoff have similar characteristics. We interpret the estimated effect as a mainly psychological one, as although bearing no direct consequence for the overall achievement, those scoring 5 or slightly above are more likely to interpret it positively and potentially modify their beliefs and subsequent study behaviour. Our results show that obtaining a passing score in at least one midterm has overall a positive and significant effect on subsequent achievement in the Math course, measured as the score in the final exam. However, we find no differences in how boys and girls react to this signal.

Although the positive effect on final exam score for those barely scoring at or above 5 in at least one midterm exam can also be explained from teachers' side, we propose study strategies as the main behavioural mechanism at play. Since the overall score in the course is a weighted average of midterm scores and the final exam, students can jointly use information on their two midterms score to compute the final exam score value that grants them a pass in the course. Our results show that students affected by the positive signal seem to use more precisely the information obtained to prepare the final exam: they can get up to one point score higher than the minimum final exam score required to pass the course. If the final exam score had been increased by teachers to benefit students who performed marginally better in the midterms, we feel that such increase would have been just enough to barely pass the overall course. Instead, our evidence suggest that students increased their effort, so their final score was well-above the required grade to pass. Again, this behavioural mechanism does not operate differently according to gender.

Our analysis may be enriched in a number of directions that we plan to pursue in future research. First, since the first Math course has a pass rate lower than 50%, we believe that it would be important to test the effect of a positive signal in achievement not only in Math but also in other first-year courses as this may influence students' overall achievement by the end of the first year and their decision to enroll in the second year in the degree, to change degree or to dropout. Second, we plan to obtain in the future data from a midterm test with multiple-choice questions which will enable us to test whether the effect differs in a setting in which there is no scope for score manipulation. Finally, it would be very informative to obtain data with enough variation in instructors' gender, ideally from several universities in Spain and in other comparable countries, to quantify whether the psychological effect is at least partially explained not only by students' gender but also by instructors'.

References

Angrist, J. and Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. The Quarterly Journal of Economics, 114(2):533–575.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (2019). What You Don't Know Can't Hurt You? A Natural Field Experiment on Relative Performance Feedback in Higher Education. Management Science, 65(8):3714–3736.

Azmat, G. and Iriberri, N. (2010). The Importance of Relative Performance Feedback Information: Evidence from a Natural Experiment using High School Students. Journal of Public Economics, 94(7-8):435–452.

Bandiera, O., Fischer, G., Prat, A., and Ytsma, E. (2018). Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments. Discussion Paper Series DP11724, Centre for Economic Policy Research.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful Ignorance? A Natural Experiment on the Effect of Feedback on Students' Performance. Labour Economics, 34:13–25. European Association of Labour Economists 26th Annual Conference.

Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India. The Quarterly Journal of Economics, 122(3):1235–1264.

Barreca, A. I., Guldi, M., Lindo, J. M., and Waddell, G. R. (2011). Saving Babies? Revisiting the effect of very low birth weight classification. Quarterly Journal of Economics, 126(4):2117–2123.

Barreca, A. I., Lindo, J. M., and Waddell, G. R. (2016). Heaping-induced Bias in Regression-Discontinuity Designs. Economic Inquiry, 54(1):268–293.

Brade, R., Himmler, O., and Jäckle, R. (2018). Normatively Framed Relative Performance Feedback – Field Experiment and Replication. MPRA Paper 88830, University Library of Munich, Germany.

Chevalier, A., Dolton, P., and Lührmann, M. (2018). "Making It Count": Incentives, Student Effort and Performance. Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(2):323–349.

De Paola, M. (2009). Does Teacher Quality Affect Student Performance? Evidence from and Italian University. Bulletin of Economic Research, 61(4):353–377.

Goodman, J., Gurantz, O., and Smith, J. (2018). Take Two! SAT Retaking and College Enrollment Gaps. Working Paper 24945, National Bureau of Economic Research.

Imbens, G. and Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. The Review of Economic Studies, 79(3):933–959.

Joensen, J. S. and Nielsen, H. S. (2009). Is There a Causal Effect of High School Math on Labor Market Outcomes? Journal of Human Resources, 44(1):171–198.

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). Noise: A Flaw in Human Judgment. Little, Brown Spark.

Koch, A., Nafziger, J., and Nielsen, H. S. (2015). Behavioral Economics of Education. Journal of Economic Behavior & Organization, 115:3 – 17.

Lavecchia, A., Liu, H., and Oreopoulos, P. (2016). Chapter 1 - Behavioral Economics of Education: Progress and Possibilities. volume 5 of Handbook of the Economics of Education, pages 1–74. Elsevier.

Levine, P. B. and Zimmerman, D. J. (1995). The Benefit of Additional High-School Math and Science Classes for Young Men and Women. Journal of Business & Economic Statistics, 13(2):137–149.

Machin, S., McNally, S., and Ruiz-Valenzuela, J. (2020). Entry through the narrow door: The costs of just failing high stakes exams. Journal of Public Economics, 190:104224.

McCrary, J. (2008). Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. Journal of Econometrics, 142(2):698–714.

Niederle, M. and Vesterlund, L. (2010). Explaining the Gender Gap in Math Test Scores: The Role of Competition. Journal of Economic Perspectives, 24(2):129–44.

Olsen, A. L. (2013). Leftmost-digit-bias in an Enumerated Public Sector? An Experiment on Citizens' Judgment of Performance Information. Judgment and Decision Making, 8(3):365.

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

Papay, J. P., Murnane, R. J., and Willett, J. B. (2014). High-School Exit Examinations and the Schooling Decisions of Teenagers: Evidence From Regression-Discontinuity Approaches. Journal of Research on Educational Effectiveness, 7(1):1–27.

Papay, J. P., Murnane, R. J., and Willett, J. B. (2015). The Impact of Test-Score Labels on Human-Capital Investment Decisions. Journal of Human Resources, 51(2):357–388.

Papay, J. P., Willett, J. B., and Murnane, R. J. (2011). Extending the Regression-discontinuity Approach to Multiple Assignment Variables. Journal of Econometrics, 161(2):203–207.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. American Economic Review, 94(2):247–252.

Rose, H. and Betts, J. R. (2004). The Effect of High School Courses on Earnings. Review of Economics and Statistics, 86(2):497–513.

Appendix

Figure A.1: $max(T_1, T_2)$ histogram with frequencies by bin

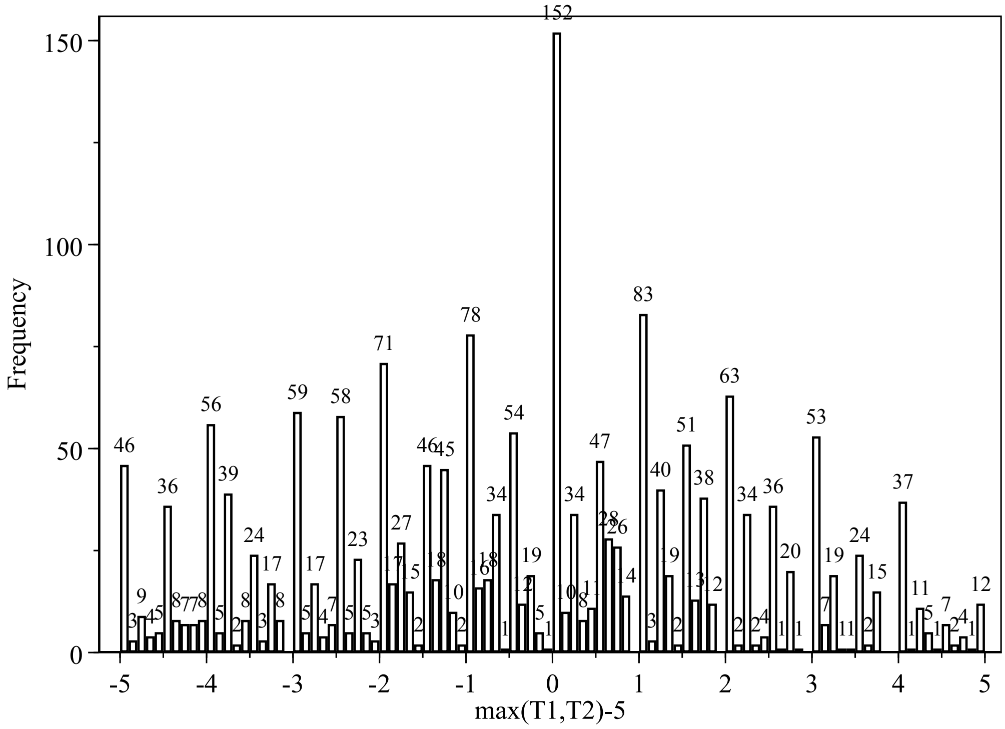
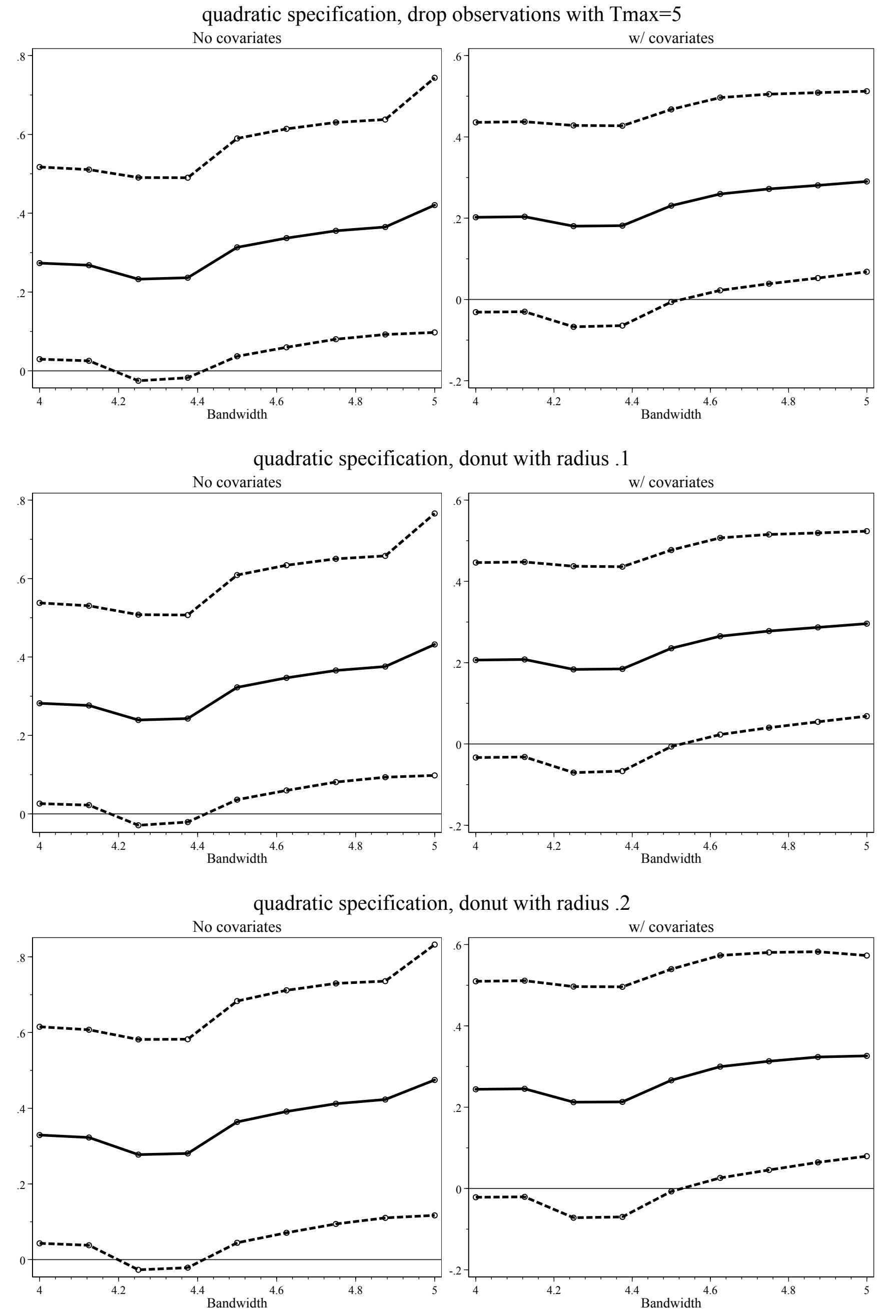


Figure A.2: Sensitivity to bandwidth values of psychological effect of tests on final exam score (quadratic specification)


quadratic specification, drop observations with Tmax=5

quadratic specification, donut with radius .1

quadratic specification, donut with radius .2

Table A.1: Cross-tabulations of number of students by whether their score in $T_1 \geq 5$ and $T_2 \geq 5$

	$4 \leq max(T_1, T_2) \leq 6$		Full sample	
	$T_1 < 5$	$T_1 \geq 5$	$T_1 < 5$	$T_1 \geq 5$
$T_2 < 5$	233	110	750	173
$T_2 \geq 5$	214	120	369	581

Table A.2: Baseline characteristics balance at $max(T_1, T_2)$ cutoff using different bandwidths and a donut with a 0.1 radius

	Female			Foreigner		
$\bar{P}$	-0.034	-0.104*	-0.036	0.022	0.017	0.025
	(0.080)	(0.061)	(0.074)	(0.049)	(0.038)	(0.047)
$\bar{T}^{*}$	-0.005	0.034	-0.039	0.019	0.001	0.025
	(0.055)	(0.029)	(0.053)	(0.032)	(0.018)	(0.037)
$\bar{T}^{*} \times \bar{P}$	0.005	0.004	0.082	-0.058	-0.020	-0.081*
	(0.072)	(0.039)	(0.072)	(0.041)	(0.023)	(0.047)
$(\bar{T}^{*})^2$			-0.011			0.009
			(0.010)			(0.008)
$(\bar{T}^{*})^2 \times \bar{P}$			0.004			0.003
			(0.014)			(0.010)
Constant	0.493***	0.529***	0.468***	0.114***	0.101***	0.112***
	(0.061)	(0.046)	(0.056)	(0.036)	(0.028)	(0.036)
N	872	1,242	1,717	872	1,242	1,717
Bandwidth	1.95	2.92	5.00	1.95	2.92	5.00

	Bus. admin (BA) degree			BA+Law degree		
$\bar{P}$	-0.160**	-0.060	-0.267***	0.028	0.056*	0.171***
	(0.078)	(0.060)	(0.073)	(0.041)	(0.033)	(0.042)
$\bar{T}^{*}$	-0.018	0.003	0.167***	0.027	-0.031*	-0.154***
	(0.053)	(0.029)	(0.051)	(0.025)	(0.018)	(0.034)
$\bar{T}^{*} \times \bar{P}$	0.178***	0.022	-0.008	-0.064*	0.022	0.122***
	(0.069)	(0.038)	(0.069)	(0.034)	(0.023)	(0.042)
$(\bar{T}^{*})^2$			0.032***			-0.019***
			(0.010)			(0.006)
$(\bar{T}^{*})^2 \times \bar{P}$			-0.071***			0.025***
			(0.014)			(0.008)
Constant	0.615***	0.630***	0.753***	0.078***	0.027	-0.073**
	(0.059)	(0.045)	(0.054)	(0.030)	(0.025)	(0.032)
N	872	1,242	1,717	872	1,242	1,717
Bandwidth	1.95	2.92	5.00	1.95	2.92	5.00

Continued on the next page

Continued from the previous page

	Econ. degree			Year 2017-18		
$\bar{P}$	0.139**	0.040	0.117**	0.028	0.087	0.223***
	(0.060)	(0.047)	(0.057)	(0.073)	(0.057)	(0.069)
$\bar{T}$	-0.019	-0.001	-0.049	0.051	-0.027	-0.205***
	(0.038)	(0.021)	(0.034)	(0.048)	(0.027)	(0.048)
$\bar{T}^* \times \bar{P}$	-0.085	-0.005	-0.035	-0.044	0.033	0.197***
	(0.052)	(0.030)	(0.054)	(0.064)	(0.036)	(0.067)
$(\bar{T}^*)^2)$			-0.016**			-0.036***
			(0.006)			(0.009)
$(\bar{T}^*)^2 \times \bar{P}$			0.039***			0.043***
			(0.011)			(0.013)
Constant	0.142***	0.160***	0.132***	0.316***	0.257***	0.125**
	(0.042)	(0.033)	(0.038)	(0.055)	(0.042)	(0.051)
N	872	1,242	1,717	872	1,242	1,717
Bandwidth	1.95	2.92	5.00	1.95	2.92	5.00

	Year 2018-19			Repeater		
$\bar{P}$	0.125	0.057	0.040	0.082	0.074*	0.122**
	(0.077)	(0.058)	(0.070)	(0.056)	(0.043)	(0.052)
$\bar{T}$	-0.063	-0.030	0.014	-0.036	-0.012	-0.077**
	(0.054)	(0.029)	(0.053)	(0.039)	(0.020)	(0.038)
$\bar{T}^* \times \bar{P}$	-0.071	-0.051	-0.143**	0.014	-0.023	0.038
	(0.068)	(0.037)	(0.068)	(0.051)	(0.027)	(0.050)
$(\bar{T}^*)^2)$			0.001			-0.014*
			(0.011)			(0.007)
$(\bar{T}^*)^2 \times \bar{P}$			0.015			0.015
			(0.013)			(0.010)
Constant	0.303***	0.327***	0.371***	0.104**	0.124***	0.078**
	(0.059)	(0.045)	(0.054)	(0.042)	(0.031)	(0.039)
N	872	1,242	1,717	872	1,242	1,717
Bandwidth	1.95	2.92	5.00	1.95	2.92	5.00

Note: The table reports in each panel estimates of regressions with two different dependent variables clearly labeled in the top row of a panel and independent variables are the same as in the RDD polynomial used in all specifications in the paper, with $\bar{T}^* = max(T_1^*, T_2^*)$ and $\bar{P} = (\bar{T}^* \geq 0)$. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.3: Pre-treated joint balance at $max(T_1, T_2)$ cutoff using optimal bandwidth by donut radius

	Female	Foreigner	BA	BA+Law	Econ.	Year 2018	Year 2019	Repeater
					Donut radius 0.1			
$\bar{P}$	-0.02	0.01	-0.10	-0.01	0.30	-0.06	0.09	-0.06
S.e.	0.15	0.09	0.15	0.07	0.12	0.14	0.14	0.11
$\bar{T}^*$	-0.10	0.00	-0.11	0.10	-0.24	0.02	0.18	0.08
S.e.	0.26	0.15	0.25	0.12	0.20	0.24	0.24	0.19
$\bar{T}^* \times \bar{P}$	0.16	0.01	0.23	-0.12	-0.06	0.24	-0.49	0.14
S.e.	0.33	0.20	0.32	0.16	0.25	0.31	0.31	0.24
$(\bar{T}^*)^2$	-0.04	-0.01	-0.04	0.03	-0.10	-0.01	0.12	0.05
S.e.	0.12	0.07	0.12	0.06	0.09	0.11	0.11	0.09
$(\bar{T}^*)^2 \times \bar{P}$	0.01	-0.02	0.07	-0.04	0.20	-0.12	-0.03	-0.17
S.e.	0.16	0.09	0.15	0.08	0.12	0.15	0.15	0.11
Constant	0.45	0.11	0.58	0.11	0.05	0.30	0.40	0.15
S.e.	0.12	0.07	0.12	0.06	0.09	0.11	0.11	0.09
N	872							
Bandwidth	1.95							
P-val $\bar{P} = 0$ jointly	0.29							
					Donut radius 0.2			
$\bar{P}$	0.03	0.06	-0.15	0.05	0.20	0.01	0.14	-0.04
S.e.	0.18	0.10	0.17	0.09	0.13	0.16	0.16	0.13
$\bar{T}^*$	-0.03	-0.03	0.04	0.15	-0.37	0.10	-0.08	0.07
S.e.	0.28	0.17	0.27	0.14	0.21	0.26	0.26	0.20
$\bar{T}^* \times \bar{P}$	-0.07	-0.01	0.01	-0.32	0.40	-0.04	-0.06	0.13
S.e.	0.37	0.22	0.36	0.18	0.27	0.34	0.34	0.27
$(\bar{T}^*)^2$	-0.01	-0.02	0.02	0.05	-0.16	0.02	0.01	0.05
S.e.	0.13	0.08	0.12	0.06	0.09	0.12	0.12	0.09
$(\bar{T}^*)^2 \times \bar{P}$	0.05	0.02	0.03	0.00	0.12	-0.06	0.01	-0.16
S.e.	0.17	0.10	0.16	0.08	0.13	0.16	0.16	0.12
Constant	0.49	0.09	0.66	0.13	-0.02	0.35	0.27	0.15
S.e.	0.14	0.08	0.13	0.07	0.10	0.13	0.12	0.10
N	842							
Bandwidth	1.95							
P-val $\bar{P} = 0$ jointly	0.79							

Note: The table reports estimates of a single regression of a seemingly unrelated regression (SUR) model with as many equations as are the predetermined characteristics in the top row in the table. Each column in the table reports estimates of one of the equations in the SUR model. The independent variables are the same as in the RDD polynomial used in all specifications in the paper, with $\bar{T}^* = max(T_1^*, T_2^*)$, $T_i^* = T_i - 5$ with $i = 1, 2$ and $\bar{P} = (\bar{T}^* \geq 0)$. The bottom row reports the p-value of the null that $\bar{P} = 0$ is jointly zero in all regressions. Robust standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: Effect of a positive signal in at least one midterm (RDD with all observations, i.e. a donut with a 0 radius)

	Final Exam Score standardised by year					
	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{P}$	-0.172	-0.157	-0.086	-0.092	0.158*	0.085
	(0.142)	(0.139)	(0.111)	(0.107)	(0.090)	(0.087)
$\bar{T}^{*}$	0.544***	0.521***	0.424***	0.434***	0.222***	0.314***
	(0.179)	(0.176)	(0.101)	(0.099)	(0.066)	(0.063)
$\bar{T}^{*} \times \bar{P}$	-0.132	-0.075	-0.067	-0.044	-0.096	-0.160*
	(0.204)	(0.199)	(0.117)	(0.114)	(0.087)	(0.083)
$\bar{T}^{*2}$					-0.003	0.012
					(0.012)	(0.011)
$\bar{T}^{*2} \times \bar{P}$					0.048***	0.027
					(0.018)	(0.017)
Constant	0.123	0.348***	0.055	0.266**	-0.115	0.122
	(0.132)	(0.133)	(0.099)	(0.104)	(0.077)	(0.081)
N	681	681	903	903	1,880	1,880
Controls	0	1	0	1	0	1
Bandwidth	1.06	1.06	1.60	1.60	5.00	5.00

Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.5: Effect of a positive signal in at least one midterm (donut with a 0.1 radius)

	Final Exam Score standardised by year					
	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{P}$	0.233*	0.163	0.253**	0.234**	0.432**	0.296**
	(0.119)	(0.124)	(0.097)	(0.089)	(0.170)	(0.116)
$\bar{T}^{*}$	0.347***	0.364***	0.290***	0.290***	0.219	0.316***
	(0.084)	(0.082)	(0.053)	(0.050)	(0.168)	(0.095)
$\bar{T}^{*} \times \bar{P}$	-0.259**	-0.186*	-0.172**	-0.142**	-0.331*	-0.350**
	(0.120)	(0.110)	(0.077)	(0.069)	(0.195)	(0.135)
$\bar{T}^{*2}$					-0.003	0.012
					(0.024)	(0.016)
$\bar{T}^{*2} \times \bar{P}$					0.091***	0.060**
					(0.033)	(0.025)
Constant	0.003	0.181	-0.050	0.145	-0.119	0.116
	(0.119)	(0.134)	(0.106)	(0.139)	(0.167)	(0.147)
N	827	827	1,182	1,182	1,723	1,723
Controls	0	1	0	0	0	1
Bandwidth	1.78	1.78	2.66	2.66	5.00	5.00

Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.