How does relative performance feedback affect beliefs and academic decisions?

Catalina Franco *

Most recent version here

January, 2019

I report on a field and lab-in-the-field experiments examining how students' beliefs, academic investments, performance, and the decision to take a college entrance exam change when students of different ability levels learn about their relative performance. In the context of a 3-month preparation course for a high-stakes college entrance exam, I elicit beliefs from all students and randomly assign them either to receive or not to receive feedback about their quartile in the score distribution in weekly practice tests. Combining elicited beliefs and administrative data, I present three main findings. First, relative performance feedback affects almost exclusively low-performing students, who become discouraged by this intervention. Compared to low-performing students who do not get relative feedback, they take fewer practice tests, get lower scores, and are less likely to take the college entrance exam they are preparing for. Second, low-performing female and male students respond in different ways, with males reducing investments and females keeping effort up but deciding to not take the college entrance exam. Third, even though low performers change behavior, they do not update beliefs in the incentivized task, suggesting potential discrepancies between elicited and true beliefs. Overall, my results shed light on the potential discouragement effects of informational interventions on students with low academic performance.

JEL codes: C91, C93, D83, D91, I21, I25, J24

*Economics Department and FAIR, Norwegian School of Economics, catalina.franco@nhh.no. I thank Tanya Rosenblat, Martha Bailey, Dean Yang and Hoyt Bleakley for their support and encouragement over the course of this project. At various stages I received helpful comments from Diego Aycinena, Mariana Blanco, John Bound, Charlie Brown, Valentina Duque, Enrique Fatas, Leonardo Garzon, Amelia Hawkins, Craig McIntosh, Meera Mahadevan, Cesar Mantilla, Markus Mobius, Ben Meiselman, Nicolas Morales, Yesim Orhun, Mounu Prem, Paul Rodriguez, Kjell Salvanes, Santiago Sautua, Juan Fernando Vargas, Gonzalo Vazquez-Bare, and many others including audiences at U. Michigan, U. del Rosario, U. de Antioquia, U. de los Andes, Advances with Field Experiments, ESA Antigua, SEEDEC at UC Berkeley, SOFI, and CMI. Laura Cortés and Luisa Oyuela provided excellent research assistance. I specially thank Paulo Rico for helping coordinate the implementation of the project at the institute. Funding for this project was generously provided by the Michigan Institute for Teaching and Learning and Economics (MITRE) and the Rackham Graduate Student Research Grant. I acknowledge receipt of funding from Fulbright Colombia during part of my PhD.

1 Introduction

Beliefs about our academic capabilities are central to our educational choices and investments (Altonji, 1993; Manski, 2004). Holding incorrect beliefs may lead students to make decisions that do not correspond with their ability and skills, with possibly high short- and long-term costs. Theoretically, accurate feedback can help correct beliefs and help students make better decisions. However, feedback can have motivational effects, in addition to its information value. Hence, the direction of student responses to an intervention providing feedback is unclear and depends on the content of the information and how much the students discount it (Brade et al., 2018). Two separate strands of the literature have studied the effect of providing *relative* performance feedback on students' grades (Azmat & Iriberri, 2010, 2016; Bandiera et al., 2015; Murphy & Weinhardt, 2018; Megalokonomou & Goulas, n.d.; Azmat et al., 2019), and the effect of eliciting and correcting students' beliefs about *absolute* performance on academic choices (Bobba & Frisancho, 2016; Gonzalez, 2017). However, there is very little work understanding the connection between these two bodies of work. I aim to close this gap by showing how providing feedback about relative performance, which is relevant for many educational settings where competition determines student success, affects students' subsequent beliefs and academic decisions.

This paper reports results from a field and lab-in-the field experiments in Colombia designed to understand how students incorporate relative performance feedback into their beliefs and decisions. I focus on a high-stakes context where relative performance beliefs are particularly consequential - college entrance exams (CEEs) - and aim to answer the following question: How do beliefs, academic investments, performance, and choices change when students of different ability levels learn about their performance relative to their peers? Ability levels matter because the information content of relative performance feedback varies and may trigger different reactions among academically stronger and weaker students. To answer my research question, I match the data collected in the experiments with administrative data of 440 students preparing for a CEE at a test preparation institute[1] to find impacts on the following outcomes: (i) belief updating, (ii) academic investments, (iii) performance, and (iv) academic decisions.

To shed light on the average effect of receiving feedback for students of different ability levels I use two experimental approaches. The first approach relies on randomly assigning students taking the test preparation course to receive or not to receive feedback about their performance in weekly practice tests relative to other students preparing for the same exam

[1]This institute offers preparation courses for standardized exams and is similar to institutions in the U.S. preparing students for the SAT, GRE, etc.

at the institute. I leverage the institute's practice test performance report to deliver this treatment. Specifically, the treatment provides students information on which quartile they lie in based on their math and reading practice test scores. By comparison with their reported beliefs, this information also allows students to know how precise they are in predicting their relative performance. Students assigned to the treatment group receive feedback over the course of the whole experiment while control students only see their absolute scores. My analysis of the treatment effects relies on using the quartile of performance in the initial practice test as a proxy for ability levels.[2]

To study belief updating, my second approach embeds a lab-in-the-field experiment into the main experiment over the course of eight weeks. I adapt an incentivized mechanism in the spirit of Mobius et al. (2014) to elicit beliefs regarding relative performance in practice tests from treatment and control students. Specifically, I ask students to assign probabilities of being in each of the four quartiles of the math and reading practice-test score distributions after each practice test.[3] I re-elicit their beliefs regarding their performance on the same practice test after they learn their absolute scores and treated students receive a "signal" indicating whether their scores are above or below the median.

This paper produces three basic results. First, low-performing students become discouraged by relative performance feedback. Compared to control students in the bottom quartile of initial practice test performance, students who receive feedback are less likely invest in preparing for the entrance exam. For example, their participation in weekly practice tests falls by 5.2 percentage points, their reported study hours fall by 30 percent, and their performance in the practice tests they take is worse than control students who were in the same initial quartile of performance but did not receive feedback. Moreover, learning that they are low-performers dissuades some students from taking the CEE they are preparing for. Low-performers who receive feedback are 11 percentage points less likely to take the exam than similar students in the control group. This difference lasts through the next admission cycle, ruling out explanations related to taking additional time to prepare after realizing that they are at the bottom of the distribution.[4]

An alternative explanation to discouragement among low-performing students is that those who reduce investments and decide not to take the CEE were less motivated from the beginning of the course. To find out whether this is the case, I analyze the first practice

[2]Because the preparation course is relatively short, it is unlikely to see big changes in performance relative to the quartiles in the initial practice test. In fact, in my data there is a high persistence of the students' initial quartile in subsequent practice tests' quartile.

[3]Quartiles are computed using scores of all students enrolled in the same preparation course even if they are not participating in the study.

[4]I do not find effects on students at other ability levels or on the actual performance in the exam among those who decide to take it.

test performance of students who later on decide to take and to not take the CEE. I find that there is no difference at baseline between them and, if anything, those who end up not taking the CEE performed slightly better in the initial practice test. Taken together, these results lend credence to the discouragement hypothesis and suggest that students who decide to quit were not performing worse since the beginning but rather become discouraged over the course of the experiment.

Second, only high-performing students seem to update beliefs in the incentivized elicitation task. To study how students incorporate feedback into their assessments of performance in subsequent practice tests, I pool beliefs across all practice tests and define categorical variables to classify them as having a correct belief or whether they over- or under-place themselves. I find that, across all rounds, about 30 percent of students below the top quartile of initial performance have a correct belief about their quartile in the most recent practice test regardless of their treatment assignment. In contrast, in the top quartile of initial performance, 53 percent of treated students can correctly assess their quartile in the most recent practice test over a base of 42 percent in the control group. This suggests that top performers are more able to incorporate the information delivered in the intervention, but it remains unclear whether it is because of their higher academic ability or because they derive utility from receiving positive feedback about it (Köszegi, 2006).

The finding that students below the top quartile of performance do not update beliefs is at odds with low-performing students' change in behavior as a result of receiving relative performance feedback. If they were not understanding the content of the information provided, there is no reason to change their effort or their academic decisions. There may be several reasons why there is a mistmach between what low-performing students report in the belief elicitation task and the beliefs revealed by their actions. One piece of evidence that is consistent with the discouragement hypothesis discussed above is that low-performing students in the treatment group are more likely to check the feedback only once out of eight times it was provided. This suggests that students are avoiding information that is potentially demotivating (Golman et al., 2017).

Third, both men and women become discouraged when receiving feedback but they adjust their behavior in different ways. Differential effects by gender are expected in this kind of intervention given the substantial gender differences found in competitiveness (Niederle & Vesterlund, 2007), college major choices (Buser et al., 2014; Reuben et al., 2015; Buser et al., 2017) and reactions to losing (Alan et al., 2016; Buser & Yuan, 2016). Women do not give up their preparation for the exam when they receive feedback, that is, they do not adjust their investments. However, women in the bottom and top quartiles of initial performance are less likely to take the CEE in the current and the next admission cycles

than women in the control group.[5] Men in the bottom quartile, on the other hand, study less and take fewer practice tests than men in the control group, so they primarily adjust their investments and, to a lesser extent, the decision to take the CEE. Speaking to the literature on gender differences in high-stakes test performance (Ors et al., 2013; Cai et al., 2016), despite similar performance in practice tests, men outperform women in the entrance exam and are substantially more likely to gain admission.

To summarize, this paper makes several contributions to existing research on the effects of correcting beliefs and providing relative performance feedback in education and experimental economics. First, a recent literature studies how correcting beliefs about academic performance improves the decision making of students (Bobba & Frisancho, 2016; Gonzalez, 2017). This literature focuses on students' beliefs about absolute performance in a mock exam, and elicits beliefs and provides feedback only once before observing performance in the real test. What remains to be understood is what drives the changes in decision making that can be observed from the mock exam to the real test. By eliciting beliefs and providing feedback more than once and in a frequent manner, I provide evidence of the drivers of students' decisions such as adjustments in effort and beliefs *before* the real test takes place, which previous papers in the literature have been unable to observe. Moreover, I analyze relative performance beliefs and feedback that are more relevant for many education contexts in which competition for slots is determined based on relative, rather than absolute, performance.

Second, this paper adds to the research finding inconclusive results of providing relative performance feedback on grades. Most work on this literature provides information about the mean or other aggregate statistic of relative performance, and finds positive effects of providing information (Azmat & Iriberri, 2010, 2016; Bandiera et al., 2015). Murphy and Weinhardt (2018) and Megalokonomou and Goulas (n.d.) analyze natural experiments in which students get to know their relative position within a class or in a national exam. Both papers find negative effects on subsequent academic performance. With the exception of Azmat et al. (2019), who also find negative effects of relative performance feedback, the papers in this literature do not consider the role of beliefs in explaining why grades change as a result of providing relative performance feedback. One on my contributions to this literature is eliciting students' beliefs from the same students who take part in the intervention so I can simultaneously study the effect of relative performance feedback on beliefs, and academic performance and decisions. Importantly, my design allows to study outcomes beyond grades

[5]Students in the top quartile of initial performance are also less likely to take the CEE. Closely examining the reason for this, I find that this effect is driven by women who performed substantially worse in the last practice test relative to their previous performances.

that can help clarify why studies in this literature find effects on grades. For example, my finding of lower academic investments among low-performing students could be thought of as a potential mechanism triggering the lower academic performance documented by some papers in this literature. In addition, my findings shed light on how students of different ability levels may be affected by an intervention providing relative performance feedback.

Third, this paper contributes to the extensive literature in experimental economics studying the effect of feedback provision (e.g. Gill et al., 2016; Azmat & Iriberri, 2010, 2016) and information processing (e.g. Eil & Rao, 2011; Grossman & Owens, 2012; Mobius et al., 2014; Ertac, 2011) in the lab. My most important contribution to these two strands of literature is to show that there is a mismatch between the beliefs reported by the students and the beliefs that can be inferred from their observed behavior. This finding suggests that elicitation tasks outside of the lab may not be capturing the same underlying behavior as in the lab. A number of features may explain this difference. The stakes in the lab are much lower relative to what is at stake for these students when they take practice tests to prepare for the CEE. Also, outside of the lab, there may be other considerations that can be more important to individuals than correctly guessing their relative performance to receive a monetary prize. For example, protecting the ego from receiving disappointing information may be much more attractive than receiving a monetary prize for correctly guessing a low performance.

Lastly, my findings shed light on ways policy makers can improve students' outcomes. Many schools around the world already provide information of the ranking of students within their class or school. This paper shows that students at the bottom of the distribution can be especially discouraged by such information. On the other hand, an implication of my results could be that students may save time and effort by adjusting their investments and possibly abandoning the idea of going for a selective college admission process. Therefore, policy makers who care about the potential psychological effects associated with discouragement face a tradeoff. They can avoid informing students with the possible consequences that they keep blindly investing in taking an exam that is very unlikely that they will pass. Or they can provide information with the risk that some of them will be discouraged from even trying. A full accounting of these considerations and a thorough analysis of the specific context should be incorporated when discussing alternatives for providing feedback in an education setting.

This paper proceeds as follows. Section 2 describes the context and experimental design. Section 3 presents summary statistics and balance of characteristics. Section 4 presents the main findings. Section 5 presents heterogeneous effects by gender, and Section 5 concludes.

2 Context and experimental design: Eliciting students' beliefs and observing decision making in the field

To study decision making and beliefs, I conduct a field experiment with students preparing to take a high-stakes college entrance exam in Colombia. Over the course of 10 weeks, I elicit beliefs from all study participants in a test preparation center about their perception of relative performance in practice tests. To test how beliefs and academic investments and decisions change when students receive feedback regarding their relative performance, I divide the sample into treatment and control groups. In 8 of the 10 weeks, I provide feedback, to the treatment group only, about the exact quartile the students' scores fall into.

2.1 College entrance exams in Colombia

In many developing countries, admissions to highly-selective public universities are based on a single factor: the score in a college entrance exam. Because public universities are usually high-quality institutions and the tuition they charge is free or highly subsidized, earning a slot is extraordinarily difficult. For many students, especially those coming from low socio-economic backgrounds, such schools may be their only opportunity to earn a college degree.

In Colombia, students graduating from high-school who are willing to enroll at a public university are, in general, required to take a university-specific college entrance exam. Every university designs its own exam, grades it, and admits students according to a pre-established mechanism for slot assignment. The number of slots by major is fixed and announced before the exam takes place. Universities administer the entrance exam once per semester, that is, there are two rounds of admissions per calendar year. For admission, universities require the entrance exam score but no letters of recommendation, high school GPA, or scores in the national standardized test.[6]

Admissions at universities using college entrance exams is highly competitive. The students of my sample are preparing for the entrance exam at Universidad de Antioquia. This is a regional university considered to be the second best public university in Colombia. It offers about 100 different majors in several campuses, the most selective of which is the Medellín campus. Every calendar year, the entrance exam takes place in April and September. Students who take the exam in April start college in August, and those who take it in September start the following February. The exam contains 80 questions divided evenly

[6]The national standardized exam is taken by all high-school graduating students and is part of the admission criteria at private universities. Most public universities will ask new admitted students to report their result in the national exam but it is not considered for admission decisions. In general, this exam tends to be less difficult that public universities CEEs.

between math and reading, and students have three hours to solve all questions. The scores of the two sections are averaged and the global score is then standardized to obtain a score between 0 and 100. To be eligible to compete for a slot, an applicant needs a minimum standardized score of 53 points for the Medellin campus and 50 for other campuses.

Gaining admission at Universidad de Antioquia is a combination of the overall score in the entrance exam and the majors students declare. Even though everyone takes the same exam in a given admission cycle, the competition every student faces is different because it depends on which one or two majors they declare when they register for the exam. In this sense, despite having high scores, if students choose a very competitive major, it is likely that they do not gain admission. Importantly for my design to study academic choices, students have to choose up to two college major options *before* they take the exam, so their actual performance does not inform them on which majors to select. Overall, admission rates are around 10 percent, but this varies substantially by major. At the Medellín campus in the April, 2018 admissions cycle, 21 of the 83 majors offered had admission rates below 5 percent (see Appendix Figure C.17). The five majors with lowest admission rates were: Surgical instrument processing (1.9 percent), nursing (2.1 percent), psychology (2.1 percent), medicine (2.2 percent), and nutrition and dietetics (2.3 percent).[7]

The slot assignment mechanism takes two pieces on information into account: the overall score in the entrance exam and the major(s) selected by the applicant. The university allows applicants to select up to two academic programs to which they would like to be admitted. This choice happens before the student takes the exam and knowing very little about potential competitors. To give an example of how the slot assignment mechanism works, in the semester in which this study takes place (first semester of 2018), there were 139 slots for medicine and about 6,300 students who declared this major as a first or second choice. After grading the exam, the university ranks the scores of all students who declare medicine as a first choice and starts assigning slots going down the list until filling all of them. For any remaining slots, the university selects applicants among those who selected medicine as a second option. It is very rare to be assigned a slot for the major selected as a second choice, especially in the most competitive majors.

Given the competitiveness of this exam, there are many institutes offering courses to help students prepare. The institutes mainly offer in-person courses lasting from 1 to 3 months on average. An online search of preparation courses for the Universidad de Antioquia entrance exam results in at least 10 of such institutes in the city of Medellín. Assuming an average of 1,000 students enrolling in these courses per admission cycle, at least 10,000 applicants are going through one of these courses every semester. The number of applicants for the

[7]Information on programs offered, cutoff scores and number of applicants can be found here.

April exam is around 35,000 while for the September exam it is 50,000. So, not less than 20 percent of applicants are obtaining some sort of exam-specific preparation every semester.

To conduct this study, I partnered with one of the most renowned test preparation institutes in Colombia. This choice of sample has the advantage that it is known that all students at the institute are willing to take the exam, which is not straightforward when sampling from high schools because some students may not be interested in applying to college or to public universities. The institute allowed me to contact all students enrolled in the preparation course taking place from January to April 2018. Students enrolled in this course attend 4 three-hour classes per week covering the two exam subjects. Besides classes, every Monday, students take a full-length practice test that is supposed to simulate the actual exam. There were 11 practice tests in total administered either in-person or online. Besides the lectures and practice tests, students obtain a workbook with practice questions, online materials and a performance report after each practice test. The cost of this course is around COP 1,000,000 (US$330), which is equivalent to 1.5 times the monthly minimum wage with the exchange rate at the time of the study.

For the intervention, the test preparation institute allowed me to survey the students, modify the performance reports, and provided administrative data. Details on the exact modifications to the performance reports are in the next subsection.

2.2 Experimental design and timeline

The experimental design consists of two parts: (i) I collect beliefs about relative performance in practice tests from all participants in a weekly lab-in-the-field experiment, and (ii) I provide relative performance feedback to a randomly selected sample of students preparing to take a college entrance exam at a test preparation institute in Medellin, Colombia. After each practice test students take as part of the preparation course, I elicit probabilities of falling in each of the four quartiles of the math and reading practice-test score distributions. I modify the institute's results report to provide relative performance feedback to treated students.

While absolute scores in practice tests are important to assess improvement, admission at the university the students in the sample are applying to is determined by the exam performance relative to other test takers declaring the same major. Hence, having access to relative comparisons can provide useful information to students beyond the absolute scores in practice tests that the institute already provides. Ideally, students would compare themselves to all other students who will declare the same major in the semester they will be taking the exam. Unfortunately, only the university knows who will be taking the exam and which

majors they declare when students register for the exam. In addition, the university only knows this information about one month before the exam is administered.

One way to overcome the impossibility of providing performance feedback relative to the actual pool of applicants who will be competing with students in my sample is to use practice test scores from students' within the test preparation institute. This is a relevant sample because it contains students who will be taking the exam, and despite being positively selected, it is not too different from other students making up the potential pool of applicants. In Figures 5 and 6, I show administrative data of a matched sample containing students at the institute and all students in Colombia and in the city of Medellín. The figures show the distribution of scores in the national standardized exam for the sample and the whole population. The two distributions overlap substantially although, on average, students at the institute perform better in math and reading than students in the whole population.

I provide feedback to students in the treatment group based on how they perform in the math and reading sections of every practice test relative to the rest of the students taking the same practice test at the institute (about 1,200 students). Hence, participating treated students receive feedback concening the rest of the students at the institute, independent of whether those students to whom they are being compared are participating or not in the research study.

To deliver the relative performance feedback, I separately compute quartiles of the math and reading practice test score distributions. The quartiles are calculated based on the scores of all students taking the same weekly practice test. To circumvent the problem of ties in practice test scores that may lead to quartiles of unequal sizes, students who are in the limit between two quartiles are randomly assigned to one or the other. The decision to provide relative performance feedback in terms of quartiles and not other finer measure is related to the belief elicitation task, which would be much more time-consuming and error-prone if students had to assign probabilities to deciles or ventiles of the score distribution.

Sample selection: The study worked with one of the most renowned test preparation centers in Colombia. All students enrolled in the first cohort of the course offered between January and April 2018 received a visit during the first week of classes.[8] In that visit, they were told about the study and signed a consent form indicating whether they wanted to participate. To promote student participation, the consent form explained that there would be raffles of cash prizes every week among students who answered the surveys. Besides explaining the study in general terms and collecting information about willingness to participate, the consent form included a question about having taken the college entrance exam in the

[8]Students who enroll after classes have started make part of other cohorts. Because they do not take practice tests at the same time as students in cohort one, they were not recruited for the study.

past, which is one of the stratification variables in the randomization procedure.

Most students expresesed interest in participating during the visit to their classroom. However, when the study started, many did not engage. In Section 3, I show that the actual sample size (440 students) is much lower than the sample who consented participation, and that it is unlikely that the treatment status affected their decision to actively participate because they did not know their treatment assignment.

Randomization: I randomly assigned half of the students who consented participation to a treatment group that received weekly relative performance feedback in the two subjects covered by the exam. To reduce sample variability and to conduct heterogeneity analysis, the randomization was stratified based on gender, whether they had taken the exam in the past, quartile in the initial practice test, and type of course they were enrolled in (morning, afternoon / evening, weekends, pre-medicine, joint preparation for two entrance exams at diffferent universities).

The randomization was performed at the individual level because the performance reports are customized for every student and to increase power. Concerns with spillover effects may arise because students are organized in classrooms at the beginning of the course. However, several features of this setting minimize the role for spillover effects. First, the performance report was designed to be delivered online and students were instructed to check it on a computer rather than on a phone to be able to complete the belief elicitation task and play the game to learn how much they could earn if selected in the weekly raffle. Because the institute does not have a computer lab that students can use, it is likely that students checked the report at home. Second, students are assigned to classrooms randomly so it is unlikely that they know each other from before the preparation course. In addition, they may make friends over time but the course is very short (3 months) so they probably do not spend too much together after class. Third, to infer one's quartile based on another students' report is difficult because the two students would basically need to have the same score.

Lab-in-the-field belief elicitation: The belief elicitation task is based on an incentive compatible mechanism developed by (Mobius et al., 2014). I elicit the probabilities of being in each quartile of the math and reading score distributions while (Mobius et al., 2014) elicit the probabilities of being above and below the median in an IQ quiz administered in the lab. Berlin and Dargnies (2016) adapts this mechanism to elicit beliefs about quartiles and I further modify it to make it easily understandable for students in my sample.[9] For simplicity and to encourage understanding of the task, I gave 12 imaginary tokens for math and 12 for reading after each practice test so students can play the "quartile game". The framing of

[9]The task elicits probabilities but I avoided asking for probabilities directly as some students may not be familiar with the concept.

the task instructed students to play the tokens by betting on the quartile in which the they thought their score would be in. They could distribute the tokens as they wished across the four quartiles but were told that their probabilities of winning a cash prize increased if they correctly guessed the quartile. Everyone received training about what a quartile was and the quartiles were defined as groups containing 25 percent of the students according to their ordered score in each exam subject. In this sense, the first group (quartile 1) contains the 25 percent of students with the highest scores, and so on.

Beliefs were elicited twice after each weekly practice test to obtain priors and posteriors. The first time students report their beliefs is right after the practice test to elicit priors based only on how they felt in the practice test. The second time students report their beliefs is when they check their experimental performance report. In the case of control students, they see their absolute scores and then report their beliefs. Treated students report their posteriors after seeing their absolute scores plus a "signal" indicating that their score was above or below the median. The reason to include this step is to study their belief updating regarding the same practice test relative to the Bayesian benchmark, i.e. to know how their posterior compares to that of a Bayseian agent with the same priors as they stated in the first belief elicitation, immediately after the practice test.

The belief elicitation task consists in incentivizing truth telling by providing weekly cash prizes to students whose guess was correct (assigned most tokens to the quartile in which their score was) and who are selected in a raffle determining who receives the prize. For each student completing the task, one of the two belief elicitations during a given week was chosen at random to be entered in the raffle (see experimental instructions). At the end of the online experimental performance report, students were guided through instructions to throw a 12-sided dice that would determine whether they receive zero or a positive amount of cash. Let y be the random draw from the dice and x the number of tokens assigned to the quartile to which the student's score belongs to. The specific procedure to determine prizes was as follows:

1. If $y \leq x$ the student wins COP 20,000 (US$7).

2. If $y > x$, the student wins COP 20,000 with $y\%$ probability. To implement this, there is a second draw to obtain a new number z. The student wins if $z \leq y$.

According to this mechanism, students had incentives to put more tokens to the quartile in which they think they are so that they maximize the probability of winning. They were also incentivized to correctly guess their number of correct answers in math and reading. If this guess was correct, the student would receive COP 5,000 for each practice test subject.

In a single round, a student whose guesses were all correct and was selected in the raffle could earn a total of COP 50,000 (almost US$ 17), which is a substantial amount for students of their age and socioeconomic status.

Relative performance feedback: Once the practice tests were graded, the institute posted a performance report in an online platform. The standard performance report provided by the institute contains the number of correct and incorrect questions in math and reading, and a global score from 0 to 100 that is meant to resemble what they would score in the actual entrance exam.[10] An example of this report is in Figure 1.

I add to this report a series of additional steps to complete the belief elicitation task and to provide relative performance feedback to students in the treatment group. I call this report the experimental performance report. In this report, control students obtain their absolute scores and report their updated beliefs after seeing this information. At the end of the report, they play the "quartile game." In general, control students could not know their quartile in the distribution through playing the game unless they assign all tokens to the quartile in which they are in so that the probability of getting a dice draw below their allocation is 1.[11]

The experimental performance report for the treatment group contains additional information. After seeing their absolute scores, treated students see a screen with two "signals," one for math and one for reading, stating whether their score in these subjects was above or below the median. After receiving the signals, the students work on the belief elicitation task. Once treated students complete the belief task, they are directed to a feedback report showing their beliefs stated in the first belief elicitation, and the quartiles in which their scores in math and reading belong to (see Figure 2). Finally, they play the "quartile game."

The key difference between the experimental report for treated and control students is that treated students get more information which compares their performance relative to the rest of the students taking the same practice test as them. The reports were short and user friendly (see experimental instructions in Appendix B). On average, students spent 40 seconds checking the experimental performance report (see Panel C of Table 2). Students could check their absolute scores without having access to the experimental performance report. This is because the institute did not want to prevent students not willing to participate or not willing to continue participating in the study from checking their scores. Because students could only access the absolute scores and not get access to the experimental report, I only include in the analysis students who checked the experimental report at least once.

[10] These scores tend to be lower and the distribution is more compressed than the scores students from the institute obtain in the actual exam.

[11] In practice, very few students allocate all tokens to a single quartile. Across all rounds, students assign all 12 tokens about 7 percent of the time for math and reading.

Section 3 below provides more details on this.

Timeline: The field experiment timeline is in Figure 3, and the timeline for the lab-in-the-field experiment used to elicit prior and posterior beliefs is in Figure 4.

2.3 Data and outcomes

The analysis uses four sources of data. Primary sources are the weekly belief elicitation surveys. Secondary sources include test preparation institute records, as well as administrative data from Colombia's testing agency and university admissions records. The main outcomes I study are whether students take the entrance exam and practice tests, performance in both, majors declared, self-reported study time, and how correct their relative performance beliefs are.

Primary data sources and outcomes: Using individual surveys, I elicited prior beliefs across 10 rounds after each practice test except after the initial one. From the experimental performance report, I obtain posterior beliefs across 8 rounds.

After every practice test, I administered a survey with questions about students' expected absolute and relative performance, hours of study in the previous week, and the perceived difficulty of the test. I provided paper or online surveys depending on the type of practice test (in-person or online). Overall, there were 10 rounds of prior belief elicitation, excluding the first practice test. I collected posterior beliefs (online only) using the experimental performance report.[12] The main outcomes from beliefs elicitation surveys include whether students are correct, underplace, overplace, have a flat prior, or have inconsistent beliefs.[13]

I administered additional surveys to collect data at midline and follow-up after the intervention. However, student participation during after class hours was low. I do not use the data collected from these surveys extensively but, nevertheless, I briefly describe what these surveys collected. The midline survey was administered online between 3 weeks and one month after the course started. The main idea was to collect information about intended majors and predicted scores. The 6-month follow-up survey inquired students about their main activity last week (studying, working, etc.), what program and institution they were attending if they were studying, whether they were beneficiaries of the government scholarship program, and a few questions related to happiness and life satisfaction. Of the students in my analytical sample, I was able to reach about 75 percent in the 6-month follow up survey.

Secondary data sources and outcomes: Participants' data from the experiment were

[12] It was only possible to collect posteriors in 8 of the 10 rounds because of technical issues with the online platform in the first two weeks of the intervention.

[13] I create indicator variables for each of these categories. See table A for definitions of these variables.

matched to administrative records from the test preparation institute, university admission statistics, and the national standardized exam administered by the Colombian agency for higher education (ICFES).

The institute provided information on practice test scores, classroom assignment, demographic and economic characteristics, contact information, type of course they enrolled in, and names of instructors. I use whether the students take the practice test and their scores in math and reading as measures of academic investments. The rest of the characteristics are part of the check for randomization balance.

From the university administrative data I obtain college major choices, overall scores and scores by section in the entrance exam, whether the applicant was admitted and to which program, whether the applicant registered for the next admission cycle and for which program. In addition, public statistics published in the university website contain admission cutoff scores for each major.

Finally, using administrative data from students in the whole country collected by ICFES, I can see how students' performance in the national standardized test compares to that of the rest of test takers in Medellin and Colombia. I use these data to analyze what kind of selection in terms of scores in the national standardized test there is at the test preparation institute relative to other high-school graduates in the country.

3 Summary statistics and balance

Students who enroll at the institute are predominantly women, low-middle income, academically better than average students in their city and Colombia, and have already taken the exam in the past.

3.1 Sample characteristics and balance

Table 1 presents means of baseline characteristics of students who checked at least one of the experimental performance reports along with p-values of individual and joint tests of differences between treatment and control. Because the basis for the empirical analysis will be the quartiles in the initial practice test, Table C.2 in Appendix C shows that characteristics are also balanced within the quartiles.

On average, students in my sample are 60 percent female, almost 18 years old, and 98 percent single. About 80 percent of them have taken the entrance exam in the past, which suggests that students who enroll in this type of institute have already tried and failed gaining admission. Based on their residential strata, a measure of socio-economic status, and their

household poverty index these students are in low and middle-low income households.

From the data collected by the institute, students obtain a score of around 37 points out of 100 possible in the initial practice test. Their number of correct answers (out of 40) in math is substantially lower than in reading. Around 42 percent of the sample is enrolled in the morning courses, 36 percent in the afternoon / evening courses, 2 percent in the weekend courses, 4 percent in the course preparing them simultaneously to two entrance exams at different universities, and 15 percent are in a pre-medicine course.

3.2 Sampling frame and sample size

Table 2 shows how many students consented participation, how many were randomly assigned, and how many checked the experimental performance report at least once. Overall, 1,024 students consented participation when they were approached in the classroom. Of these students, 512 were assigned to the treatment and control groups, respectively. However, 56 percent of those who consented participation never checked the performance reports. The main reason why many students did not participate despite consenting participation was that it was difficult to interact with them after they left the institute. In general, students in the top quartile of initial performance were more likely to check at least one experimental performance report (57 percent), while between 35 to 40 percent of students in other quartiles did so. This does not mean that students did not know their absolute scores. As mentioned before, the institute did not want to prevent students from receiving absolute scores so they could get this information without having to look at the experimental report.

All analyses are performed using the sample of students who checked at least one of the experimental performance reports. Students who did not check the reports did not fill out the belief elicitation tasks and seem to be more disengaged in general, with lower participation rates in the college entrance exam (6 pp lower than that of students who checked at least one performance report). Among those who checked at least once, some did not check all 8 reports delivered in the intervention (Panel C of Table 2). Again, this is partly due to the fact that it proved extremely difficult to engage them in the study when they were not present at the institute. It could also be because some students intentionally avoided information that could discourage them. On average, treated and control students checked 2.5 experimental performance reports and spent about 40 seconds going through all the screens of the report. Basic statistic on report checking by quartile of initial performance are in Table 2.

While the smaller sample size reduces power to detect effects, it does not seem to threat the internal validity of the study because none of the attrition sources is correlated with the treatment assignment. In fact, it was not possible for the students to know which group they

were assigned to before they checked the report for the first time. Even after checking the report they may not know that other students may receive different pieces of information in the reports as explained in Subsection 2.2.

3.3 How different is the sample from average students?

Linking participant IDs with administrative data from the national agency in charge of testing all high-school graduates in Colombia (ICFES) shows that students in my sample are positively selected.

Because there is very little information on who enrolls in this type of college preparation courses, a natural question is how representative of the general student population are the students in the sample. Figures 5 and 6 show the distributions of math and reading scores in the national standardized test for students in the sample and all students in Colombia and the city of Medellín. In both cases, the distributions of scores of students in my sample is shifted to the right of the scores of all other high-school graduates, although there is substantial overlap between the two. The support of the distributions of Colombia high-school graduates goes from zero to 100 while the support of students in my sample goes from around 20 to 80. That is, the average student in my sample scores higher than the average student in Colombia and the variances in the scores are lower.

One implication of the positive selection is that by providing feedback about relative performance, students at the bottom of the distribution in the preparation institute may get the misleading message that they are not good while in fact they are but they are being compared to students who are at the top of the applicant pool distribution. However, this is not the case as there is good overlap between the two distributions. As I explain in the results section, students in the bottom two quartiles have very low admission rates, suggesting that they are not very good performers when comparing them to the actual applicant pool.

4 Findings: Effect of relative performance feedback on academic decisions and beliefs

This section shows the effects of the relative performance feedback intervention on students' beliefs and academic outputs and inputs. It also discusses whether students' actions are consistent with what students reported in the incentivized belief task. As has been found in previous work in the lab (Gill et al., 2016), I observe that top and bottom performers are the most responsive to feedback.[14] Bottom performers receiving feedback are 5 percentage

[14]However, in the lab, both top and bottom subjects increase effort but this is not the case in my setting.

points less likely to show up to practice tests and 11 percentage points less likely to take the entrance exam. This behavior is consistent with discouragement and is observed in spite of poor-performing students' apparent inability to update beliefs reflecting their low performance. Top performers receiving feedback become more accurate in their beliefs but are also 6 percentage points less likely to show up to the entrance exam.

Most of previous work providing feedback in the field focuses on the effects on students' grades and GPA but often cannot look at students investments and effort because they do not survey students.[15] The papers focusing on eliciting and correcting students' beliefs conduct surveys but are unable to look at how the feedback they provide to students affects their investments in preparing for an important exam (Bobba & Frisancho, 2016; Gonzalez, 2017). An exception is Azmat et al. (2019) who have measures of study hours and satisfaction. Relative to their paper, I can study other investments students make besides study time such as taking practice tests and the dynamics involved over time. I also provide evidence on how feedback affects extensive-margin decisions such as whether or not to take an important exam.

4.1 Estimation strategy

I show that there is a high degree of persistence in the quartiles the students are classified in at the beginning of the experiment. That is, students' initial quartile is highly predictive of their quartile in future practice tests and, hence, of the content of the feedback they receive if assigned to the treatment group or would have received if assigned to the control group. For this reason, the estimation is based on initial quartiles of performance.

To obtain treatment effects, I first run a regression of the outcome of interest on an indicator T_i (equal to one if the student was assigned to the treatment group), indicators for quartile in the initial practice test (Q_i), interactions between the treatments and quartile indicators, randomization strata fixed effects ($strata_i$), and baseline covariates ($\mathbf{X_i}$) (see equation 1). The excluded quartile is the bottom quartile so the interactions coefficients from this specification (τ_i) are the difference-in-differences estimates relative to the worst-performing students.

$$y_i = \beta_1 + \beta_2 T_i + \sum_{q=1}^{3} \alpha_q Q_i + \sum_{q=1}^{3} \tau_q Q_i * T_i + \rho strata_i + \mathbf{X_i}\gamma + \varepsilon_i \tag{1}$$

To obtain treatment effects by quartile, i.e., the difference in means between treated

[15] Measures of study time, class attendance, effort, and so on are usually not available in administrative records of institutions.

and control students in a specific initial quartile, I perform the following calculation for all quartiles except the bottom quartile, which is obtained directly from the point estimate of β_2 in equation 1:

$$\mathbb{E}[y_i|T_i = 1, Q_i = q] - \mathbb{E}[y_i|T_i = 0, Q_i = q] = \beta_2 + \tau_q \tag{2}$$

Where, for quartile q, the treatment effect is computed as the coefficient indicating treatment plus the interaction coefficient between the treatment and quartile q.

To show the high persistence in quartiles over the course of the experiment, Tables 4 and 5 reveal that the proportion of students who were initially classified in the top quartile are in the same quartile in about 50 percent of the subsequent practice tests and in the top two quartiles 75 percent of the time. Persistence among students who initially were classified in the bottom quartile is lower but still sizeable, with scores in the bottom two quartiles over 50 percent of the time. This persistence makes clearer the interpretation of the feedback students received in most reports: top performers received information that they were performing relatively well on practice tests and poor performers learned that they were at the bottom of the distribution.

4.2 Result 1: Low-performing students become discouraged by relative performance feedback

Students in the bottom quartile of initial performance who receive feedback take practice tests less often, study fewer hours, and perform worse in practice tests than similar students in the control group. This discouragement effect seems to carry over to the decision of taking the entrance exam. My evidence supports the hypothesis of discouragement rather than selection of students who may be have been demotivated from the beginning.

Table 6 pools data from all practice test rounds and shows that low performers (in bottom quartile of initial performance) in the treatment group are 5.2 percentage points less likely to take practice tests than students in the same level of performance in the control group. In the control group, students take 95.6 percent of all practice tests. The difference between the two groups means that treated students take one full practice test less out of 10 possible than the control group. Figure 9 shows that this effect is visible in one specific week in round 8 (practice test during Holy Week). Because students attend only 3 of the 5 days of the week during Holy Week, the fact that low-performing treated students do not show up may suggest that they do not have the same level of commitment than control students and prefer to use their time in activities outside the institute.

The likelihood of taking practice tests does not vary with treatment assignment in other

quartiles. On average, students in initial quartiles above the bottom take almost all practice tests and there is no difference between the treatment or the control group.

Turning to study time, in contrast with Azmat et al. (2019), I find that low-performing students receiving relative performance feedback study fewer hours per week than control students. Study time is self reported and collected in weekly surveys. The question is intended to elicit time spent working on exam-related problems, excluding class and practice test time. Low performers study 2 fewer hours for math and 1.5 fewer hours for reading than control students who study 6.3 and 5.2 hours per week, on average, respectively. These reductions, despite being only significant at the 10 percent level due to the small sample size in the bottom quartile, are substantial and economically significant at around 30 to 32 percent of weekly study time. Figures 10 and 11 show the weekly dynamics. In this case, the drop in study time is more notorious along the whole period than it was in the outcome measuring whether students took practice tests.

Once again, I do not find differences in study hours and performance in practice tests among treated and control students in quartiles of initial performance above the bottom. In fact, it is interesting to note that study time is relatively constant for students of different ability levels. No matter where in the performance distribution students lie in the first practice test, they spend between 5 and 6 hours studying for the math section, and between 4 and 5 hours studying for the reading section per week.

As expected from investing less time studying for the entrance exam, low-performing students receiving feedback obtain fewer correct practice test questions in math and reading (Table 6). Bottom performers have 1.7 and 1.3 fewer correct out of 40 questions in practice tests than students in the control group, who obtain 15.1 correct questions in math and 17.6 in reading, on average across all practice tests. The dynamics of these two variables by quartile are in Figures 12 and 13. In the rest of quartiles, performance in practice tests also does not differ by treatment status. It closely follows the level of ability observed in the initial practice tests, with scores increasing monotonically for students in higher quartiles of initial performance.

One dimension in which my research differs from the previous literature is that I can move forward from the analysis of performance and grades to study students' academic decisions such as whether or not to show up to an important scholastic exam. This type of decision is highly consequential and may determine not only the academic path that the students will follow but also many other life-time outcomes. I present evidence that top and bottom performers receiving feedback are less likely to show up to the entrance exam in two consecutive admission cycles. Among those who show up, performance is not statistically different between treatment and control.

Table 7 presents treatment effects for each of the four quartiles of the initial practice test score distribution. Not all students are equally affected by feedback. The stronger responses come from the top and bottom quartiles in which students receiving feedback are 10.8 and 5.8 percentage points less likely to take the April, 2018 entrance exam than students of similar ability in the control group, respectively. In both quartiles, 100 percent of students in the control group took the exam. Consistent with the findings described above, I do not find a significant effect in the middle quartiles, although treated students in quartile 2 (second best) are 4 percentage points less likely to show up.

In both cases, the top and bottom of the distribution, students may react to feedback by postponing because they feel they are not prepared enough. If students think they need extra time to prepare better for the exam, we may see them registering for the next admission cycle. Column 2 of Table 7 shows the results for an outcome variable equal to one if the students does not register in two consecutive admissions cycles (April and September, 2018). The results show that students who receive feedback are 7.8 and 11.6 percentage points more likely to never register if they are in the top and bottom quartiles, respectively. These results suggest that the effects of feedback is not related to gaining extra time to prepare for the exam, or at least not if the extra time is five more months.

Performance in the entrance exam is not affected by feedback but keep in mind that this is already mediated by the decision to take the exam. Columns 3 to 5 in Table 7 show that the scores students obtain in the exam increase monotonically with quartile in jumps of almost 10 points but do not differ statistically between treatment and control. Further, in line with no substantial differences in performance, admission rates are not statistically different across treatment assignments (table 8). Recall that admission at this university is highly selective so we do not expect many students gaining admission. Admission rates vary substantially between students above and below the median in the initial practice test at the institute. Students above the median have admission rates from 13 to 31 percent, while students below the median are very unlikely to be admitted with rates between 4 and 7 percent.

Another margin students can adjust is the choice of major and whether to declare a second choice major in case admission to the first choice does not work out. Recall that admission to the university depends on the score in the entrance exam as well as what major(s) students declare. Column 2 of Table 8 shows the treatment effects when the outcome is the cutoff score of the first choice major students declare based on the scores obtained by the cohort in the previous admission cycle.[16] Directionally, students seem to adjust in the right direction:

[16]The reasoning behind using the scores from a previous cycle is that cutoff scores are endogenous and may change in every cycle. So, students in a given semester cannot know the cutoff scores for the cycle they

treated students in quartile 2 and the bottom quartile choose majors with lower cutoff scores but this change (of about two points) is not enough to make a difference in the admission rates and not statistically significant. In fact, there is little variation in the cutoff scores of the first choice major students select, with students in all quartiles choosing majors whose cutoff scores were between 79 or 80 points (out of 100 possible points). Column 3 shows that admission to a second choice major is not very common for students in any quartile of initial performance.

To summarize the findings so far, low-performing students who receive relative performance feedback seem to get discouraged by this information and reduce investments, performance in practice tests, and are less likely to take the college entrance exam they are preparing for. An alternative explanation to these findings is that it is not discouragement but that I am simply picking up low-performers who were not very motivated from the start and had a higher likelihood to quit regardless of the treatment assignment. To study this alternative hypothesis, in Figure 14, I look at performance in the initial practice test and in subsequent practice tests for control students, and students in the treatment group who decided to take and not take the entrance exam, separately. If this hypothesis was correct, I would expect to see that students who end up deciding not to take the exam were performing worse than those who decided to take it. The figure shows the opposite: students who end up not taking the exam performed, if anything, slightly better than those who decided to take it. The decrease in performance comes in subsequent practice tests, where the density of those who ended up not taking the exam is substantially different and shifted to the left of the control and treated students who decided to take the exam. All in all, the evidence supports the discouragement hypothesis.

4.3 Result 2: Elicited beliefs do not correspond with beliefs revealed by behavior

Even though low-performing students change their behavior and decisions as a result of feedback, they do not seem to update their beliefs accordingly in the incentivized belief elicitation task. Top performers, on the other hand, report more accurate beliefs relative to control students with similar performance.

Since the 1960s, work in economics and psychology has documented the overconfidence phenomenon, the tendency of people to think that they performed better than they did or that they performed better than others. In general, overconfidence arises because people have

are applying at the time they declare what majors they want admission for (this is beofre they take the exam). However, cutoff scores from previous admission cycles are completely observable to them.

imperfect information about their own abilities or performances and know even less about others (Moore & Healy, 2008). It is often believed that correcting this imperfect information may result in better decision making. However, other effects, such as discouragement (as this paper finds), may take place. In fact, some research on overconfidence shows that people thinking that they are better than they actually are make them work harder (Chen & Schildberg-Hörisch, 2018). One advantage of my research is that I take belief elicitation outside the lab setting, which allows me to track how students update beliefs and can compare with the behaviors they engage in in their day-to-day academic life.

In the weekly lab-in-the-field experiment, beliefs are elicited twice: right after the students take the practice test (elicits priors) and once again after students check the performance report (elicits posteriors). Both, treatment and control students report relative performance beliefs twice per week. The difference between treated and control students is the information they see in the performance report. Control students see the standard report provided by the institute that contains the number of correct questions in math and reading and a global score that is supposed to resemble the score they would obtain in the real entrance exam (see Figure 1). The report for treatment students includes a "signal", an additional piece of information telling them whether their scores are above or below the median. The timeline of how belief elicitation works in a given week is in Figure 4.

Across all rounds of prior belief elicitation, less than 35 percent of students have correct relative performance beliefs. Figure 7 shows the classification of prior beliefs resulting from comparing the quartile to which the students assigned the highest probability relative to the actual quartile in which their performance lies. For example, students with correct priors are those who assigned most tokens to the quartile in which their score is. They are classified as over- (under-) placing when they assign most tokens to a quartile that is above (below) their performance quartile. Students could also report a flat prior if they did not know where there performance was or an inconsistent prior if they assign most tokens to non-consecutive quartiles.

In lab experiments most people overplace their performance relative to others (Moore & Healy, 2008). However, in my setting overplacing is as likely as underplacing. In Figure 7, about 25 percent of the students overplace their performance and another 25 percent underplace it. This is not surprising given that it has been found that more people think their performance is lower than others when the task is difficult than when it is easy (Moore & Healy, 2008). Less than 10 percent of the students have a flat prior and less than 5 percent report an inconsistent belief.

At the posterior stage, combining treatment and control students, belief accuracy improves, with almost 50 percent of students reporting a correct belief (see Figure 8). The

fraction of students overplacing is reduced to about 20 percent while the fraction underplacing remains similar to the prior stage.

Students in the best quartile of performance are between 25 and 30 percent more likely to hold correct priors across all rounds than students in the control group. Tables 9 and 10 present the effect of receiving feedback on students' prior beliefs. Across all rounds, about 41 percent of top-performing control students have correct beliefs in reading, 23 percent overplace and 27 percent underplace their score. Treated students are 10 percentage points or 25 percent more likely to be correct at the prior stage (before receiving the above / below median signal). Most of the change in beliefs comes from lower overplacement, that is, students are less likely to think that they are in a higher quartile than they are when their practice test score is below the top quartile. The patterns for math in table 10 are similar in magnitude.

Low performers' beliefs, however, do not change substantially relative to their peers in the control group. The bottom panel of Tables 9 and 10 presents the treatment effect on prior beliefs for students in the bottom quartile of initial practice test performance. Roughly 30 percent of these students have correct beliefs, 36 percent overplace, and about 16 percent underplace in reading and 25 percent in math. In contrast with top performers, there are no statistical differences for the treatment group, which suggests that bottom performers are not very successful at incorporating the information they receive through feedback. There is, nevertheless, some indication that low performers become less likely to overplace their score as a result of receiving feedback.

Students in the middle quartiles of middle performance do not become more correct in their assessments of performance. The only difference between treatment and control in these quartiles is that treated students in quartile 2 of initial performance are more likely to overplace their score in reading (but not in math).

To sum up, top performing treated students seem to be more successful at incorporating feedback in their belief elicitation responses that students in other quartiles of initial performance. The question of why this is the case can be approached from different angles. One is that top-performing students are more cognitively able and can understand the elicitation task better. The fact that almost 42 percent of control students in the top quartile (vs. about 30 percent of control students in other quartiles) have correct beliefs provides some evidence for this hypothesis (see Tables 9 and 10). Another hypothesis is that, when students are not at the top, receiving information is discouraging and they may prefer to avoid it. Figure 15 shows the distribution of the number of times (out of a total of 8) that students from each quartile check the performance report. While there are no differences in the distribution of report checking for treated and control students in quartiles 1 and 2 of

initial performance, treated students in quartiles 3 and 4 are 62 and 76 percent more likely to check only one performance report than control students in the same quartiles, respectively. Section 5 provides a conceptual framework to understand the discrepancy between elicited beliefs and beliefs revealed by students' observed behavior.

4.4 Result 3: Female and male students adjust different margins

This section shows that discouragement can manifest in different ways. While low performing treated men's effort suffers as a result of receiving relative performance feedback, treated women keep effort up but are more likely to decide against taking the college entrance exam. To estimate treatment effects by gender, I add a female dummy and interactions with the terms in equation 1 to obtain difference-in-differences (DiD) estimates, and compute within gender treatment effects using the analogous to equation 2.

Studying differential responses by gender is natural in this context given the vast evidence that female students tend to perform better than males in areas like reading, while male students tend to perform better than females in areas like math. The OECD shows that, for the Program for International Student Assessment (PISA) test, there are few countries around the world where the gender gap in these subjects does not exist or is reversed (OECD, 2015). Furthermore, a growing literature has documented a gender gap favoring male students in high-stakes, competitive environments such as college entrance exams (e.g. Jurajda & Münich, 2011; Ors et al., 2013; Cai et al., 2016).

From my intervention, the largest response in terms of academic investments is observed among low-performing male students (Table 11). Treated men in the bottom quartile are 6.7 points less likely to attend practice tests, and study math about half of the time per week than male control students. Low-performing female students have effects going in the same direction, but power is not enough to detect statistical significance of the within female effect or the DiD effect.

In a previous section I found lower scores in practice tests among students in the bottom quartile of initial performance but decomposing the effect by gender does not allow to establish if the effects is driven by students of a specific gender. I only find strong negative within gender and DiD effects on the number of correct answers in math and reading in practice tests among females in quartile 2. This is likely explained by the fact that treated students see a signal indicating that their scores are above the median but subsequently receive feedback saying that they are at the "bottom of the top". In fact, Tables 14 and C.6 show that both, men and women in quartile 2, tend to overplace their performance. Why it affects more women's performance could be related to women reacting more strongly to the

negative news and feeling extra pressure to perform well on the tests.[17]

Table 13 shows that, even though the treatment effects for women and men tend to go in the same direction, the effect on exam taking is stronger for women. Treated women in the top quartile are 7.2 percentage points less likely to take the entrance exam and 7.5 percentage points less likely to register for the exam over two admission cycles relative to control women. A closer examination of this effect indicates that these women who did not take the exam were performing better in math (but not in reading) than females who were in the same quartile of initial performance but decided to take the exam.[18] On average, across all practice tests, treated women who decided to not take the exam were placed 17 percentiles above women who decided to take it. Moreover, these women were more accurate in their quartile predictions. They were 28 and 20 pp more likely to provide a correct assessment of their practice test performance over a base of 48 and 58 percent, for math and reading respectively, than treated women who decided to take the exam. However, they do not know that their performance is in the higher percentiles because they do not check the detailed information more often than women who decided to take the exam.[19] All in all, it seems that these high-performing women are pessimistic because, despite their good performance and accurate beliefs, they are less likely to report that they are very confident in their likelihood of gaining admission to the university.

Finally, while performance in the last practice test, which took place one week before the entrance exam, is not very different across genders, men outperform women in the entrance exam (Figure 16). Explaining this pattern is out of the scope of this paper but it may lend support the hypothesis of women being more likely to "choke under pressure" as they are able to perform equally well as men in a practice test but not in the real exam.

5 Conceptual framework

In this section I use a conceptual framework with the aim of revisiting the findings suggesting a discrepancy between elicited beliefs and beliefs revealed by students' behavior and propose some hypothesis to explain the discrepancies.

In Bayesian learning, individuals have beliefs about their ability that evolve according

[17] Research in educational psychology has found that women tend to report higher test anxiety (e.g. Bors et al., 2006) and that female students who are high-test anxious perform worse than those who are low-test anxious while there is not difference for men (e.g. Chapell et al., 2005).

[18] The evidence presented here is meant to be descriptive only because the decision to take the exam is endogenously determined and the sample sizes for students who did not take the exam is very low.

[19] The experimental protocol in Appendix B shows that after treated students see the quartile feedback, they can click on a button to obtain more "detailed information," which is basically their percentiles in the distribution.

to signals that they receive from different sources. Let α_i be individual i's true ability, and $\mu_i = \alpha_i + \varepsilon_i$ be the belief the individual holds about her ability with $\varepsilon_i \sim N(0, \sigma^2)$.[20] This belief is formed along time based on the individual's past experiences such as in the schooling system. In my setting, students are preparing for a college entrance exam in which what matters is two things: the ranking of their absolute score and the college majors they declare before taking the exam (and knowing their performance). In this sense, to obtain a slot at the university, what is important is their relative performance.

The intervention I conduct at the institute consists in providing students with signals of their relative ability specific to the entrance exam. Because practice tests only measure ability imperfectly, each new signal s_i will have a random component: $s_i = \alpha_i + \varepsilon_i^s$, $\varepsilon_i^s \sim N(0, \sigma_s^2)$. If individuals are Bayesians, priors are suffiicient statistics for past information so we can write the information content of the signal as: $I_{i,t+1} = s_i - \mathbb{E}[s_i|\Omega_{i,t}]$, with $\Omega_{i,t}$ being the set of information available to the individual at time t. Because priors contain all past information relevant to individuals, they only use new information to update relative ability beliefs:

$$\mathbb{E}[\alpha_i] = \gamma\mu_i + \rho I_{i,t+1} \tag{3}$$

Where γ and ρ are relative weights given information up to time t and new information received in $t+1$, respectively.

One of the main findings of this paper is that, even though students seem to be using the new information received as revealed by their choices related to the entrance exam, they do not necessarily express their change in beliefs in the lab elicitation task. That is, had I not elicited beliefs, it would be straightforward to conclude that individuals seem to behave according to Bayesian learning. However, this is not what they express when stating their beliefs in the lab elicitation task.[21] In the framework of this section, students not updating would be equivalent to $\mathbb{E}[\alpha_i] \approx \gamma\mu_i$, that is, either the weight given to $I_{i,t+1}$ is very low or studnets simply decide to opt out of new information. I discuss several hypotheses that may be behind this inconsistency. My design did not intend to disentangle different hypotheses, but it is potentially an interesting area for future work.

The first hypothesis is that learning information that one is not a good performer may affect students' ego or impose a psychological cost that motivates them to report beliefs consistent with optimistic self-deception (Bénabou & Tirole, 2002). This hypothesis would suggest that, even though they know their performance is not good, they do not want to

[20] The framework presented here follows Gonzalez (2017).

[21] To my knowledge, there is only one other paper that finds a disconnect between reported beliefs and actual choices. In a lab experiment, (Sautua, 2018) shows that individuals express a belief consistent with the gambler's fallacy, but their choices reflect a belief that is contrary to the fallacy.

feel bad by learning this information so decide to not report the truth in the lab task. In fact, some theoretical models include a direct belief utility component in the utility function (Köszegi, 2006; Mobius et al., 2014) that captures that the individuals care about their belief about how good they are. Alternatively, they could set $\rho = 0$ by deciding to not look at the information provided or have a very low ρ, by which the value of new information has a very low weight in their new belief formation. My experiment cannot provide direct evidence of optimistic self-deception, but it has certainly been documented than when individuals care about their ego their information processing differs than when the task is ego irrelevant (Ertac, 2011). I do provide some evidence and discuss in subsection 4.3 and Figure 15 that low-performing students engage in information avoidance by looking at feedback only once and stopping thereafter.

A second hypothesis suggests that the value of ρ can vary according to the informational content of $I_{i,t+1}$. Rabin and Schrag (1999) propose a model of confirmatory bias where individuals are more likely to take into account signals confirming their prior and give less importance to signals disconfirming their prior. Because I elicit posteriors right after providing the above-/below-median signal to treated students, I am able to present evidence regarding confimation bias. Appendix Figure C.18 shows how much individuals update relative to a Bayesian in four scenarios.[22] In the left panel, the left bar shows how much students update when they are above the median and this information confirms their belief. The right bar shows a disconfirming signal, that is, they thought they were below the median but receive the signal that they are above. The right panel shows the confirming signal that they are below the median on the left bar. The right bar shows the disconfirming signal that they are below the median when they thought they were above. In both cases students update more in line with the Bayesian benchmark when confirming rather than disconfirming signals are received, providing support for confirmation bias.

A third hypothesis, outside of the conceptual framework, involves individuals' potential low ability to understand the task or being inattentive when solving it (Gabaix, 2017; Chetty et al., 2009; DellaVigna, 2009). Complexity of the task or limited attention could be particularly relevant in the case of individuals with lower ability. In fact, from paper surveys, where there is no way to make the sum of tokens add to 12, I find that students in the bottom quartile are more likely to make mistakes in assigning the 12 tokens. I create an indicator variable for the sum of the tokens assigned across quartiles not being equal to 12. About 16 percent of top performers but 34 percent of bottom performers make this type of

[22]Updating is computed according to Bayes' formula using students' priors collected after each weekly practice test. In the figure, I compare what fraction of the updating I see in students' posteriors compared to a Bayesian with the same priors as the students'.

mistake. Indeed, lower ability students had a harder time understanding the task. There is also less consistency in how students assign tokens in the belief elicitation task which could be related with lack of understanding of the task (see Appendix Figure C.19).

The fourth hypothesis relates to the stakes of the lab task and of the decisions students face. As in standard laboratory experiments, the stakes of the lab-in-the-field belief elicitation task are relatively small and focused on monetary incentives. The stakes of the real decisions, on the contrary, are quite high, with some of the decisions the students make influencing their future in terms of employment prospects, income and economic mobility. Because the likelihood of winning a prize was relatively small and they may care more about their ego than about winning a cash prize as hypothesis 1 states, they may have decided to not update beliefs in the lab task but update their beliefs for the decisions that matter. This type of behavior may be consistent with a "dual beliefs" model in which one self acts according to one set of beliefs in real-life decision making and the other self acts according a another set of beliefs for other - probably lower stakes - decisions.

Finally, there is incomplete evidence on how well belief elicitation tasks perform at eliciting the beliefs that people use to make decisions outside of the lab. Even though there is general consensus that belief elicitation in the lab is a good approximation to turn latent into observable beliefs (Schotter & Trevino, 2014), there is still lack of evidence on how good these elicitation mechanisms are outside of the lab generates data that is meaningful and relevant in ego-relevant contexts as the one I study.

Regardless of their motivation to report beliefs incoherent with their behavior, low performers receiving feedback classify in the definition of "dropouts" (Müller & Schotter, 2010). The term refers here to individuals being dissuaded by their low probability of gaining admission. Other examples of this behavior have been documented in elementary school kids stopping when running a race when it is clear they have no chance at winning (Fershtman & Gneezy, 2011), and disadvantaged students not being willing to invest in an SAT preparation course after they learn their ability (Benoit, 1999). Top performers are more accurate in predicting where their scores lie in the distribution but are still less likely to take the entrance exam. Moreover, it seems that those who decide not to take the exam are among the best performers but tend to be pessimistic. Again, this behavior can be consistent with "workaholics" in the sense introduced by Müller and Schotter (2010). These are individuals who seem unable to stop working because they do not think they are ever good enough.

6 Conclusion

Information failures in the context of education are widespread and research shows that they are sizable enough to affect students' decisions and outcomes. It has been found that having access to information about the returns to education (Jensen, 2010; Nguyen, 2008), school quality (J. S. Hastings & Weinstein, 2008; Mizala & Urquiola, 2013), application procedures (Hoxby, Turner, et al., 2013), financial aid (Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2012; Dinkelman & Martínez, 2014), and future earnings (Attanasio & Kaufmann, 2014; Wiswall & Zafar, 2015a, 2015b; J. Hastings, Neilson, & Zimmerman, 2015) helps students make decisions that better correspond to their academic abilities.

A more recent literature acknowledges that schooling choices are made under uncertainty (Altonji, 1993; Altonji et al., 2016), and that one source of uncertainty is the lack of information about students' own ability (Bobba & Frisancho, 2016; Gonzalez, 2017) or parents' knowledge about their children's ability (Dizon-Ross, 2018). These papers find positive effects on academic decisions when information about *absolute* performance is provided to students or parents. It follows that a straightforward policy recommendation would be to provide more information. This paper argues, nevertheless, that more information may have unexpected effects in domains beyond academic performance. Previous literature has focused on academic performance and has highlighted that providing detailed *relative* performance information such as class rank or decile in the grade distribution reduces test scores (Murphy & Weinhardt, 2018; Megalokonomou & Goulas, n.d.; Azmat et al., 2019). Other studies have found positive effects on grades when relative information refers to a summary statistic (Azmat & Iriberri, 2010; Tran & Zeckhauser, 2012), to low-stakes tests (Dobrescu et al., 2019), or framed in a non-negative way (Brade et al., 2018). My paper provides evidence on changes in students' belief updating, and academic investments and decisions in a college entrance exam setting, where information about relative performance is relevant.

I design a field and lab-in-the-field experiments to understand how relative performance feedback affects students' beliefs, decisions and academic performance. Apart from the random assignment of students, I assemble a panel dataset from student surveys and use administrative data to reach three main findings. First, low-performing students receiving feedback become discouraged and take fewer practice tests, study fewer hours and perform worse than students of the same ability who do not receive feedback. The treatment also seems to dissuade them from taking the exam they are preparing for, an effect that can also be observed for top-performing students. Second, beliefs elicited using a lab-in-the-field task do not change by treatment status except for students who are at the top of the distribution. This suggests a mismatch between elicited beliefs and beliefs revealed by observed behavior.

Third, men and women adjust different margins, with men reducing effort and women shying away from taking the college entrance exam.

Contrary to the traditional idea that "information can't hurt", I present evidence that students across the whole distribution of academic performance are not affected equally by relative performance feedback, and that there are important changes in decision making in a high-stakes setting. Providing relative performance information can discourage students at the bottom of the score distribution to try harder and to opt out of taking an important exam. On the one hand, this could be thought of as efficient, since higher ability students who have higher chances of gaining admission will be the ones competing for the university slots. On the other hand, it may not be ideal in other contexts or from a policy perspective seeking to reward effort rather than achievement.

One of the main strengths of my paper relates to the fact that most studies in the previously discussed strands of the literature can only study effects at two points in time and are therefore only able to capture effects on final outcomes such as performance in an exam. By eliciting beliefs and providing high-frequency (weekly) feedback, I generate insights on how beliefs and investments change in the period between the intervention and observing final academic outcomes, and on the margins beyond grades in which students are affected by feedback.

The main limitation is that I study a very particular setting in which relative performance beliefs are important, but settings in other countries or with other populations may vary substantially so it is hard to generalize. Another limitation is that, due to small sample sizes, power may be limited to detect some effects that are economically meaningful but not statistically significant.

An important policy implication of this paper is that educational institutions and testing agencies must be cautious on how they provide relative performance feedback to students if they want all students, regardless of their ability level, to try harder. In this sense, my findings should raise awareness on how students are being informed of their performance, which is especially important given that providing rank within a class is a widespread practice across the world.

References

Alan, S., Boneva, T., & Ertac, S. (2016). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit.

Altonji, J. G. (1993). The demand for and return to education when education outcomes are uncertain. *Journal of Labor Economics*, *11*(1, Part 1), 48–83.

Altonji, J. G., Arcidiacono, P., & Maurel, A. (2016). The analysis of field choice in college and graduate school: Determinants and wage effects. In *Handbook of the economics of education* (Vol. 5, pp. 305–396). Elsevier.

Attanasio, O. P., & Kaufmann, K. M. (2014). Education choices and returns to schooling: Mothers' and youths' subjective expectations and their role by gender. *Journal of Development Economics*, *109*, 203–216.

Azmat, G., Bagues, M., Cabrales, A., & Iriberri, N. (2019). What you know can't hurt you: A natural field experiment on relative performance feedback in higher education. *Management Science*.

Azmat, G., & 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.

Azmat, G., & Iriberri, N. (2016). The provision of relative performance feedback: An analysis of performance and satisfaction. *Journal of Economics & Management Strategy*, *25*(1), 77–110.

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

Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, *117*(3), 871–915.

Benoit, J.-P. (1999). Color blind is not color neutral: testing differences and affirmative action. *Journal of Law, Economics, and Organization*, *15*(2), 378–400.

Berlin, N., & Dargnies, M.-P. (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization*, *130*, 320–336.

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the h&r block fafsa experiment. *The Quarterly Journal of Economics*, *127*(3), 1205–1242.

Bobba, M., & Frisancho, V. (2016). Learning about oneself: The effects of signaling ability on school choice. *Inter-Am. Dev. Bank, Discuss. Pap*, *450*.

Bors, D. A., Vigneau, F., & Kronlund, A. (2006). L'anxiété face aux examens: Dimension-

nalité, similitudes et différences chez les étudiants universitaires. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *38*(2), 176.

Brade, R., Himmler, O., & Jäckle, R. (2018). Normatively framed relative performance feedback–field experiment and replication.

Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, *129*(3), 1409–1447.

Buser, T., Peter, N., & Wolter, S. C. (2017). Gender, competitiveness, and study choices in high school: Evidence from switzerland. *American economic review*, *107*(5), 125–30.

Buser, T., & Yuan, H. (2016). Do women give up competing more easily? evidence from the lab and the dutch math olympiad.

Cai, X., Lu, Y., Pan, J., & Zhong, S. (2016). Gender gap under pressure: Evidence from china's national college entrance examination.

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test anxiety and academic performance in undergraduate and graduate students. *Journal of educational Psychology*, *97*(2), 268.

Chen, S., & Schildberg-Hörisch, H. (2018). *Looking at the bright side: The motivation value of overconfidence* (Tech. Rep.). DICE Discussion Paper.

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *American economic review*, *99*(4), 1145–77.

DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic literature*, *47*(2), 315–72.

Dinkelman, T., & Martínez, C. (2014). Investing in schooling in chile: The role of information about financial aid for higher education. *Review of Economics and Statistics*, *96*(2), 244–257.

Dizon-Ross, R. (2018). Parents' perceptions and children's education: Experimental evidence from malawi. *American Economic Review (forthcoming)*.

Dobrescu, L. I., Faravelli, M., Megalokonomou, R., Motta, A., et al. (2019). *Rank incentives and social learning: Evidence from a randomized controlled trial* (Tech. Rep.). Institute of Labor Economics (IZA).

Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–38.

Ertac, S. (2011). Does self-relevance affect information processing? experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, *80*(3), 532–545.

Fershtman, C., & Gneezy, U. (2011). The tradeoff between performance and quitting in high

power tournaments. *Journal of the European Economic Association*, *9*(2), 318–336.

Gabaix, X. (2017). *Behavioral inattention* (Tech. Rep.). National Bureau of Economic Research.

Gill, D., Kissová, Z., Lee, J., & Prowse, V. L. (2016). First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision.

Golman, R., Hagmann, D., & Loewenstein, G. (2017). Information avoidance. *Journal of Economic Literature*, *55*(1), 96–135.

Gonzalez, N. (2017). *How learning about one's ability affects educational investments: Evidence from the advanced placement program* (Tech. Rep.). Mathematica Policy Research.

Grossman, Z., & Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, *84*(2), 510–524.

Hastings, J., Neilson, C. A., & Zimmerman, S. D. (2015). *The effects of earnings disclosure on college enrollment decisions* (Tech. Rep.). National Bureau of Economic Research.

Hastings, J. S., & Weinstein, J. M. (2008). Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly journal of economics*, *123*(4), 1373–1414.

Hoxby, C., Turner, S., et al. (2013). Expanding college opportunities for high-achieving, low income students. *Stanford Institute for Economic Policy Research Discussion Paper*(12-014).

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, *125*(2), 515–548.

Jurajda, Š., & Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to czech universities. *American Economic Review*, *101*(3), 514–18.

Köszegi, B. (2006). Ego utility, overconfidence, and task choice. *Journal of the European Economic Association*, *4*(4), 673–707.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, *72*(5), 1329–1376.

Megalokonomou, R., & Goulas, S. (n.d.). Knowing who you are: the effect of feedback on short and long term outcomes.

Mizala, A., & Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics*, *103*, 313–335.

Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2014). *Managing self-confidence* (Tech. Rep.). National Bureau of Economic Research.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502.

Müller, W., & Schotter, A. (2010). Workaholics and dropouts in organizations. *Journal of*

the European Economic Association, *8*(4), 717–743.

Murphy, R., & Weinhardt, F. (2018). *Top of the class: The importance of ordinal rank* (Tech. Rep.). National Bureau of Economic Research.

Nguyen, T. (2008). Information, role models and perceived returns to education: Experimental evidence from madagascar. *Unpublished manuscript*, *6*.

Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, *122*(3), 1067–1101.

OECD. (2015). *The abc of gender equality in education: Aptitude, behaviour, confidence.* http://dx.doi.org/10.1787/9789264229945-en.

Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*, *31*(3), 443–499.

Rabin, M., & Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The Quarterly Journal of Economics*, *114*(1), 37–82.

Reuben, E., Wiswall, M., & Zafar, B. (2015). Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *The Economic Journal*.

Sautua, S. (2018). *When diversification clashes with the reinforcement heuristic: an experimental investigation* (Tech. Rep.). unpublished working paper.

Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.*, *6*(1), 103–128.

Tran, A., & Zeckhauser, R. (2012). Rank as an inherent incentive: Evidence from a field experiment. *Journal of Public Economics*, *96*(9-10), 645–650.

Wiswall, M., & Zafar, B. (2015a). Determinants of college major choice: Identification using an information experiment. *The Review of Economic Studies*, *82*(2), 791–824.

Wiswall, M., & Zafar, B. (2015b). How do college students respond to public information about earnings? *Journal of Human Capital*, *9*(2), 117–169.

7 Figures

EVALUACIÓN	FECHA	PUNTAJE
PRE-U DE A 2018-I	ENERO 23, 2018	2.5
NOMBRES - APELLIDOS	**CÓDIGO**	**# DE LISTA**	**TOTAL UdeA**
ESTUDIANTE 001 PRUEBA	1000001	1	25.0

RAZONAMIENTO LÓGICO MATEMÁTICO				
CATEGORÍA	# PREGUNTAS	CORRECTAS	INCORRECTAS	PUNTAJE
ANALISIS	10	2	8	2.0
TOTAL	10	2	8	2.0

COMPETENCIA LECTORA				
CATEGORÍA	# PREGUNTAS	CORRECTAS	INCORRECTAS	PUNTAJE
INTERPRETACION	10	3	7	3.0
TOTAL	10	3	7	3.0

Figure 1: Screenshot results report to all students

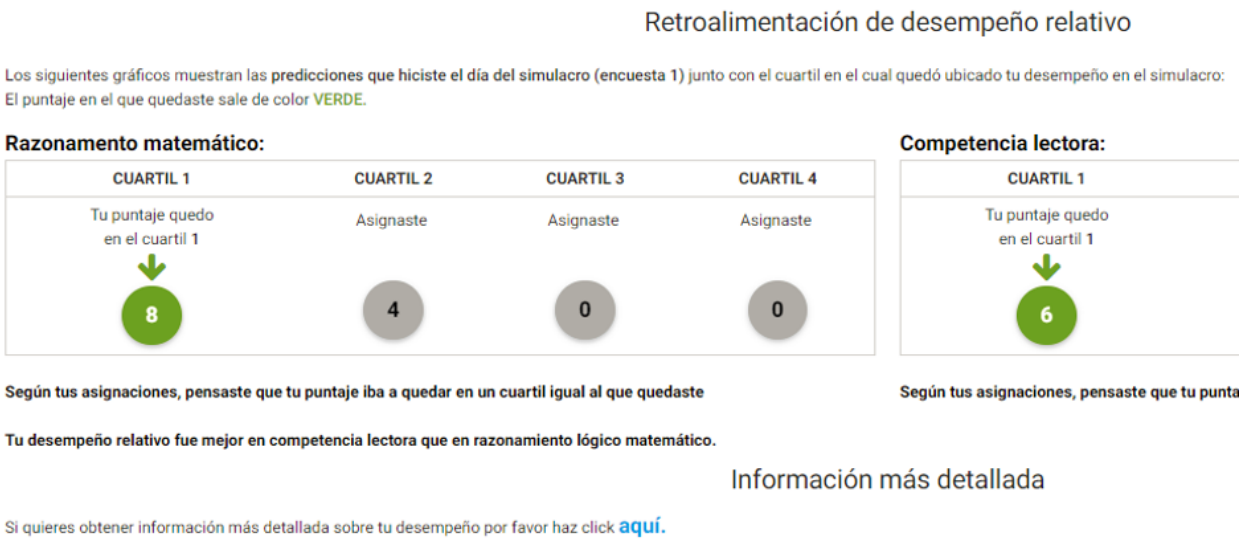


Figure 2: Screenshot relative performance feedback to treated students

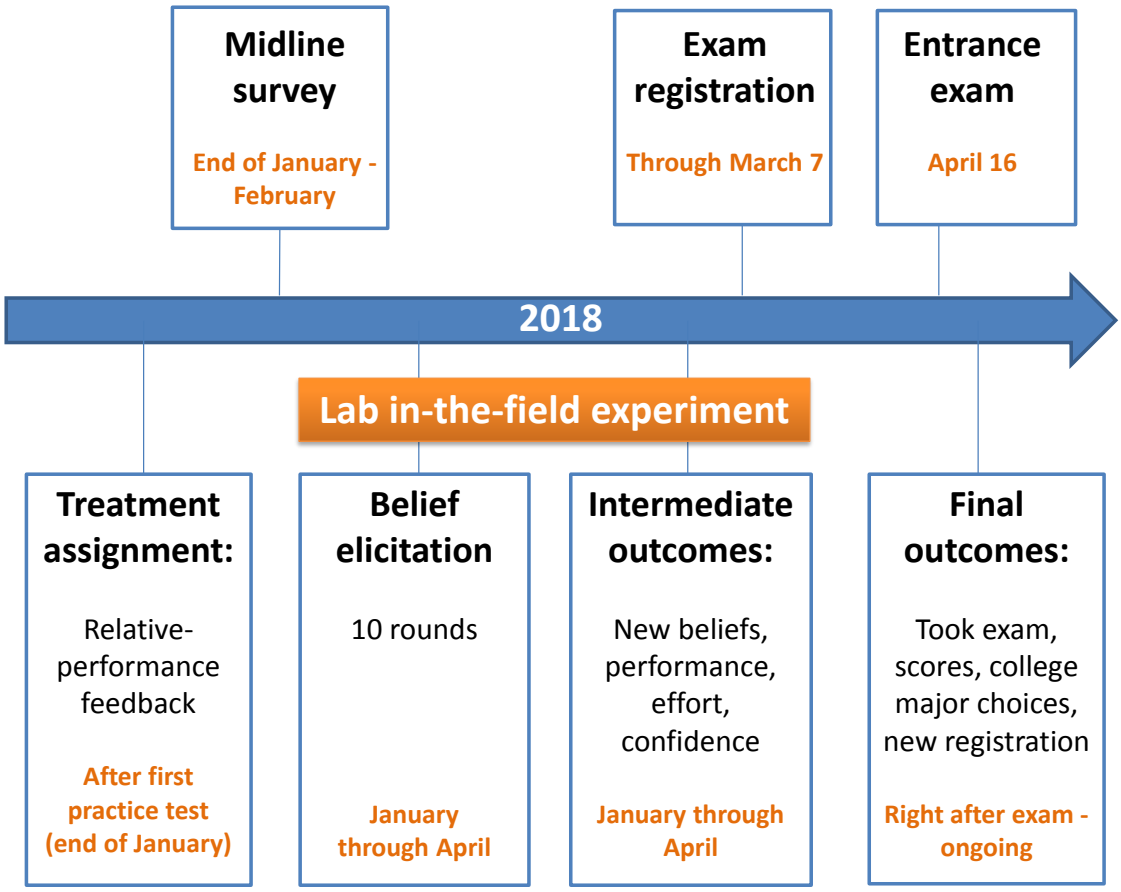


Figure 3: Timeline

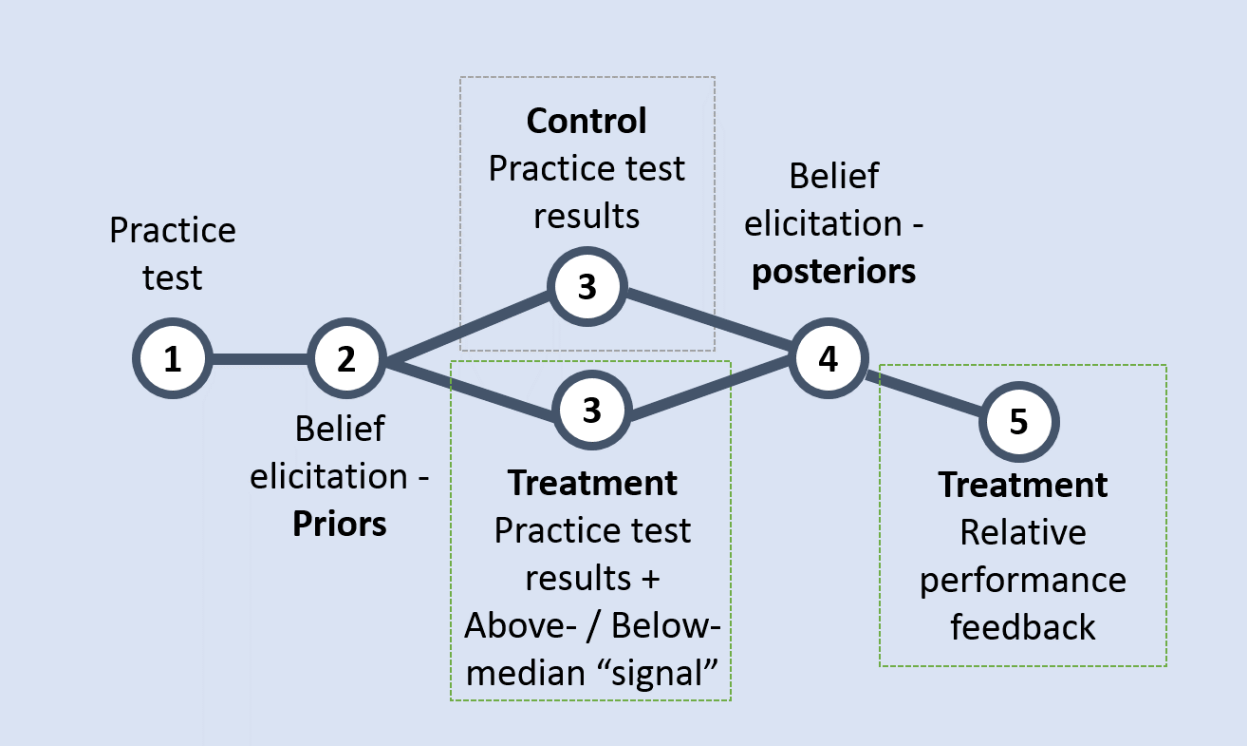


Figure 4: Lab-in-the-field experiment timeline

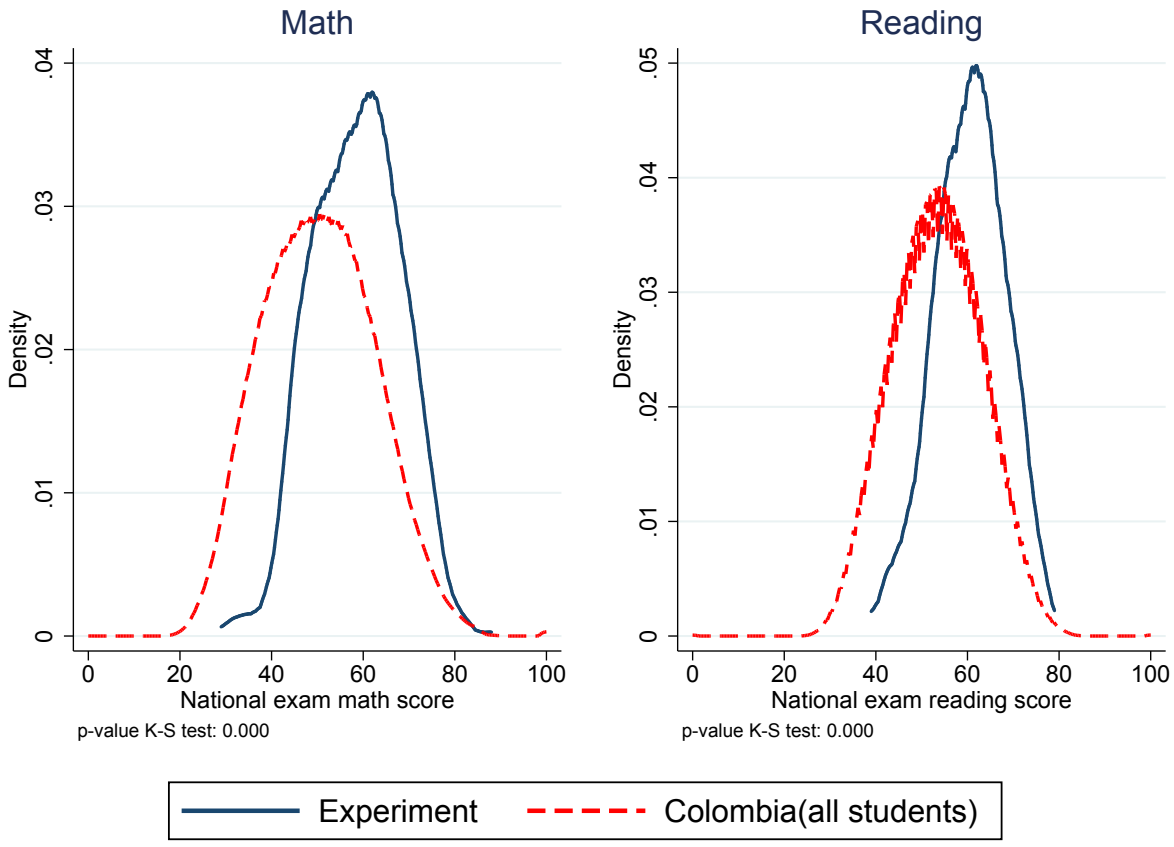


Figure 5: Positive selection of students in the sample relative all high-school graduates in Colombia

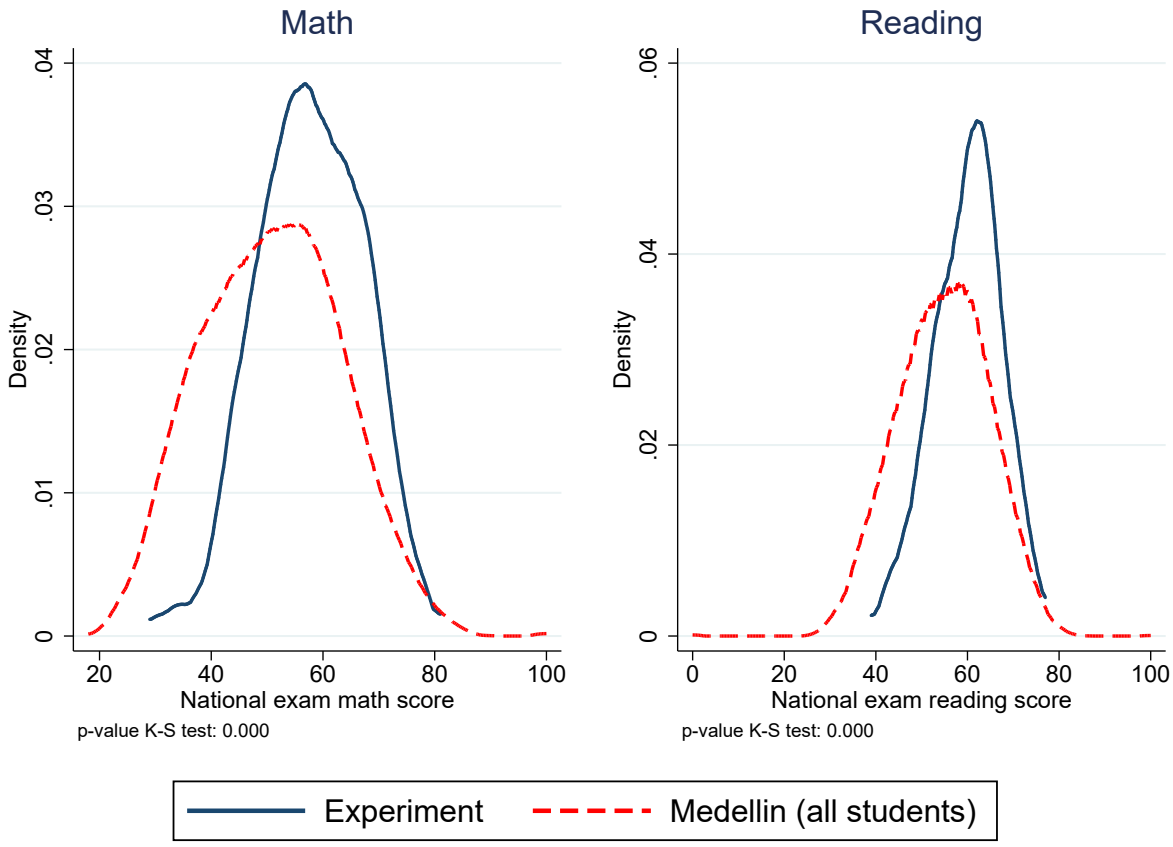


Figure 6: Positive selection of students in the sample relative all high-school graduates in Medellin

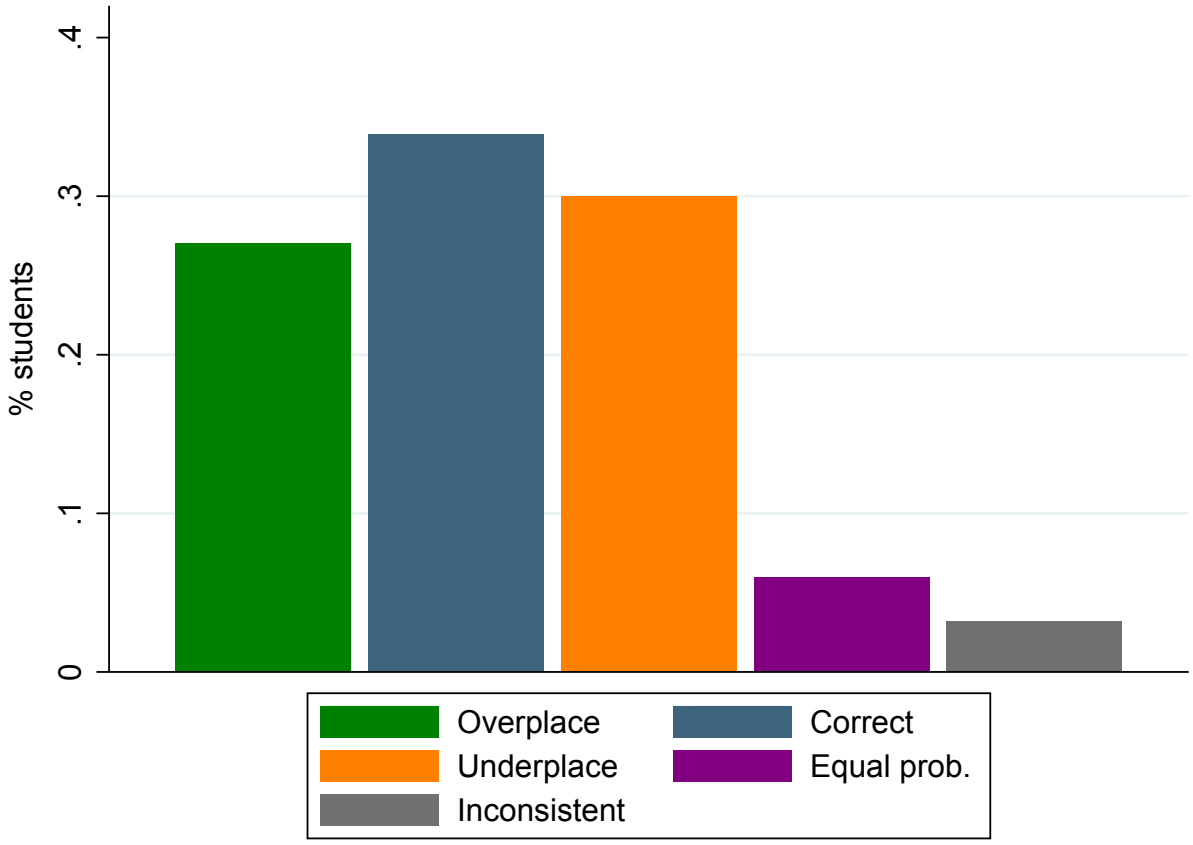


Figure 7: Classification of prior beliefs depending on quartile with highest probability assignment and actual quartile - all students

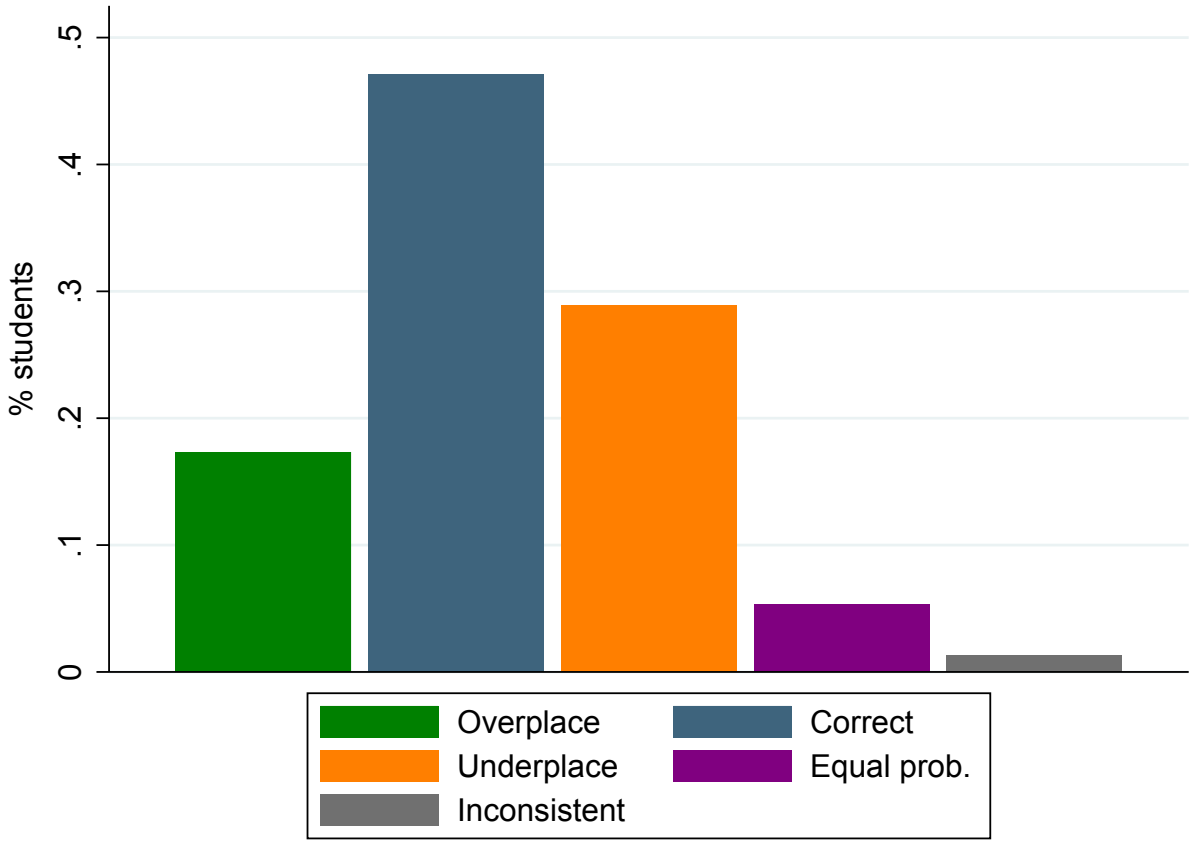


Figure 8: Classification of posterior beliefs depending on quartile with highest probability assignment and actual quartile - all students

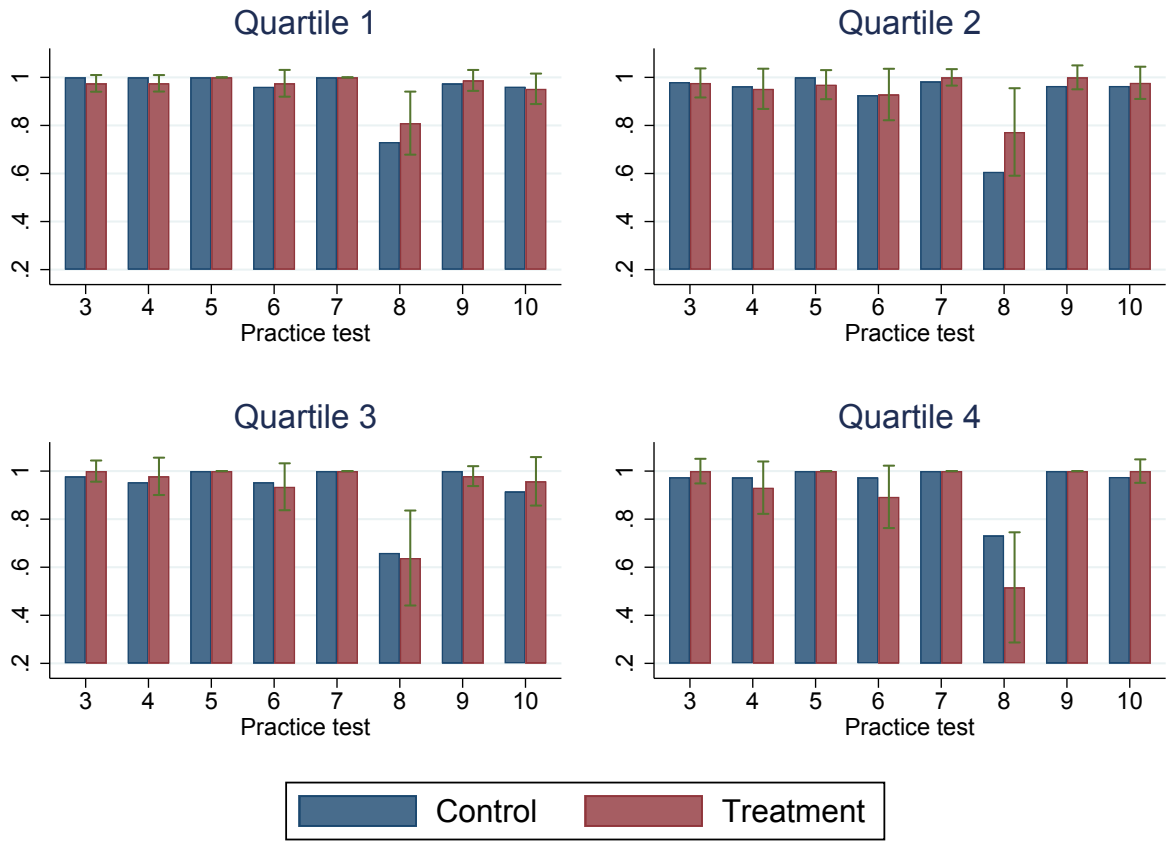


Figure 9: Fraction of students taking practice tests by week and quartile

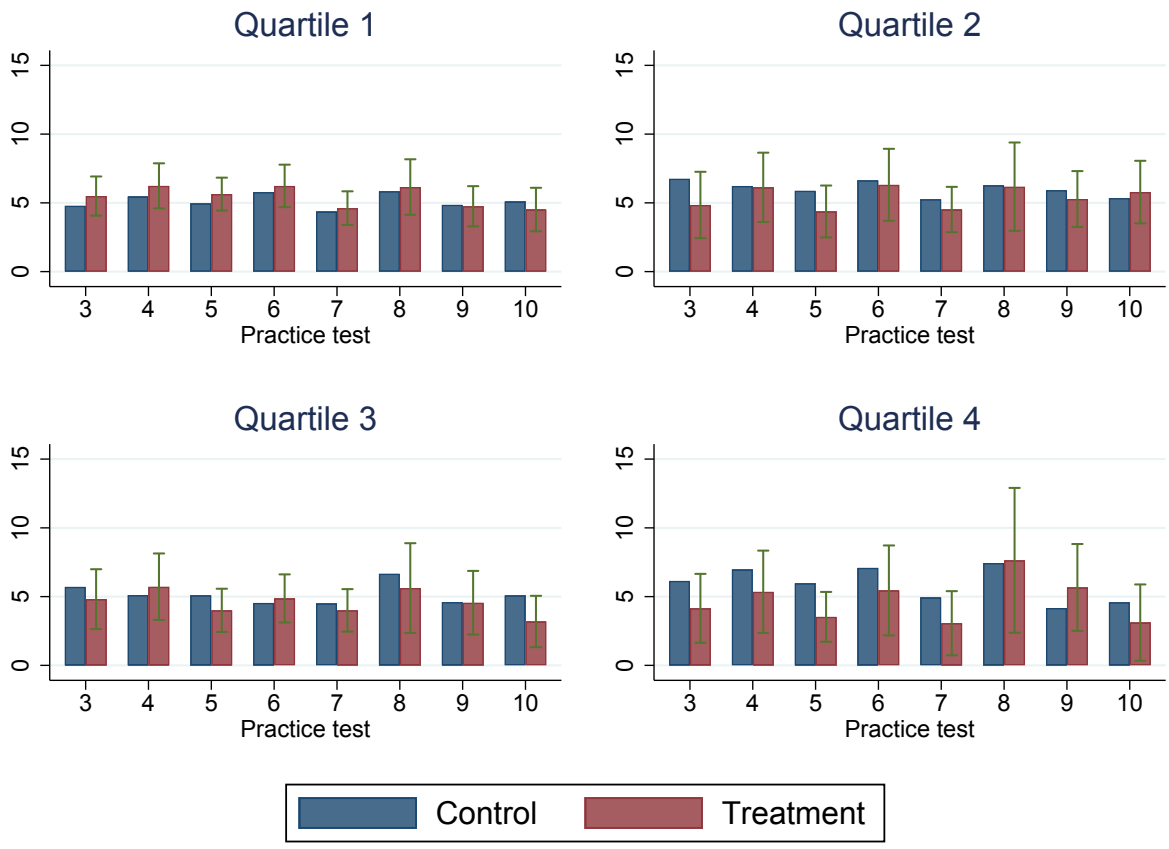


Figure 10: Average study hours in math by week and quartile

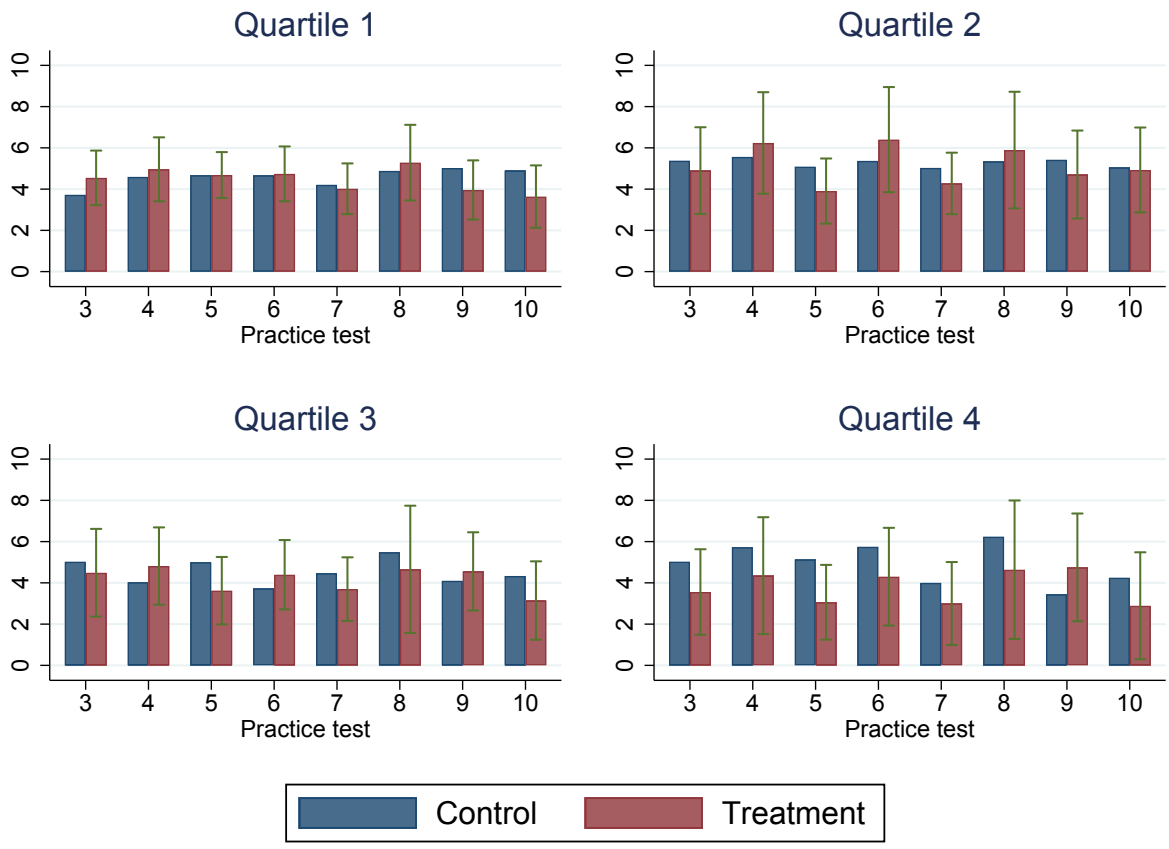


Figure 11: Average study hours in reading by week and quartile

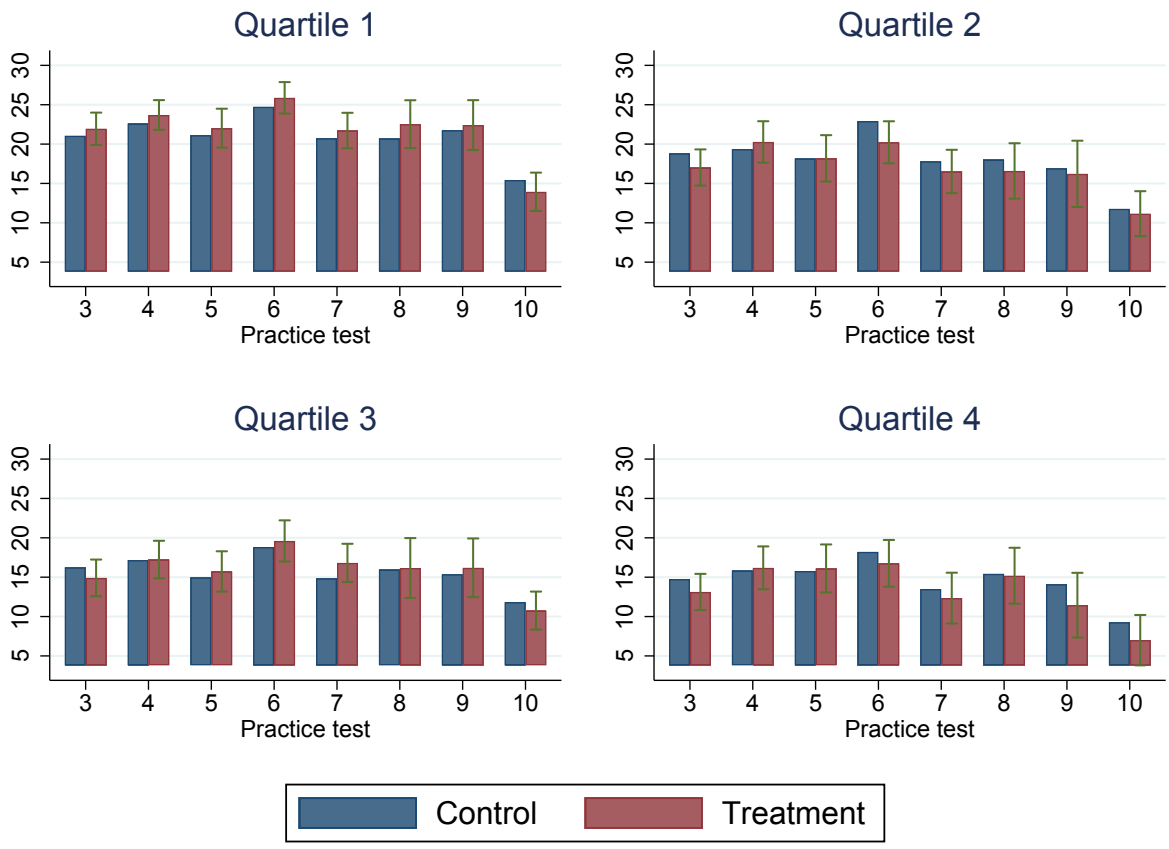


Figure 12: Practice tests: Number of correct answers in math by week and quartile

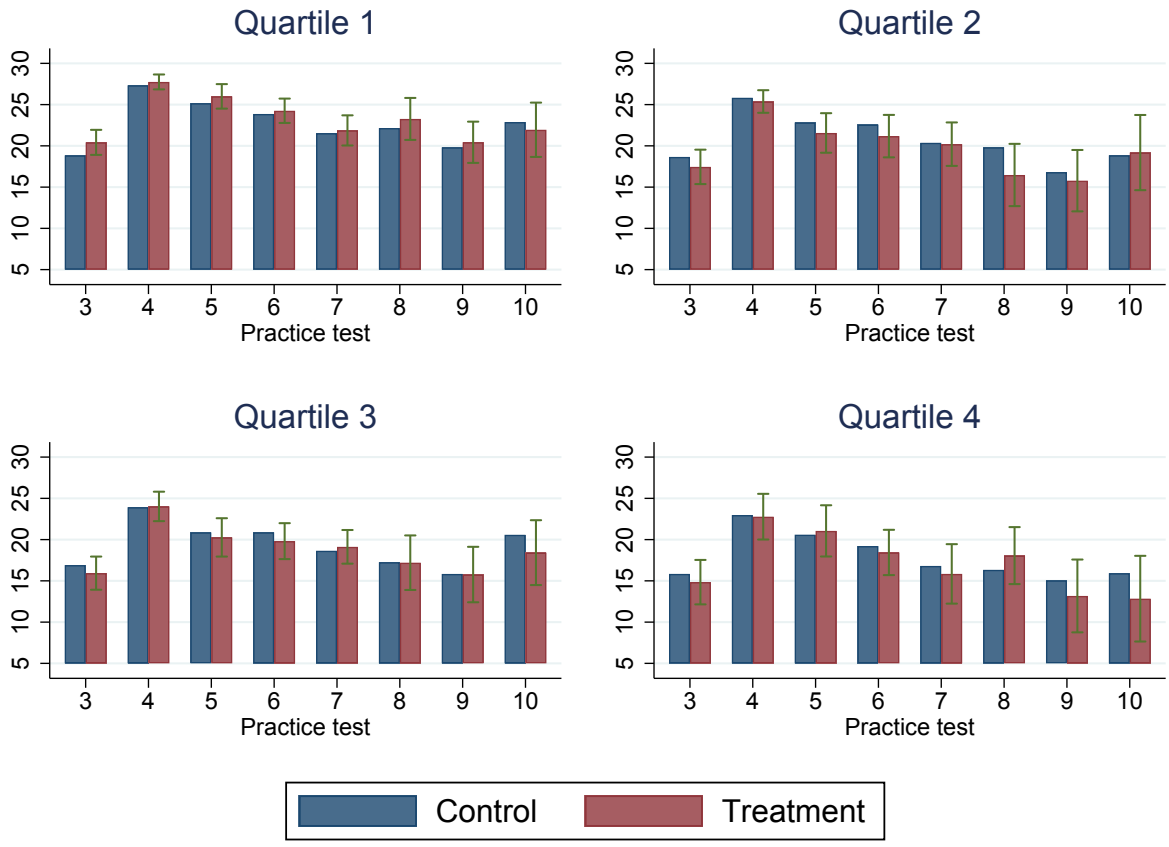


Figure 13: Practice tests: Number of correct answers in reading by week and quartile

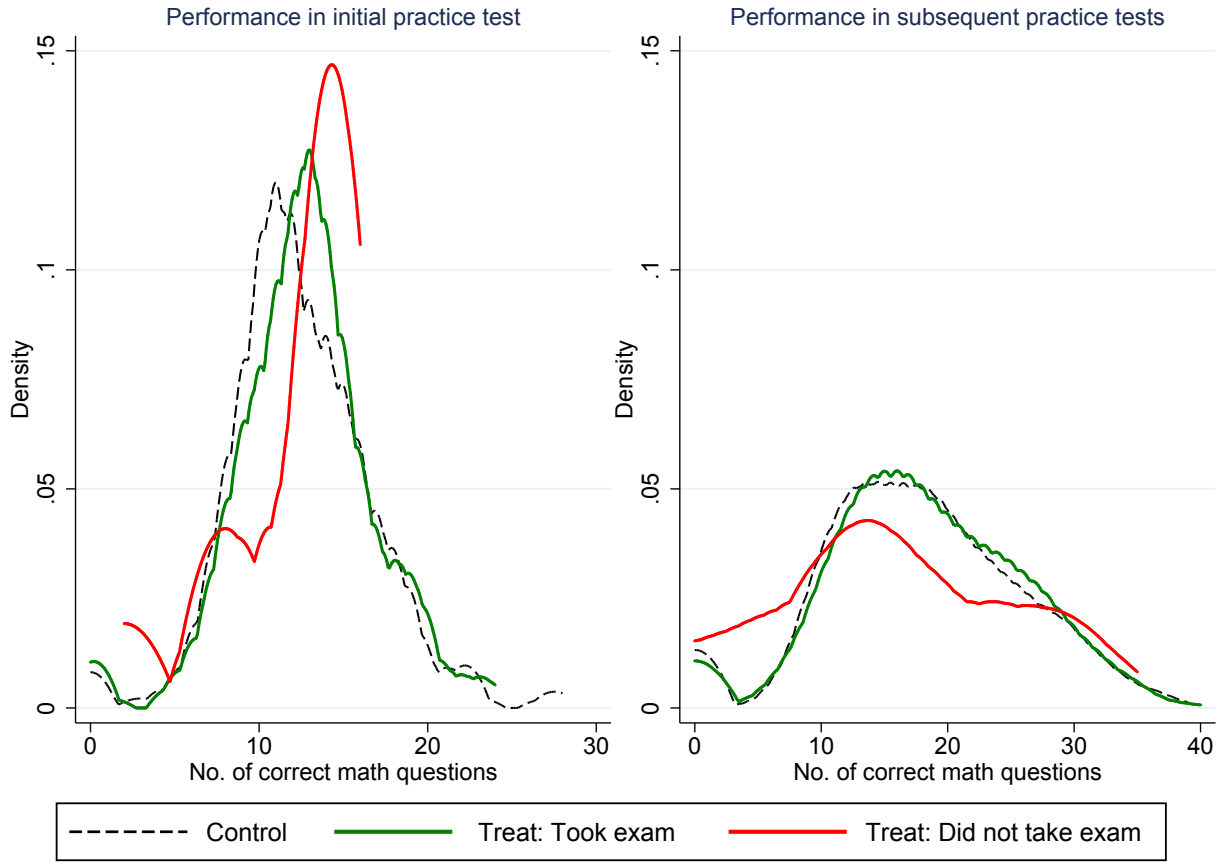


Figure 14: Bottom quartile: Number of correct answers in practice tests by treatment and by decision to take exam

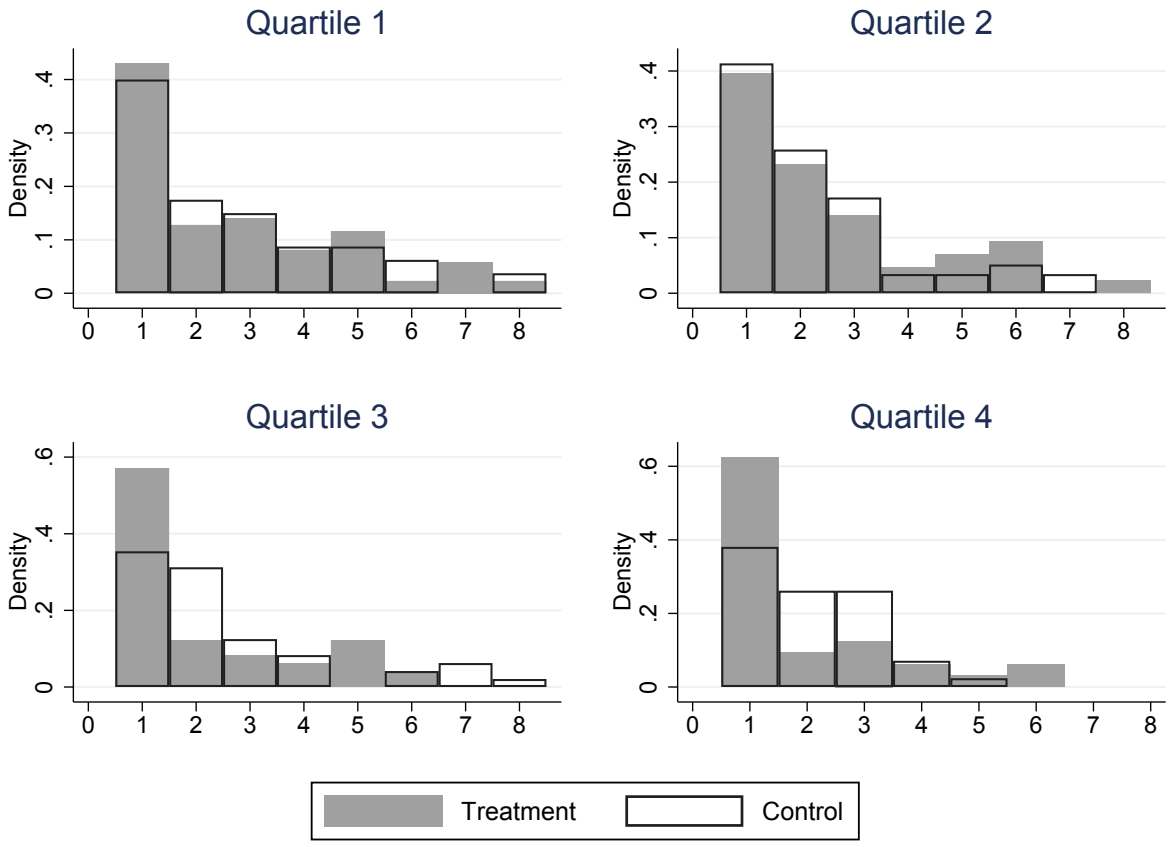


Figure 15: Number of times students check performance report by quartile

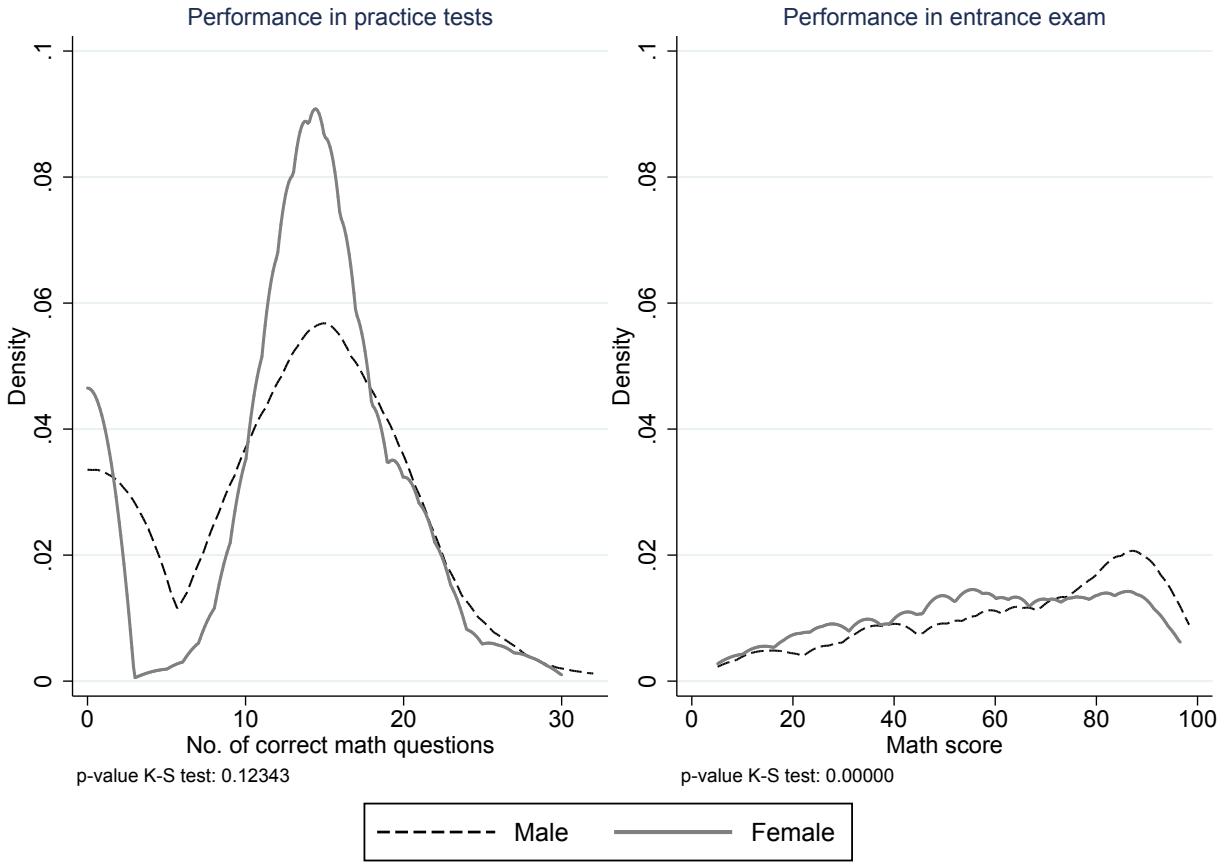


Figure 16: Performance in last parctice test and entrance exam by gender

8 Tables

Table 1: Balance of baseline characteristics

	Control	**Treatment**	**P-value (T-C)**	**No. obs**
Stratification variables				
Female	0.613	0.600	0.780	440
Q1 = top in initial practice test	0.351	0.410	0.208	438
Q2 in initial practice test	0.254	0.204	0.218	438
Q3 in initial practice test	0.211	0.234	0.567	438
Q4 = bottom in initial practice test	0.184	0.152	0.374	438
Previously taken entrance exam	0.795	0.810	0.699	439
AM course	0.426	0.414	0.803	440
PM course	0.357	0.372	0.746	440
Integrated UdeA - UNAL	0.043	0.042	0.975	440
Pre-medicine	0.148	0.148	0.995	440
Weekend course	0.026	0.024	0.879	440
Demographic variables				
Age	17.858	17.383	0.027	434
Single	0.982	0.985	0.782	433
Student	0.717	0.760	0.312	434
Disability	0.018	0.029	0.445	434
Underrepresented minority	0.164	0.101	0.053	434
Urban	0.889	0.903	0.621	434
Residential strata	2.518	2.596	0.432	434
SISBEN score (poverty index)	23.774	24.026	0.919	434
Academic performance variables (initial practice test)				
Math no. correct (out of 40)	12.432	12.671	0.544	439
Reading no. correct (out of 40)	19.808	20.486	0.280	439
Avg. practice test score in classroom	37.413	37.489	0.764	440
P-value of F-test of joint significance			0.331	439

Notes: Figures under the control and treatment column headings are the means of the baseline covariates on the left for treated and control students who checked the experimental performance report at least once during the course of the study. Column 4 shows the p-values of the differences in means between treated and control students. The joint orthogonality test does not include stratification variables. Integrated UdeA-UNAL are students who are taking a preparation course for two college entrance exams for admission to two different public universities. Average practice test score in classroom is the mean score in the initial practice test for all students in the same classroom as participants in the study.

Table 2: Sampling frame and attrition

	Q1 = top	**Q2**	**Q3**	**Q4 = bottom**	**All**
Panel A. Students who consented participation					
Assigned to control	145	130	132	105	512
Assigned to treatment	147	126	132	107	512
TOTAL	292	256	264	212	1,024
Fraction of those consenting	28.5%	25.0%	25.8%	20.7%	100%
Panel B. Students who checked at least one experimental performance report					
Assigned to control	80	58	48	42	228
Assigned to treatment	86	43	49	32	210
TOTAL	166	101	97	74	438
Fraction of all participants	37.9%	23.1%	22.1%	16.9%	100%
Total Panel B / Total Panel A	56.8%	39.5%	36.7%	34.9%	42.8%
Panel C. Statistics on report checking (conditional on checking at least one report)					
Average (out of 8)	2.70	2.42	2.35	2.04	2.45
Standard deviation	1.96	1.73	1.77	1.29	1.77
Minimum	1	1	1	1	1
Maximum	8	8	8	6	8
Average seconds spent in report	41.01	34.06	41.32	36.69	39.15

Notes: Quartiles are calculated based on scores in the initial practice test of all students at the test preparation institution, not only participants. Attrition from the moment students consented to participate to the rounds in which they checked performance reports is detailed in Panels A and B. The second source of attrition is in Panel C. Most of the students who checked the performance report at least once did not check the 8 reports but 2.5 on average, and spent about and average of 40 seconds checking them.

Table 3: Balance of characteristics among attitors

	Control	**Treatment**	**P-value (T-C)**	**No. obs**
Stratification variables				
Female	0.553	0.575	0.592	605
Q1 = top in initial practice test	0.234	0.203	0.363	605
Q2 in initial practice test	0.251	0.278	0.459	605
Q3 in initial practice test	0.295	0.274	0.573	605
Q4 = bottom in initial practice test	0.220	0.245	0.471	605
Previously taken entrance exam	0.797	0.793	0.910	604
AM course	0.447	0.461	0.733	605
PM course	0.237	0.242	0.894	605
Integrated UdeA - UNAL	0.058	0.062	0.849	605
Pre-medicine	0.061	0.064	0.859	605
Weekend course	0.197	0.171	0.417	605
Demographic variables				
Age	18.154	18.117	0.881	569
Single	0.960	0.966	0.719	569
Student	0.718	0.733	0.681	570
Disability	0.014	0.024	0.407	569
Underrepresented minority	0.119	0.123	0.892	570
Urban	0.877	0.891	0.615	570
Residential strata	2.545	2.519	0.769	570
SISBEN score (poverty index)	23.288	24.034	0.730	567
Academic performance variables (initial practice test)				
Math no. correct (out of 40)	11.522	11.484	0.901	605
Reading no. correct (out of 40)	18.864	18.609	0.617	605
Avg. practice test score in classroom	37.760	38.014	0.238	605

Notes: Figures under the control and treatment column headings are the means of the baseline covariates on the left for student who were initially assigned to treatment or control groups but did not check the experimental performance report at least once during the course of the study. Column 4 shows the p-values of the differences in means between treated and control students. The joint orthogonality test does not include stratification variables. Integrated UdeA-UNAL are students who are taking a preparation course for two college entrance exams for admission to two different public universities. Average practice test score in classroom is the mean score in the initial practice test for all students in the same classroom as participants in the study.

Table 4: Persistence of quartile in initial practice test - reading

	Proportion of practice tests in reading quartile:			
	Q1=top	Q2	Q3	Q4=bottom
Q1 = top	0.089**	-0.056**	-0.043*	0.010
	(0.043)	(0.027)	(0.023)	(0.020)
Mean control	0.489	0.279	0.152	0.080
Q2	-0.071	0.015	0.054	0.002
	(0.055)	(0.037)	(0.035)	(0.031)
Mean control	0.364	0.270	0.217	0.149
Q3	-0.018	0.032	-0.004	-0.010
	(0.040)	(0.039)	(0.041)	(0.038)
Mean control	0.193	0.260	0.311	0.236
Q4 = bottom	-0.035	0.036	-0.012	0.011
	(0.042)	(0.042)	(0.041)	(0.059)
Mean control	0.151	0.241	0.313	0.295
N	3515	3515	3515	3515
N_clust	438	438	438	438

Notes: Each column shows the treatment effect in equation 2 by the quartile in the initial practice test in column 1. The outcomes in the columns are the proportion of times the students were placed in Q1, Q2, Q3 and Q4 after for practice tests 3 to 10. For example, on average, control students who were classified in quartile 1 in the initial practice test are classified in the same quartile in about 49% of the subsequent practice tests. Treated students are 8.9 pp pmore likely to be classified in Q1 if they were in Q1 in the initial practice test. Standard errors are clustered at the individual level. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 5: Persistence of quartile in initial practice test - math

	Proportion of times in math quartile:			
	Q1=top	Q2	Q3	Q4=bottom
Q1 = top	0.064	-0.021	-0.032	-0.011
	(0.054)	(0.034)	(0.025)	(0.021)
Mean control	0.528	0.244	0.135	0.093
Q2	-0.068	0.028	0.049	-0.009
	(0.060)	(0.043)	(0.036)	(0.041)
Mean control	0.355	0.307	0.167	0.171
Q3	0.033	0.021	0.002	-0.056
	(0.057)	(0.040)	(0.039)	(0.047)
Mean control	0.188	0.279	0.290	0.244
Q4 = bottom	-0.068	-0.034	0.045	0.058
	(0.044)	(0.052)	(0.043)	(0.066)
Mean control	0.151	0.289	0.268	0.292
N	3517	3517	3517	3517
N_clust	438	438	438	438

Notes: Each column shows the treatment effect in equation 2 by the quartile in the initial practice test in column 1. The outcomes in the columns are the proportion of times the students were placed in Q1, Q2, Q3 and Q4 after for practice tests 3 to 10. For example, on average, control students who were classified in quartile 1 in the initial practice test are classified in the same quartile in about 53% of the subsequent practice tests. Treated students are 6.4 pp pmore likely to be classified in Q1 if they were in Q1 in the initial practice test. Standard errors are clustered at the individual level. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 6: Effect of relative performance feedback on academic investments and practice test scores

	Takes practice tests	**Math study hours**	**Reading study hours**	**Math correct answers**	**Reading correct answers**
Q1 = top	0.011	0.819	0.278	0.696	0.490
	(0.011)	(0.594)	(0.569)	(0.736)	(0.518)
Mean control	0.953	5.018	4.449	21.688	22.856
Q2	0.010	-0.791	-0.114	-1.004	-1.290
	(0.019)	(0.856)	(0.792)	(0.875)	(0.793)
Mean control	0.926	6.179	5.348	18.640	20.831
Q3	0.011	-0.580	-0.291	0.391	-0.593
	(0.019)	(0.806)	(0.745)	(0.881)	(0.688)
Mean control	0.931	5.140	4.455	16.285	19.231
Q4 = bottom	-0.052***	-2.011*	-1.537*	-1.717*	-1.279
	(0.019)	(1.107)	(0.871)	(1.020)	(1.047)
Mean control	0.956	6.303	5.236	15.120	17.557
N	3645	2289	2285	3442	3442
N_clust	438	425	425	438	438

Notes: Each point estimate is the treatment effect on the outcome in the column heading within the quartile of the initial practice test (column 1). Standard errors clustered at the individual level in parenthesis. Outcomes are, from left to right: Fraction of practice tests taken out of 8, average number of self-reported weekly study hours over 8 practice tests in math and reading, and average number of correct answers in math and reading over 8 practice tests. For reference, the mean of the control group in the quartile is reported below the standar error. Controls in this regression include randomization strata and the baseline covariates presented in the balance table. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 7: Effect of relative performance feedback on taking entrance exam and exam performance

	Did not take exam	**Never registered**	**Math score**	**Reading score**	**Total score**
Q1 = top	0.056**	0.059**	1.632	-2.561	-0.375
	(0.025)	(0.025)	(3.176)	(2.637)	(2.325)
Mean control	0.000	0.000	70.888	73.867	72.266
Q2	0.042	-0.000	0.807	-1.725	-0.351
	(0.052)	(0.044)	(4.766)	(4.539)	(3.450)
Mean control	0.052	0.052	60.644	63.163	61.849
Q3	-0.016	-0.016	1.052	-7.674	-5.221
	(0.024)	(0.024)	(4.955)	(4.950)	(4.196)
Mean control	0.021	0.021	50.553	53.538	53.319
Q4 = bottom	0.106*	0.104*	-0.653	3.736	1.581
	(0.057)	(0.056)	(5.862)	(6.120)	(4.974)
Mean control	0.000	0.000	42.377	46.339	44.360
N	438	438	421	421	421

Notes: Each point estimate is the treatment effect on the outcome in the column heading within the quartile of the initial practice test (column 1). Robust standard errors in parenthesis. Outcomes are, from left to right: indicator for not taking the college entrance exam in the first admission cycle, indicator for not taking the exam in the first cycle and not registering for the exam in the following cycle, and exam scores in math, reading and total (only for those who took the exam). For reference, the mean of the control group in the quartile is reported below the standard error. Controls in this regression include randomization strata and the baseline covariates presented in the balance table. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 8: Effect of relative performance feedback on admission and cutoff scores of majors chosen

	Admitted to first choice	**Cutoff score first choice**	**Admitted to second choice**
Q1 = top	-0.076	0.899	0.021
	(0.071)	(1.587)	(0.030)
Mean control	0.313	80.267	0.025
Q2	0.135	-1.927	-0.045
	(0.084)	(2.062)	(0.031)
Mean control	0.130	79.484	0.037
Q3	0.004	0.041	-0.005
	(0.050)	(1.963)	(0.040)
Mean control	0.043	78.918	0.043
Q4 = bottom	0.003	-1.986	-0.014
	(0.068)	(2.417)	(0.020)
Mean control	0.071	79.439	0.024
N	421	421	421

Notes: Each point estimate is the treatment effect on the outcome in the column heading within the quartile of the initial practice test (column 1). Robust standard errors in parenthesis. Outcomes are, from left to right: indicator for not taking the college entrance exam in the first admission cycle, indicator for not taking the exam in the first cycle and not registering for the exam in the following cycle, and exam scores in math, reading and total (only for those who took the exam). For reference, the mean of the control group in the quartile is reported below the standard error. Controls in this regression include randomization strata and the baseline covariates presented in the balance table. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 9: Effect of relative performance feedback on prior beliefs - reading

	Correct	**Overplace**	**Underplace**
Q1 = top	0.101**	-0.062*	-0.006
	(0.039)	(0.034)	(0.036)
Mean control	0.417	0.230	0.274
Q2	0.007	0.117**	-0.075
	(0.041)	(0.046)	(0.050)
Mean control	0.295	0.222	0.345
Q3	0.004	-0.008	-0.032
	(0.047)	(0.047)	(0.037)
Mean control	0.309	0.320	0.206
Q4 = bottom	0.044	-0.140*	0.048
	(0.060)	(0.071)	(0.059)
Mean control	0.288	0.365	0.160
N	2551	2551	2551
N_clust	433	433	433

Notes: Each column shows coefficients of a regression of a dummy indicating whether the student was correct, underplaced or overplaced their prior belief on a treatment dummy and randomization strata. Each panel indicates the quartile in which students were in the intial practice test. For example, column 1 in Panel A shows that students who were in the top quartile in the initial practice test had a correct prior in 41-51% of all subsequent practice tests. Overplace (underplace) means that the student assigned the highest probability to a higher (lower) quartile than her score was in. Standard errors are clustered at the individual level. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 10: Effect of relative performance feedback on prior beliefs - math

	Correct	**Overplace**	**Underplace**
Q1 = top	0.124***	-0.070**	-0.016
	(0.042)	(0.031)	(0.043)
Mean control	0.395	0.179	0.341
Q2	0.003	0.078*	-0.013
	(0.053)	(0.047)	(0.058)
Mean control	0.363	0.154	0.323
Q3	0.014	-0.047	-0.033
	(0.051)	(0.055)	(0.045)
Mean control	0.328	0.280	0.232
Q4 = bottom	0.051	-0.154**	0.001
	(0.066)	(0.070)	(0.052)
Mean control	0.301	0.365	0.187
N	2551	2551	2551
N_clust	433	433	433

Notes: Each column shows coefficients of a regression of a dummy indicating whether the student was correct, underplaced or overplaced their prior belief on a treatment dummy and randomization strata. Each panel indicates the quartile in which students were in the intial practice test. For example, column 1 in Panel A shows that students who were in the top quartile in the initial practice test had a correct prior in 41-51% of all subsequent practice tests. Overplace (underplace) means that the student assigned the highest probability to a higher (lower) quartile than her score was in. Standard errors are clustered at the individual level. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 11: Effect of relative performance feedback on academic investments by gender

	Takes practice tests		**Math study hours**		**Reading study hours**	
	Female	Male	Female	Male	Female	Male
Q1 = top	0.018	0.002	0.848	0.696	0.077	0.495
	(0.014)	(0.017)	(0.829)	(0.801)	(0.805)	(0.720)
Mean control	0.956	0.946	5.312	4.456	4.763	3.852
DiD F vs. M	0.015		0.151		-0.418	
	(0.022)		(1.149)		(1.072)	
Q2	-0.026	0.069**	-0.760	-0.781	-0.063	-0.121
	(0.024)	(0.027)	(1.175)	(1.173)	(1.109)	(1.033)
Mean control	0.950	0.890	6.731	5.153	5.923	4.286
DiD F vs. M	-0.095***		0.020		0.058	
	(0.036)		(1.667)		(1.520)	
Q3	0.019	-0.009	0.092	-2.169	0.116	-1.254
	(0.021)	(0.037)	(0.997)	(1.315)	(0.975)	(0.991)
Mean control	0.929	0.934	5.204	4.985	4.713	3.824
DiD F vs. M	0.028		2.261		1.370	
	(0.044)		(1.672)		(1.413)	
Q4 = bottom	-0.038	-0.067**	-0.737	-3.727**	-1.476	-1.504
	(0.025)	(0.030)	(1.516)	(1.520)	(1.251)	(1.198)
Mean control	0.949	0.963	6.010	6.611	5.390	5.074
DiD F vs. M	0.029		2.989		0.028	
	(0.039)		(2.149)		(1.739)	

Notes: Each point estimate is the treatment effect on the outcome in the column heading for females and males within the quartile labeled in each panel. For reference, the mean of the control group in the quartile is reported below the standar error. The DiD coefficient shows the difference-in-differences coefficient between females and males. Robust standard errors in parenthesis. Controls in the regression include random strata, age, poverty index, marital status, students at another institution, underrepresented minority status, residential strata, urban, scores obtained in math and reading in initial practice test, average score in initial practice test in student's classroom. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 12: Effect of relative performance feedback on performance in practice tests by gender

	Math correct answers		**Reading correct answers**	
	Female	Male	Female	Male
Q1 = top	0.750	0.612	0.617	0.345
	(0.945)	(1.198)	(0.693)	(0.757)
Mean control	21.383	22.278	23.002	22.571
DiD F vs. M	0.139		0.272	
	(1.525)		(1.023)	
Q2	-3.151***	1.015	-2.710***	1.060
	(1.030)	(1.390)	(1.016)	(1.168)
Mean control	18.438	19.348	21.694	19.410
DiD F vs. M	-4.165**		-3.770**	
	(1.733)		(1.550)	
Q3	0.717	-0.349	-0.217	-1.323
	(1.042)	(1.634)	(0.776)	(1.384)
Mean control	15.882	17.123	18.852	20.018
DiD F vs. M	1.067		1.106	
	(1.944)		(1.608)	
Q4 = bottom	-2.026	-1.331	-0.560	-2.164
	(1.275)	(1.692)	(1.509)	(1.416)
Mean control	14.449	15.829	17.251	17.880
DiD F vs. M	-0.695		1.604	
	(2.127)		(2.072)	

Notes: Each point estimate is the treatment effect on the outcome in the column heading for females and males within the quartile labeled in each panel. For reference, the mean of the control group in the quartile is reported below the standar error. The DiD coefficient shows the difference-in-differences coefficient between females and males. Robust standard errors in parenthesis. Controls in the regression include random strata, age, poverty index, marital status, students at another institution, underrepresented minority status, residential strata, urban, scores obtained in math and reading in initial practice test, average score in initial practice test in student's classroom. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 13: Effect of relative performance feedback on exam taking and cutoff score of major first choice

	Did not take exam		**Never registered**		**Cutoff score first choice**	
	Female	Male	Female	Male	Female	Male
Q1 = top	0.072***	0.035	0.075**	0.037	-0.540	3.029
	(0.036)	(0.029)	(0.036)	(0.029)	(1.967)	(2.685)
Mean control	0.000	0.000	0.000	0.000	81.231	78.571
	0.048					
DiD F vs. M	0.037		0.038		-3.569	
	(0.045)		(0.045)		(3.324)	
Q2	0.069	-0.004	0.042	-0.067	-3.140	0.175
	(0.051)	(0.105)	(0.044)	(0.088)	(2.749)	(3.153)
Mean control	0.000	0.130	0.000	0.130	80.085	78.377
DiD F vs. M	0.073		0.109		-3.315	
	(0.116)		(0.099)		(4.193)	
Q3	-0.018	-0.021	-0.020	-0.016	-0.784	1.728
	(0.035)	(0.023)	(0.035)	(0.023)	(2.402)	(3.591)
Mean control	0.032	0.000	0.032	0.000	79.674	77.584
DiD F vs. M	0.003		-0.004		-2.512	
	(0.040)		(0.040)		(4.367)	
Q4 = bottom	0.119	0.091	0.123	0.081	-0.237	-4.103
	(0.081)	(0.075)	(0.080)	(0.072)	(3.350)	(3.706)
Mean control	0.000	0.000	0.000	0.000	79.323	79.567
DiD F vs. M	0.029		0.042		3.867	
	(0.109)		(0.106)		(5.176)	

Notes: Each point estimate is the treatment effect on the outcome in the column heading for females and males within the quartile labeled in each panel. For reference, the mean of the control group in the quartile is reported below the standar error. The DiD coefficient shows the difference-in-differences coefficient between females and males. Robust standard errors in parenthesis. Controls in the regression include random strata, age, poverty index, marital status, students at another institution, underrepresented minority status, residential strata, urban, scores obtained in math and reading in initial practice test, average score in initial practice test in student's classroom. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 14: Effect of feedback on prior beliefs by gender - reading

	Correct		**Overplace**		**Underplace**	
	Female	Male	Female	Male	Female	Male
Q1 = top	0.155***	0.040	-0.074*	-0.046	-0.063	0.064
	(0.050)	(0.060)	(0.043)	(0.055)	(0.049)	(0.052)
Mean control	0.389	0.458	0.229	0.217	0.325	0.205
DiD F vs. M	0.115		-0.028		-0.127*	
	(0.078)		(0.069)		(0.071)	
Q2	0.055	-0.076	0.115*	0.125*	-0.120*	0.003
	(0.056)	(0.061)	(0.066)	(0.064)	(0.072)	(0.061)
Mean control	0.246	0.398	0.208	0.230	0.391	0.265
DiD F vs. M	0.131		-0.010		-0.123	
	(0.084)		(0.093)		(0.095)	
Q3	-0.023	0.086	-0.011	-0.029	-0.025	-0.005
	(0.054)	(0.087)	(0.056)	(0.091)	(0.046)	(0.060)
Mean control	0.281	0.372	0.335	0.295	0.205	0.192
DiD F vs. M	-0.108		0.018		-0.020	
	(0.102)		(0.108)		(0.077)	
Q4 = bottom	0.004	0.040	-0.072	-0.188*	0.026	0.063
	(0.075)	(0.094)	(0.096)	(0.103)	(0.083)	(0.082)
Mean control	0.277	0.314	0.339	0.373	0.179	0.147
DiD F vs. M	-0.036		0.117		-0.036	
	(0.121)		(0.139)		(0.113)	

Notes: Each point estimate is the treatment effect on the outcome in the column heading for females and males within the quartile labeled in each panel. For reference, the mean of the control group in the quartile is reported below the standar error. The DiD coefficient shows the difference-in-differences coefficient between females and males. Robust standard errors in parenthesis. Controls in the regression include random strata, age, poverty index, marital status, students at another institution, underrepresented minority status, residential strata, urban, scores obtained in math and reading in initial practice test, average score in initial practice test in student's classroom. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

A Variable coding

Table A.1: Coding of beliefs elicitation token allocation

Possible token allocations:				**Variable coding if actual quartile is:**			
Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
6	6	0	0	Correct	Correct	Overplace	Overplace
6	0	6	0	Inconsistent	Inconsistent	Inconsistent	Inconsistent
6	0	0	6	Inconsistent	Inconsistent	Inconsistent	Inconsistent
0	6	6	0	Underplace	Correct	Correct	Overplace
0	0	6	6	Underplace	Underplace	Correct	Correct
0	6	0	6	Underestimate	Inconsistent	Inconsistent	Inconsistent
5	5	x	x	Correct	Correct	Overplace	Overplace
x	5	5	x	Underplace	Correct	Correct	Overplace
x	x	5	5	Underplace	Underplace	Correct	Correct
5	x	x	5	Inconsistent	Inconsistent	Inconsistent	Inconsistent
5	x	5	x	Inconsistent	Inconsistent	Inconsistent	Inconsistent
x	5	x	5	Underestimate	Inconsistent	Inconsistent	Inconsistent
4	4	4	0	Correct	Correct	Correct	Overplace
0	4	4	4	Underplace	Correct	Correct	Correct
4	0	4	4	Inconsistent	Inconsistent	Inconsistent	Inconsistent
4	4	0	4	Inconsistent	Inconsistent	Inconsistent	Inconsistent
4	4	x	x	Correct	Correct	Overplace	Overplace
x	x	4	4	Underplace	Underplace	Correct	Correct
4	x	x	4	Inconsistent	Inconsistent	Inconsistent	Inconsistent
x	4	4	x	Underplace	Correct	Correct	Overplace
3	3	3	3	Equal probability	Equal probability	Equal probability	Equal probability

Notes: This table shows how the beliefs variables are coded in cases of repeated numbers. I define five beliefs variables: correct, overplace, underplace, equal probability and inconsistent. Where there is a clear maximum probability allocation, the variable correct takes the value of 1 if the student's scores lies in that quartile. Overplace equals one if the quartile with the largest probability allocation is less than the actual quartile (e.g., if the student is in Q3 and she allocates most probability to Q1 or Q2). Underplace equals to 1 if the largest probability is in a quartile larger than the actual quartile (e.g., if the student is in Q3 and she allocates most probability to Q4). Inconsistent is when the largest probabilities are assigned to non-adjacent quartiles.

B Experimental protocol

B.1 Lab-in-the-field belief elicitation

Students take a weekly practice test that can be administered on paper or online. Immediately after the practice test, they receive a paper or online survey with the questions below. A few days later, when they receive the performance report, the answer the belief elicitation (The quartiles game) once again. As explained in the main text, this is intended to understand how students form posterior beliefs given the information provided: absolute scores in the case of control students, and above-/below- median signal in the case of treated students.

In case of questions students had access to a FAQ sheet that they could access from the website containing the survey. They were also provided with a phone number and email so they could ask a specific question.

To determine how much cash they would win if they were selected as one of the weekly winners, a computerized dice was thrown in the last page of the performance report. See below for the instructions students had to follow.

Beliefs survey instructions

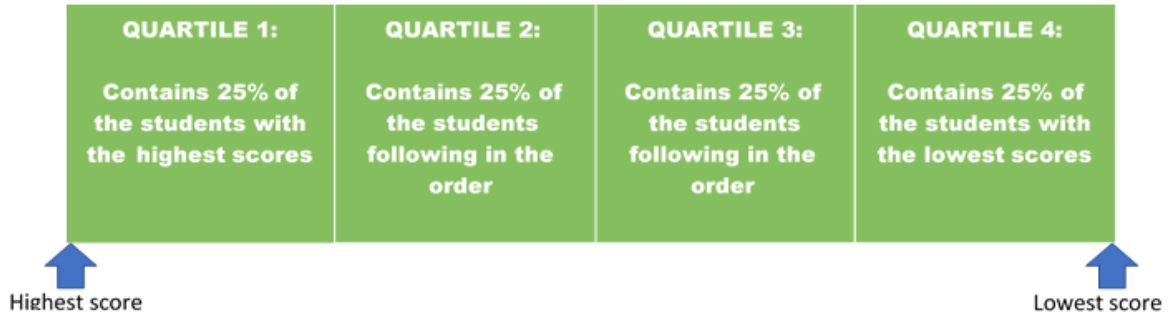
The following questions are related to today's practice test. Based on your answers, you may win on of the cash prizes. The winners will be chosen based on the distance of their ID's last four digits and the Loteria de la Cruz Roja jackpot that plays next week. The IDs that are closest to the jackpot will receive cash prizes until 300,000 pesos are awarded. Recall that you must complete all surveys to enter the raffle of 6 laptops. Winners will be contacted by email.

Q1: How many questions of the practice test do you think you will answer / have answered correctly? For every correct guess, you will receive 5,000 pesos if you are selected in the draw.
Mathematical logical reasoning: _____ correct out of 40 questions
Reading competency: _____ correct out of 40 questions

Q2: The Quartiles Game
Imagine that you enter a casino and are given 24 tokens to play. You choose to bet them in the "Quartiles Game."

The bet you will make is regarding how you think you performed in the practice test of this week. To make your bet, the dealer explains to you that there are 4 sections in the betting table called "quartiles." Each quartile contains a subgroup of the scores of students

who took the practice test ordered from highest to lowest. This shows how the table looks like:



You bet consists in assigning the 24 tokens to the quartiles in the two sections of the practice test (12 tokens to mathematical and logical reasoning and 12 tokens to reading).

Tip: Assigning more tokens to the quartile(s) in which you truly believe your score is will maximize your chances of winning one of the prizes. Remember that no one besides you will see your answers.

For each section of the practice test, please assign 12 tokens to the quartiles. If you think it is unlikely that your score will be in one or more of the quartiles, please assign zero tokens to those quartiles. Make sure that the sum of your allocations is equal to 12 in each column:

	Mathematical Logical Reasoning	**Reading Competency**
Bet to quartile 1 (group with highest scores)	__ tokens	__ tokens
Bet to quartile 2	__ tokens	__ tokens
Bet to quartile 3	__ tokens	__ tokens
Bet to quartile 4 (group with lowest scores)	__ tokens	__ tokens
Sum of tokens	__ tokens	__ tokens

To determine if you win your bet, the dealer will look at the quartile your score is in and roll a 12-sided die like this:

If the result of the die roll is equal or below your bet in the quartile your score is in, you will win 20,000 pesos in case you are selected in the draw. For example, you bet 8 tokens and the result of the die roll is 6, you will be eligible for a prize.

If the die roll result is above your bet, the dealer will roll the die again. You will win 20,000 pesos if the new result is below the result of the first die roll and you are selected in the draw. For example, if the result of the first roll is 6 and the result of the second is 3.

Q3: 3. Please rate the level of difficulty of each section of the practice test, where 1 is extremely easy and 7 extremely hard.

Q4: 4. Approximately, how many hours did you study last week for each section of the practice test? Include time you dedicated to solve practice questions and review materials. Do NOT include class time and homework. Choose zero if you did not study for that section of the test last week.

Options: integers from 0 to 8+

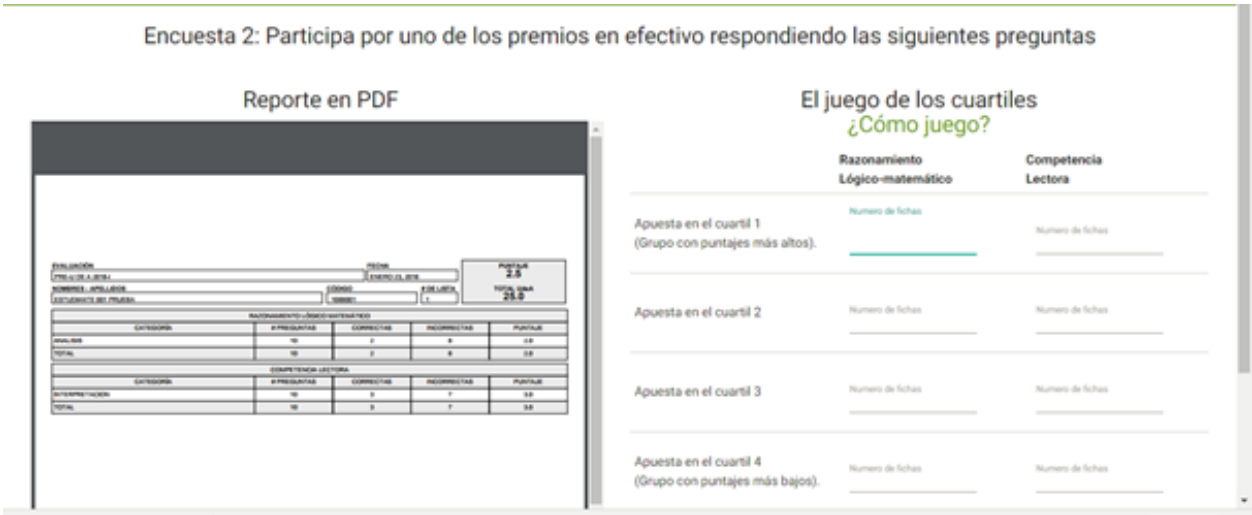
Q5: 5. Based on how you feel today, how confident do you feel that you will gain admission to Universidad de Antioquia?

B.2 Feedback provision

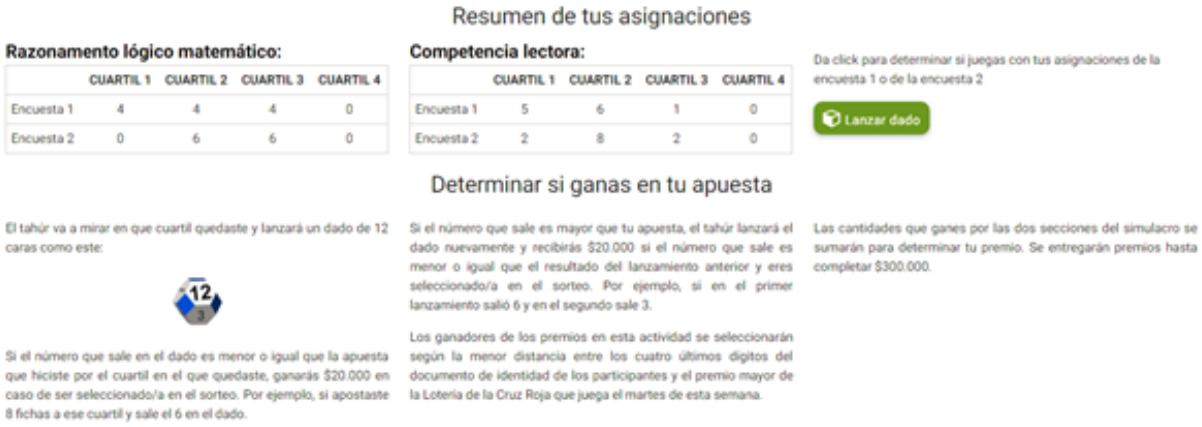
A. Students in the control group

The following screenshots show the different steps students went through while checking the online performance report (in Spanish with comments in English).

Step 1: PDF with absolute scores on the left and The Quartiles Game on the right.



Step 2: On top, summary of token allocations in both surveys, right after the practice test and after checking the performance report. Below, instructions on how to win in The Quartiles Game.



Step 3: A survey (between survey 1 and 2) is selected to play The Quartiles Game. Students throw the dice separately for math and reading.

Step 4: Students are informed how much they will earn if selected in the weekly draw.

B. Students in the treatment group

Treated students follow the same steps with an additional step after Step 1:

Step 1a: Students see relative performance feedback report as discussed in the main text.

C Additional figures and tables

Major	Applicants	Admitted	Admission rate (%)
Surgical instrument processing	1,855	36	0.0194
Nursing	2,590	55	0.0212
Psychology	2,192	47	0.0214
Medicine	6,277	139	0.0221
Nutrition and dietetics	1,632	37	0.0227
Audiovisual and multimedia communication	797	22	0.0276
Business administration (Health)	1,162	35	0.0301
Veterinary medicine	1,968	60	0.0305
Dentistry	1,659	54	0.0325
Foreign languages	1,132	39	0.0345
Translation	1,100	41	0.0373
Agricultural inputs technology	229	9	0.0393
Social work	1,447	57	0.0394
Journalism	541	24	0.0444
Zootechnics	1,221	56	0.0459
Athletic training	910	42	0.0462
Communications	533	25	0.0469
Business administration	1,803	86	0.0477
Civil engineering	1,557	76	0.0488

Figure C.17: Admission rates for most competitive majors to start a program in the 2018-II semester

Table C.2: Balance of baseline characteristics by quartile

	Q1 = top		**Q2**		**Q3**		**Q4 = bottom**	
	Control	Treat	Control	Treat	Control	Treat	Control	Treat
Startification variables								
Female	0.638	0.570	0.603	0.604	0.646	0.674	0.524	0.563
Previously taken entrance exam	0.850	0.849	0.810	0.790	0.750	0.837	0.714	0.687
AM course	0.513	0.442	0.431	0.442	0.396	0.388	0.310	0.344
PM course	0.287	0.313	0.362	0.372	0.396	0.429	0.405	0.438
Integrated UdeA - UNAL	0.013	0.035	0.034	0.046	0.083	0.020	0.071	0.093
Pre-medicine	0.163	0.187	0.138	0.140	0.104	0.122	0.190	0.093
Weekend course	0.025	0.023	0.034	0.000	0.021	0.041	0.024	0.031
Demographic variables								
Age	17.494	17.186	17.877	17.262	17.870	17.715	18.049	17.563
Single	0.974	0.976	1.000	1.000	1.000	0.980	0.952	1.000
Student	0.696	0.837**	0.750	0.667	0.660	0.715	0.786	0.742
Disability	0.000	0.023	0.036	0.096	.	.	0.048	0.000
Underrepresented minority	0.127	0.059	0.143	0.143	0.191	0.102	0.238	0.161
Urban	0.911	0.895	0.893	0.881	0.851	0.939	0.881	0.903
Residential strata	2.633	2.872	2.482	2.357	2.298	2.531	2.595	2.258
SISBEN score (poverty index)	25.001	26.028	28.115	25.731	19.544	21.809	20.421	19.671
Academic performance variables (initial practice test)								
Math no. correct (out of 40)	15.625	15.349	12.259	12.442	11.063	11.062	8.190	8.250
Reading no. correct (out of 40)	25.550	26.186	20.776	20.954	16.708	17.143	11.262	9.656
Avg. practice test score in classroom	38.052	38.157	37.443	37.720	37.616	36.797	36.127	36.440

Notes: Each column contains the mean of the variable on the left-hand-side in the control and treatment groups by quartile. Asterisks in the treatment mean indicate that the difference in means between treatment and control is significant at the 1% level (***), 5% level (**), or 10% level (*).

Table C.3: Effect of relative performance feedback on posterior beliefs - reading

	Correct	**Overplace**	**Underplace**
Q1 = top	0.148***	-0.037	-0.114**
	(0.055)	(0.035)	(0.052)
Mean control	0.488	0.127	0.338
Q2	0.012	0.125**	-0.039
	(0.067)	(0.059)	(0.062)
Mean control	0.328	0.194	0.328
Q3	0.046	-0.007	-0.048
	(0.065)	(0.076)	(0.062)
Mean control	0.377	0.279	0.246
Q4 = bottom	0.033	-0.176**	-0.007
	(0.094)	(0.087)	(0.081)
Mean control	0.307	0.398	0.227
N	1072	1072	1072
N_clust	438	438	438

Notes: Each column shows coefficients of a regression of a dummy indicating whether the student was correct, underplaced or overplaced their prior belief on a treatment dummy and randomization strata. Each panel indicates the quartile in which students were in the intial practice test. For example, column 1 in Panel A shows that students who were in the top quartile in the initial practice test had a correct prior in 41-51% of all subsequent practice tests. Overplace (underplace) means that the student assigned the highest probability to a higher (lower) quartile than her score was in. Standard errors are clustered at the individual level. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table C.4: Effect of relative performance feedback on posterior beliefs - math

	Correct	**Overplace**	**Underplace**
Q1 = top	0.149***	0.017	-0.154***
	(0.051)	(0.034)	(0.051)
Mean control	0.490	0.096	0.351
Q2	0.040	0.014	0.076
	(0.073)	(0.053)	(0.070)
Mean control	0.422	0.141	0.273
Q3	0.101	-0.055	-0.059
	(0.071)	(0.066)	(0.074)
Mean control	0.391	0.227	0.300
Q4 = bottom	0.123	-0.163*	-0.097
	(0.098)	(0.095)	(0.077)
Mean control	0.289	0.361	0.253
N	1018	1018	1018
N_clust	419	419	419

Notes: Each column shows coefficients of a regression of a dummy indicating whether the student was correct, underplaced or overplaced their prior belief on a treatment dummy and randomization strata. Each panel indicates the quartile in which students were in the intial practice test. For example, column 1 in Panel A shows that students who were in the top quartile in the initial practice test had a correct prior in 41-51% of all subsequent practice tests. Overplace (underplace) means that the student assigned the highest probability to a higher (lower) quartile than her score was in. Standard errors are clustered at the individual level. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table C.5: Effect of relative performance feedback on entrance exam scores by gender

	Math score		**Reading score**		**Total score**	
	Female	Male	Female	Male	Female	Male
Q1 = top	4.716	-2.888	0.869	-7.521*	3.028	-5.326
	(4.136)	(5.077)	(3.432)	(4.271)	(3.003)	(3.800)
Mean control	68.051	75.878	72.151	76.884	69.927	76.380
DiD F vs. M	7.604		8.390		8.354	
	(6.587)		(5.552)		(4.902)	
Q2	-0.834	2.647	-2.927	0.435	-1.800	1.688
	(6.203)	(7.066)	(6.099)	(6.551)	(4.748)	(4.664)
Mean control	54.600	71.777	64.042	61.545	59.322	66.503
DiD F vs. M	-3.481		-3.361		-3.488	
	(9.403)		(8.851)		(6.665)	
Q3	-1.197	4.819	-7.559	-7.558	-6.811	-2.256
	(6.006)	(8.764)	(6.199)	(8.387)	(5.116)	(7.422)
Mean control	49.169	52.994	53.221	54.098	52.694	54.420
DiD F vs. M	-6.016		-0.001		-4.555	
	(10.618)		(10.428)		(9.020)	
Q4 = bottom	-6.260	5.828	4.199	3.098	-1.083	4.600
	(8.022)	(8.566)	(8.081)	(9.838)	(6.743)	(7.752)
Mean control	43.240	41.427	45.983	46.731	44.612	44.082
DiD F vs. M	-12.088		1.101		-5.682	
	(11.647)		(12.976)		(10.365)	

Notes: Each point estimate is the treatment effect on the outcome in the column heading for females and males within the quartile labeled in each panel. For reference, the mean of the control group in the quartile is reported below the standar error. The DiD coefficient shows the difference-in-differences coefficient between females and males. Robust standard errors in parenthesis. Controls in the regression include random strata, age, poverty index, marital status, students at another institution, underrepresented minority status, residential strata, urban, scores obtained in math and reading in initial practice test, average score in initial practice test in student's classroom. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table C.6: Effect of feedback on prior beliefs by gender - math

	Correct		**Overplace**		**Underplace**	
	Female	Male	Female	Male	Female	Male
Q1 = top	0.151***	0.109*	-0.074*	-0.059	-0.047	0.010
	(0.056)	(0.064)	(0.043)	(0.042)	(0.059)	(0.060)
Mean control	0.360	0.440	0.201	0.120	0.369	0.319
DiD F vs. M	0.042		-0.014		-0.057	
	(0.085)		(0.060)		(0.083)	
Q2	-0.075	0.068	0.180*	-0.035	-0.046	0.061
	(0.059)	(0.098)	(0.065)	(0.055)	(0.068)	(0.110)
Mean control	0.329	0.442	0.159	0.124	0.343	0.292
DiD F vs. M	-0.144		0.215**		-0.107	
	(0.115)		(0.085)		(0.130)	
Q3	-0.008	0.081	-0.052	-0.053	-0.045	0.032
	(0.056)	(0.106)	(0.066)	(0.100)	(0.055)	(0.085)
Mean control	0.293	0.410	0.293	0.256	0.255	0.167
DiD F vs. M	-0.089		0.001		-0.077	
	(0.120)		(0.120)		(0.103)	
Q4 = bottom	0.026	0.070	-0.133	-0.196*	0.002	0.006
	(0.090)	(0.091)	(0.097)	(0.101)	(0.069)	(0.079)
Mean control	0.268	0.343	0.384	0.343	0.179	0.196
DiD F vs. M	-0.044		0.064		-0.004	
	(0.125)		(0.141)		(0.104)	

Notes: Each point estimate is the treatment effect on the outcome in the column heading for females and males within the quartile labeled in each panel. For reference, the mean of the control group in the quartile is reported below the standar error. The DiD coefficient shows the difference-in-differences coefficient between females and males. Robust standard errors in parenthesis. Controls in the regression include random strata, age, poverty index, marital status, students at another institution, underrepresented minority status, residential strata, urban, scores obtained in math and reading in initial practice test, average score in initial practice test in student's classroom. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

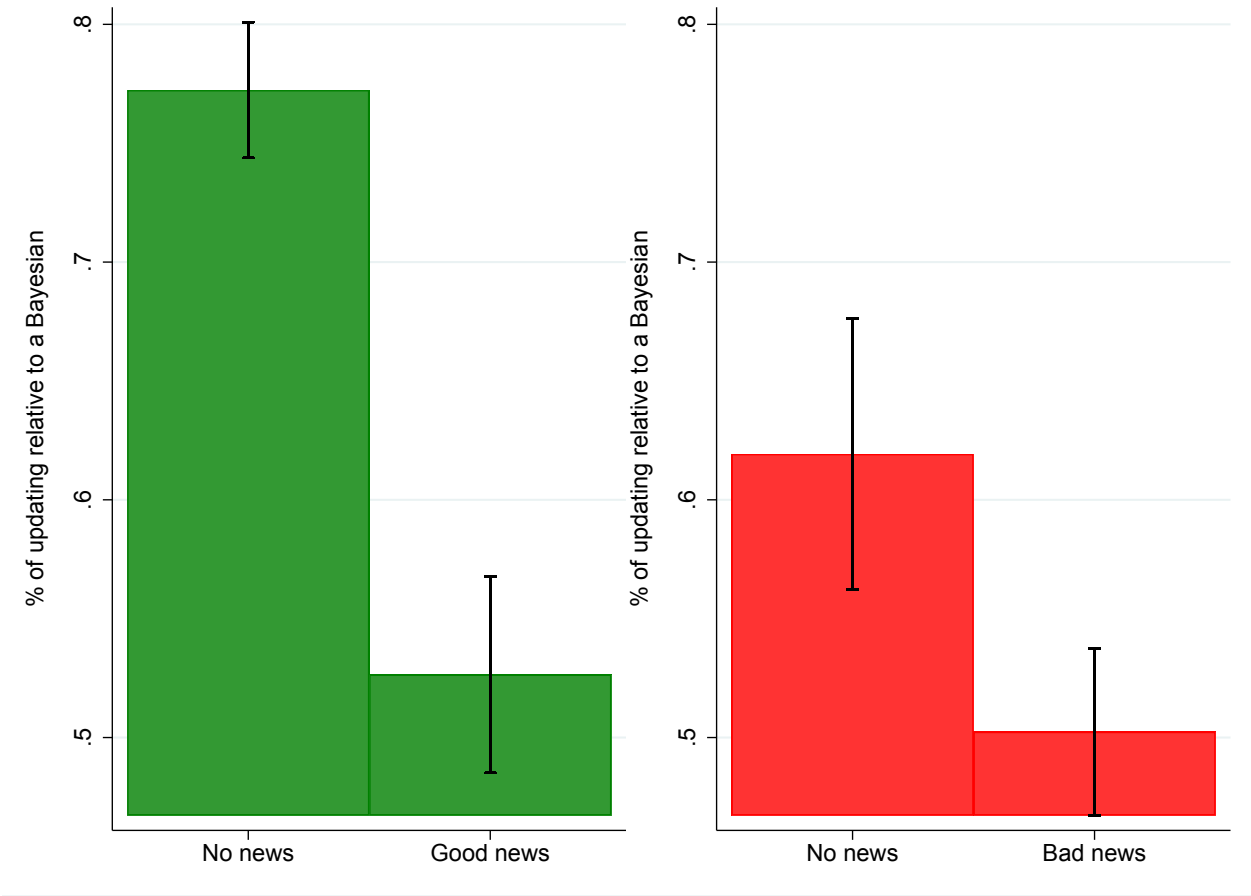


Figure C.18: Confirmatory bias in reading

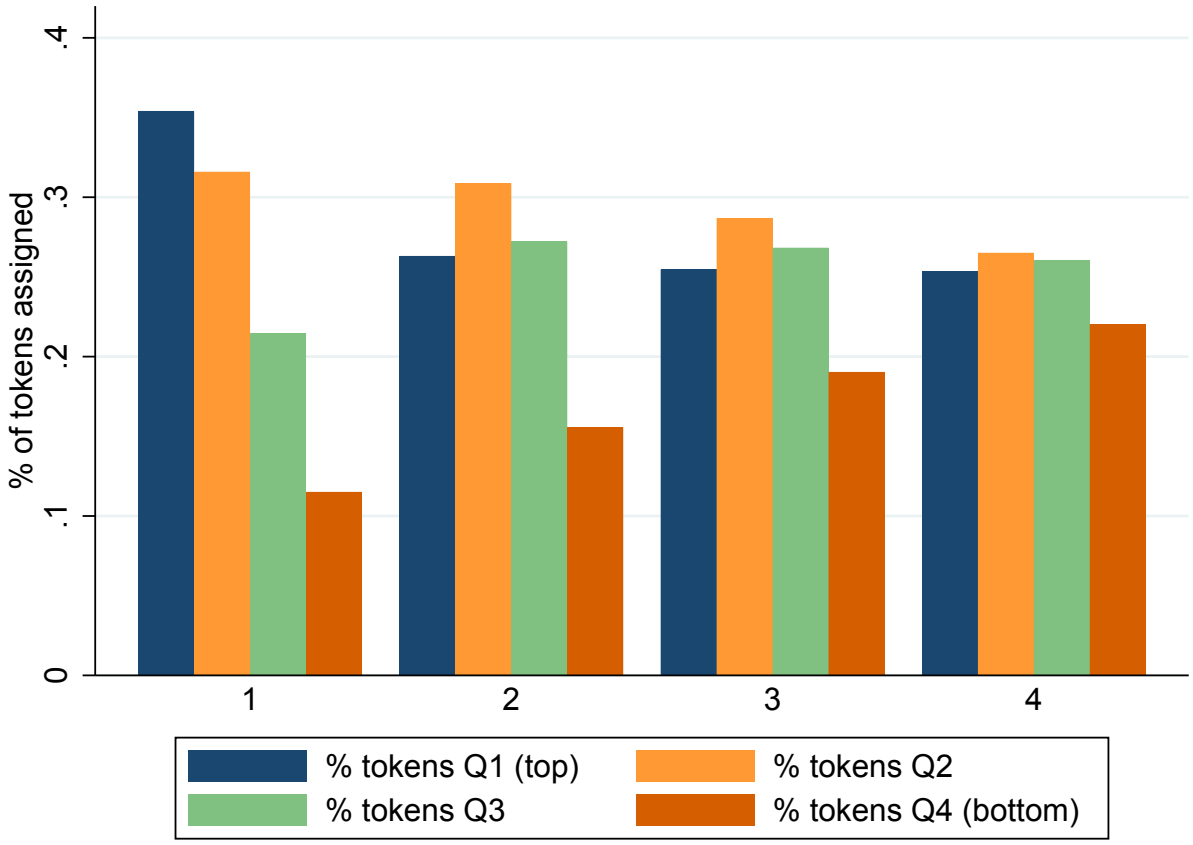


Figure C.19: Token allocation by quartile in initial practice test