How does relative performance feedback affect beliefs and academic decisions? Evidence from a field experiment

Catalina Franco [*]

Most recent version here

January 31, 2019

I conduct field and lab-in-the-field experiments, with students preparing for a college entrance exam, to identify how receiving relative performance feedback affects students' beliefs, performance and academic decisions. I elicit beliefs from all students about relative performance in weekly practice tests and provide feedback to treated students about their actual standing in the score distribution at a test preparation center in Colombia. Combining the panel dataset collected from the experiment with administrative data, I study impacts on: (i) relative performance beliefs, (ii) academic investments, (iii) academic decisions, and (iv) performance. First, feedback makes low-performing students invest less in academic inputs like taking practice tests and study time. Second, I find that high- and low-performing students receiving feedback are less likely to take the entrance exam. Third, heterogeneous effects by gender indicate that women do not change investments but lead the negative effect on exam taking, and are much less likely to gain admission despite similar performance in practice tests. Fourth, beliefs elicited with an incentive compatible task do not match the beliefs revealed by students' actions. Overall, my results shed light on the potential discouragement effects of informational interventions on students with low academic performance.

[*] Job market candidate, Economics Department, University of Michigan, Ann Arbor, cfrancob@umich.edu. Post-doctoral researcher, Universidad del Rosario, catalina.francob@urosario.edu.co. I thank my advisors 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, 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 and ESA Antigua. 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

It is often believed that information cannot hurt and, in general, helps individuals make better decisions. For example, providing information about performance can correct the tendency that most people have to overestimate their absolute and relative abilities when solving a task in the lab (e.g., Hoelzl & Rustichini, 2005; Moore & Healy, 2008) or when making subjective assessments (e.g., Svenson, 1981; Englmaier, 2006). Outside of the lab, providing information to correct biased beliefs has the potential to avoid sub-optimal decisions. However, only a handful of studies examine the effect of this approach in real-life settings.[1] Moreover, little is known about how correcting beliefs may have unintended consequences if some individuals are hurt by the content of the information received.[2] Considering that many institutions today already provide some sort of relative performance feedback (e.g., ranking of students in schools), examining the effects of this type of policy in real-life scenarios becomes imperative.

This paper examines the effects on beliefs, performance and academic decisions of providing relative performance feedback to students. I focus on a high-stakes context where relative performance beliefs are particularly consequential - college entrance exams - and aim to answer the following two questions: How do beliefs, academic investments, performance, and choices change when students learn about their relative performance? And, are beliefs elicited using an incentive compatible task coherent with the beliefs revealed by real-life behavior? To answer these questions, I use field and lab-in-the-field experiments, and administrative data of 440 students preparing for a college entrance exam at a test preparation institute in Colombia[3] to find impacts on the following outcomes: (i) evolution of beliefs over time, (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. My first empirical strategy relies on randomly

[1] In the education context, Bobba and Frisancho (2016) and Gonzalez (2017) elicit absolute performance beliefs regarding a mock exam and provide information on actual performance. Dizon-Ross (2018) elicits parents' beliefs about the performance of their children and provides them with clear information about their kids' school performance. These studies find that correcting beliefs affects individuals' decisions.

[2] Relative performance feedback generally has positive effects on effort in lab settings (Gill et al., 2016; Eriksson et al., 2009; Kuhnen & Tymula, 2012), as well as on academic performance (Tran & Zeckhauser, 2012; Azmat & Iriberri, 2010, 2016; Bandiera et al., 2015; Jalava et al., 2015). Recent evidence has started to show that relative performance feedback may have negative effects on academic performance (Murphy & Weinhardt, 2018; Azmat, Bagues, Cabrales, & Iriberri, in press).

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

assigning students taking the test preparation course to receive feedback about their relative performance in weekly practice tests. 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 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.

To study belief updating, my second approach embeds a weekly lab-in-the-field experiment into the main experiment. I adapt an incentive-compatible mechanism in the spirit of Mobius et al. (2011) 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.[4] 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.

My first finding is that bottom performers in the preparation course make adjustments to their academic investments, in response to learning that they lie in the lower end of the performance distribution. Poor performers seem discouraged by relative performance feedback. Even though most treated students exert similar levels of effort and are equally likely to take practice tests than control students, students in the bottom quartile of initial practice test performance invest less in preparing for the entrance exam. Self-reported weekly hours of independent study fall by about 30 percent in math and reading in the bottom quartile. Information from the institute's administrative records shows that participation in weekly practice tests also falls by 5.2 percentage points from a base of 96 percent in the control group.

My second finding is that the discouragement effect is so strong that it even dissuades some students from taking the college entrance exam they are preparing for. Using administrative data from university records, I am able to observe who takes the exam and which major(s) they declare. Again, low-performers receiving feedback are the most discouraged. Those who were at the bottom of the distribution of scores within the test preparation course are 11 percentage points less likely to take the exam that similar students in the control group. What is most surprising is that there is also a negative effect on exam-taking

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

rates among high-performing students. Students who were in the top quartile of initial practice test performance are 5.8 percentage points less likely to take the exam than control students of the same ability.

The effect of providing relative performance feedback on the probability of taking the entrance exam lasts through the next college admission cycle. Students who decided not to take the exam right after finishing the preparation course may postpone taking the exam if they feel that they are not well prepared for the major they intend to apply to. University records from the following entrance exam shows that top- and bottom-performing students receiving feedback are more likely to never register for either of the two admission cycles than students of similar ability in the control group.

Among those who take the exam, I can compare students' intended major reported in a midline survey to the major they actually declared when registering for the entrance exam. I find that, consistent with learning their ability, bottom performers are 25 percentage points more likely to switch to an easier major relative to no one changing in the control group.

Third, I find that there are no statistical differences in admission rates or performance in the entrance exam across treated and control groups. This suggests that there is a change in the composition of the students who decide to take the entrance exam. In fact, comparing the practice test scores of students who take and do not take the entrance exam, I find that the highest and lowest performers are those who choose not to take the entrance exam. In the first case, I present evidence that those who decide not to take the exam were not scoring lower than other treated or control students in the initial practice test but have lower scores in subsequent practice tests. Hence, not taking the exam is a rational decision based on their weak performance. In the case of top performers, the evidence suggests that they may be eligible for scholarships and may have better outside options in general.

Even though performance in the entrance exam does not differ statistically by treatment, the treatment group experiences a small negative effect on math and reading scores in practice tests. Consistent with lower investments in study time, bottom performers have about 1.8 and 1.3 fewer correct questions in math and reading, respectively, across all tests. Treated students in quartile 2 also perform worse in practice tests, which can reflect the anxiety generated by learning that they are located at the bottom of the top group.

Fourth, 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), I study how my treatment affects men and women differentially. Women do not give up their preparation for the exam when they receive feedback, that is, they do not adjust their investments. However, they are less likely to take the entrance exam and to take the entrance exam in the next admission cycle than control women. Men in the bottom quartile, on the other hand, study less and take fewer practice tests than men in the control group. They also adjuct their decision to take the exam but to a less extent. 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. Overall, the effect of feedback reduces effort but not performance among low-performing men, and discourages high-performing women from taking the entrance exam.

Having established that feedback had substantial effects on students' decisions, I now turn to the question of whether the beliefs students report in the lab-in-the-field experiments are consistent with the beliefs revealed from their observed actions. Top performers receiving feedback are about 30 percent more likely to be accurate in predicting their relative performance relative to the control group across all rounds. Because they know they are at the top of the distribution and reflect this knowledge in the incentive compatible task, it could be expected that they are at least as likely to take the entrance exam as students in the control group. This is because the feedback is providing them additional clues about their probability of gaining admission. In the case of bottom performers, there are no differences between the beliefs reported by students in the treatment or the control group, which suggests that students in this quartile may have not internalized the information provided. Thus, it could be expected that there were no differences in exam taking rates between treatment and control. However, I find that students in these two quartiles make decisions that would be expected from individuals who internalize the information received through feedback.

This paper contributes to four bodies of work in education and experimental economics. First, a recent literature studies how correcting beliefs about academic performance improves decision making of students (Bobba & Frisancho, 2016; Gonzalez, 2017) and parental investments in their kids (Dizon-Ross, 2018). I contribute to this literature by adding the dynamic dimension on belief elicitation and feedback provision, the relative instead of absolute nature of the feedback, and by connecting beliefs elicited with an incentive compatible task commonly used in the lab with real-life behavior. Second, this paper adds to the research finding inconclusive results of providing relative performance feedback on grades (Azmat & Iriberri, 2010, 2016; Bandiera et al., 2015; Azmat et al., in press; Murphy & Weinhardt, 2018) by shedding light on which students are hurt by feedback and by highlighting other margins in which students are affected. I find that low performers get discouraged and reduce effort, and that relative performance feedback may affect other margins besides academic performance such as the decision to take a college entrance exam and college majors choices. Third, this paper contributes to the large literature studying the effect of feedback provision in the lab (e.g. Gill et al., 2016; Azmat & Iriberri, 2010, 2016) that finds that individuals exerting more effort when feedback is provided and greater responses from individuals at the extremes of the performance distribution. I add to this literature by showing that feedback affects important real-life outcomes and that the direction of the effects is not the same as in the lab. Finally, my paper contributes to the literature studying information processing in the lab (e.g. Eil & Rao, 2011; Grossman & Owens, 2012; Mobius et al., 2011; Ertac, 2011) by linking responses using a lab task with real-life behavior.

Lastly, my findings shed light on ways policy makers can improve students' outcomes. Many schools around the world provide information of the ranking of students within their class. This paper shows that students at the bottom of the distribution can be especially discouraged by such news. On the other hand, this paper also establishes that feedback helps align students' decisions with their abilities, which is likely to have positive benefits for the students themselves and for society. 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 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 students about relative performance in practice tests. To test how beliefs and actions change when students receive feedback regarding their relative performance, I divide the sample in treatment and control groups. In 8 of the 10 weeks, I provide feedback about the exact quartile the students' scores fall in to the treatment group only.

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. For many students, especially those coming from low socio-economic backgrounds, such schools are their only chance of earning a college degree. Because tuition at public universities is free or highly subsidized, earning a slot is extraordinarily hard.

In Colombia, students graduating from high-school who are willing to enroll at a public university are 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.

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 Medellin campus. Every calendar year, the entrance exam takes place in April and September. The exam contains 80 questions divided 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. Overall, admission rates are around 10 percent but this varies substantially by major. At the Medellin campus in the April, 2018 admissions cycle, 21 of the 83 majors offered had admission rates below 5 percent. 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).[5]

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. Importantly, 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.

Given the competitiveness of this exam, there are many institutes offering courses to help prepare students. 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 Medellin. 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 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 in-

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

stitutes 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. The institute allowed me to contact all students enrolled in the preparation course taking place from January to April, 2018. In total, 1,045 students consented to participate in the study. 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.

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 relative performance beliefs 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.

Admission is determined by the exam performance relative to other test takers. 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, and they get access to this information about one month before the exam is administered. One way to provide this type of relative performance information is by comparing students' performance within the test preparation institute, which is the basis of my design.

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 preparation course, regardless of participation in the study. 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.

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

Of the nearly 1,200 students enrolled in the institute for a preparation course for admission at Universidad de Antioquia, 1,045 accepted to participate in the study. In Section 3, I will show that the actual sample size is much lower because some students did not actively participate in the study.

Randomization: Of the 1,045 students who consented to participate, I randomly assigned half of them 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 also to increase power. Students are organized in classrooms at the beginning of the course after they choose the time slot in which they want to take the course. Because it is unlikely that most of them know each other from before the course and the treatment is based on each student's individual performance, it is unlikely to find important spillover effects in this setting.

Belief elicitation task: The belief elicitation task is based on an incentive compatible mechanism designed by (Mobius et al., 2011) to elicit the probabilities of being in each quartile of the math and reading score distributions. 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. The framing used for the task was that of receiving 12 tokens per test subject to play at a casino by betting on the quartile in which the students thought their score would be in. 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: the first time right after the practice test (priors), and the second time right after control students see their absolute scores and treated students see their absolute scores plus a "signal" indicating that their score was above or below the median (posteriors). The purpose of the second belief elicitation was to understand how students use information immediately after receiving it and compare their updating with that of a Bayesian agent with the same priors as they reported in the first elicitation. They were instructed that after the second belief elicitation, they would throw a dice to determine how much they would earn if they were selected in the weekly raffle.

The incentive compatibility of the belief elicitation consists in incentivizing truth telling by providing weekly cash prizes on one of the two belief elicitations chosen at random. Once completing the second belief elicitation and feedback (in the case of the treatment group), 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 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 from the dice draw. 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 large 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. Students in the control group saw the standard performance report containing 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.[6] An example of this report is in Figure 1. Students in the treatment group were directed to a feedback report showing their beliefs and actual standing in the distribution right after completing the second round of belief elicitation.

The key difference between the report treated and control students see is that control students only get access to absolute scores and report posterior beliefs right after seeing their scores. Along with absolute scores, treated students receive a “signal” indicating whether their scores are above or below the median, report posteriors after seeing the signal, and then see complete information of their relative performance.

Timeline: The field experiment timeline is in Figure 2, and the timeline for the lab-in-the-field experiment used to elicit prior and posterior beliefs is in Figure 3.

2.3 Data and outcomes

The analysis uses four sources of data. Primary sources are the weekly belief elicitation surveys, and a midline and two follow-up 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: I collected data from participants using belief elicitation surveys as well as midline and follow-up surveys. I elicited priors across 10 rounds

[6] These scores tend to be lower and the distribution is more compressed than the scores they obtain in the actual exam.

after each practice test except the first. Students reported 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, the perceived difficulty of the test, and how confident they felt about gaining admission. 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) after students checked the absolute scores in the performance report. 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.

The main outcomes from beliefs elicitation surveys include whether students are correct, underplace, overplace, have a flat prior, or have inconsistent beliefs.[7] I create indicator variables for each of these categories. For example, according to instructions given to students on how to allocate the tokens among the quartiles, the correct belief variable is coded as one if students assign most tokens to the quartile in which their scores lies in or the assign equal number of tokens to 2 or 3 quartiles including the one in which their score is. In addition, for the treatment group, I can compute how the students' posteriors relate to what a Bayesian would have updated after receiving the above / below median signal.

The midline and first follow-up were administered online. Between 3 weeks and one month after the course started, participants filled out a survey asking about their intended majors, predicted scores, among others. The main outcome I use from the midline survey is their intended first choice major. Two follow-up surveys were conducted a few days and six months after the entrance exam of April, 2018. 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 427 students actively participating in the experiment, I was able to reach about 75 percent of the sample in the 6-month follow up survey.

Secondary data sources and outcomes: Participants' data from the experiment were matched to administrative records from the test preparation institute, university admission statistics, and the national standardized exam administered by the Colombian agency for

[7] A belief is inconsistent if students seem to be assigning tokens at random. For example, if they assign most tokens to non-adjacent quartiles.

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 most of these characteristics as variables to 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. Even though not all students checked the performance report modified by the experiment, there is no reason to expect selective attrition because it is unlikely that they knew their treatment assignment.

3.1 Attrition

There are two sources of attrition. The first is related to students who never engaged with the experiment and did not check the performance report. The second is among students who checked some but not all reports.

Overall, 56 percent of students who consented to participate never checked the performance report. This does not mean that students did not know their absolute scores. They could access them right after finishing the online practice tests and as soon as the institute graded their individual practice test when it was administered on paper. Because the intervention required having the scores of all students available, there was a delay between the moment in which the students could check the absolute scores and the distribution of the performance report. This feature of how the institute reports scores may have discouraged students to check the performance reports and participate in the weekly raffles of cash prizes.

The other source of attrition is that very few students accessed all 8 performance reports. It proved extremely difficult to engage them in the study when they were not present at the institute. Thus, because the report was available a few days after they took the practice tests, many students would not check it consistently because it required effort at times where they were not present at the institute.

While the attrition 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. Across all quartiles and within quartiles the difference in the proportion of students checking at least one performance report is not statistically different. Nevertheless, between quartiles there are substantial differences. A higher fraction (55 percent) of students in the best-performing quartile checked the report at least once relative to about 40 percent in other quartiles. These differences do not matter for the analyses because they are performed within quartile.

The number of times students checked the report by quartile conditional on having checked at least once does not differ by treatment. In all quartiles, students check the report 2.4 times on average. These varies from 2.1 times for students in the bottom quartile to 2.7 for students in the top quartile. The overall mean and means by quartile are in Table 3. Graphical evidence on the number of times checking the performance report is in Figure 4. These histograms are almost equal for treatment and control in the quartiles above the median but have a higher mass at one performance report for treated students in the quartiles below the median. This suggests that students realizing that they were not performing relatively well are more likely to check the reports only once.

3.2 Final sample characteristics and balance

Table 1 presents baseline characteristics of students in all quartiles who checked at least one of the experiment performance reports along with p-values of individual and joint tests of differences between treatment and control. None of the characteristics has statistical differences. Because the basis for the empirical analysis will be the quartiles in the initial practice test, Table 2 in the appendix shows that characteristics are also balanced within the quartiles.

On average, students in my sample are 59 percent female, almost 18 years old, and single. Based on their SISBEN score and residential strata, both measures of socio-economic status, these students are in low and middle-low income households. 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.

From the data collected by the institute, students obtain a score of around 38 points out of 100 possible in the first practice test. Their scores in math are substantially lower than in reading. Almost 50 percent of the sample is enrolled in the morning courses, 32 percent in the afternoon / evening courses, 4 percent in the weekend courses, 4 percent in the course preparing them simultaneously to two entrance exams at different universities, and 11 percent are in a pre-medicine course.

All in all, despite attrition, the samples used in the analysis have internal validity. I perform the analysis with interaction terms between treatment and quartile indicators and compute treatment effects within quartile.

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 Medellin. In both cases, the distributions of scores of students in my sample is notoriously to the right of the scores of all other high-school graduates. 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 that the sample is positively selected 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 the are not good while in fact they are but they are being compared to students who are much better. 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 decision making and beliefs

This section shows the effects of the relative performance feedback intervention on academic outputs and inputs, and on students' beliefs, and discusses whether students' actions conform with expressed beliefs. 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. Bottom performers receiving feedback are 6 percentage points less likely to show up to practice tests, 11 percentage points less likely to take the entrance exam, and 25 pp more likely to switch to majors requiring lower cutoff scores. These lower participation rates and major changes are 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.

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 own ability. Bobba and Frisancho (2016) and Gonzalez (2017) elicit beliefs about students' performance in a mock exam and provide information about their performance. Both studies find that students are more likely to choose academic options more in line with their ability if they receive this information. Dizon-Ross (2018) elicits parent's beliefs about their children academic performance and find that parents who receive this information adjust schooling inputs and children's enrollment. Another strand of the literature highlights that certain types of information may not have the expected effects. Murphy and Weinhardt (2018); Azmat et al. (in press) find that providing relative performance information such as class rank discourages effort and aspirations of students. My study draws from both strands as is able to provide insights on how beliefs and investments change in the period between the intervention and observing final academic outcomes, and on the characteristics and margins in which students are hurt by feedback.

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. Hence, the estimation is based on initial quartiles which are good predictors of the type of feedback treated students receive (control students would have received) across most rounds.

To obtain treatment effects, I first run a regression of the outcome of interest on an indicator for treatment (T_i), 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 in 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 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 5 and 6 reveals 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 Effects on relative performance beliefs

Students in the bottom quartile do not seem to incorporate the feedback they receive in their relative-performance assessments. Top performers, on the other hand, become more accurate over time 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 they have 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.[8] In this research, I distinguish between what Moore and Healy (2008) call overestimation (people think their ability or performance is better than it is) and overplacement (people believe they are better than others).

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

[8] Research on overconfident individuals, however, shows that thinking that we are better than we actually are make people work harder (Chen & Schildberg-Hörisch, 2018).

timeline of how belief elicitation works in a given week is in Figure 3.

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 over- (under-) estimate 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.

Different from lab experiments in which most people overplace their performance relative to others(Moore & Healy, 2008), I find that overplacing is as likely as underplacing. 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 perfect of top-performing control students have correct beliefs in reading, 23 percent overplace and 27 percent underplace. 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 come from less overplacement. The patterns for math are similar in magnitude.

Poor performers' beliefs, however, do not change relative to their peers in the control group. The bottom panel of Tables 9 and 10 presents the treatment effect for students in the bottom quartile of initial practice test performance. Around 30 percent of these students have correct beliefs, 33 percent overplace in reading and 40 percent overplace in math, and about 16 percent underplace in either subject. In contrast with top performers, there are no statistical differences for the treatment group, which suggests that bottom performers are not very successful incorporating the information they receive through feedback.

After students receive the above / below median signal, updating follows the same pattern for priors. Top-performing treated and control students become more correct in their relative performance predictions but treated students are about 14 percentage points more likely to hold and accurate belief in both, math and reading, over a base of 51 and 47 percent, respectively. This means that relative to the prior stage, control students update their beliefs by 20 percentage points just by looking at the absolute scores they obtained in the practice test.

Less than 12 percent of treated and control top-performing students overplace. What becomes more prevalent after students check the performance report is to underplace. Over 30 percent of control students underplace in reading and 36 percent underplace in math. Top-performers in the treatment group are 10 and 15 percentage points less likely to underplace in reading and math, respectively. Posterior beliefs of bottom performers in the control and treatment groups are nearly the same as at the prior stage.

A central question for this paper is: What would we expect to see in real life behavior given how students incorporated the information provided into their reported beliefs? Because bottom performers in the treatment group do not seem to be incorporating the information, I do not expect them to have a different behavior relative to control students. That is, if their reported beliefs correspond to their actions, they should not show signs that they update when receiving feedback. Top performers, on the other hand, seem to be incorporating the information. Hence, I expect that they have similar or higher levels of investments and are at least as likely to take the entrance exam. I discuss these issues in subsection 4.6.

4.3 Effects on academic decisions and exam performance

In this subsection 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. Poor performers receiving feedback are more likely to switch to majors with lower cutoff scores relative to the control group.

Existing empirical studies in the education context have found mixed results regarding the effects of providing relative performance feedback on academic performance. Some studies find increases in performance when providing information to students about their scores relative to the mean (Azmat & Iriberri, 2010) or information of the rank within class (Tran & Zeckhauser, 2012). In contrast, two recent studies find that relative performance feedback may have negative effects on student academic performance when analyzing secondary school grades for those who are in lower ranks within their primary school class (Murphy & Weinhardt, 2018), and when providing information about the decile of the distribution in which students' scores lie (Azmat et al., in press).

Given the ambiguous results from previous studies, the effect providing feedback is still an open question. Relative performance beliefs and feedback have only been studied by Azmat et al. (in press) but only indirectly because the beliefs they elicit are not from the same students receiving the intervention. Moreover, the effects of feedback on academic choices has only been assessed in lower-stakes environments such as choosing academic subjects in secondary school (Murphy & Weinhardt, 2018). 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.

The first main finding is that feedback affects the decision of taking the college entrance exam that students in my sample are preparing for. Table 11 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 worst and best 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, the top and bottom quartiles, 100 percent of students in the control group took the exam. For the second quartile of performance (below the top but above the median), the effect is -4.4 percentage points over a base of 94.6 percent. However, I do not have enough power to find this effect to be statistically significant although its magnitude is not small.

The seemingly discouraging effect of feedback could be due to students postponing because they feel they are not prepared enough; however, they do not take it in the next admission cycle. Column 2 of Table 11 shows the effect of never registering for the exam in two consecutive admission cycles. This variable is an indicator equal to one if the students did not register for the April, 2018 and did not register for the September, 2018 exam. 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 may last for at least two admission cycles.

In the case of bottom performers, those receiving feedback and who took the exam in April seem to be more resilient than those in the control group. Analyzing bottom performers who did take the exam in April but decided not to take it again in September, I find that students receiving feedback are less likely to give up. Column 3 of Table 11 shows that bottom performers receiving feedback are 23.1 percentage points less likely to give up than control students. From a base of 65.8 percent of the control not taking the exam in September, this coefficient represents a change of 65 percent. My interpretation of this finding is that students who were receiving feedback knew that they were not performing relatively well before the actual exam while students in the control may have learned this just when they received the exam results. In this sense, treated students may have been better prepared for the bad news and less discouraged to keep trying. For other quartiles, I do not see any significant difference on giving up.

Turning to performance in the entrance exam and admission, the scores students obtain in the exam increases monotonically with quartile in jumps of almost 10 points but do not differ statistically between treatment and control (Table 12). There is a small disadvantage in math scores for treated students in quartiles 1 to 3 and a small positive advantage in quartile 4. The differences in reading are very close to zero.

There are no statistical differences in admission rates between treatment and control in Table 12. However, recall that there has been some selection into who we see taking the exam.[9]. In the control group, admission rates for the first option students declare increase almost monotonically with quartile of initial practice test performance from 7.3 percent in the bottom quartile to 4.4 percent in quartile 3, 13.5 percent in quartile 2 and 31.6 in the top quartile. We see a negative coefficient on admission rates for top performers receiving feedback, and positive coefficients for students in quartiles 2 and 3.

Because the outcome of being admitted is a combination of the score students obtain and majors they declare, top performers having (not statistically significant) lower admission rates than the control group could be the result of: (i) Students who took the exam and received feedback declared majors that were very hard to be admitted into; (ii) The

[9] This selection will be discussed in more detail in the next subsection

applicant pool of treated students that we see taking the exam excludes the best among the top performers. In Table 13 we see that top performers tend to choose majors with cutoff scores that are above two standard deviations of the mean cutoff scores at slightly higher rates than control students.[10] In Subsection 4.4 I will discuss selection into taking the exam.

Students in quartiles 2 and 4 have non-significant positive admission coefficients of 11.5 and 3.1 , respectively. In this case the explanations behind the positive coefficients could be again related to the pool of applicants and the majors they select. There is suggestive evidence (Table 13) that treated students in these quartiles are less likely to choose extremely competitive majors. Also, in quartile 4 students who were performing at the bottom were the ones who decided not to take the exam.

Finally, another way bottom performers were affected by feedback is the decision to switch to an easier or harder major. With the subsample of students who completed the midline survey (172), I calculate the difference between the cutoff scores of the major they declared when registering for the exam and the major they intended at the time of answering the survey. I take the average of the cutoff scores in the previous two admission cycles which is the information students would likely use when making such decisions. The outcome measuring switching to a harder major is an indicator equal to one when the major they declare has a higher cutoff score than the major they intended. Analogously, switching to an easier major is equal to one when the cutoff score is lower in the declared major relative to the intended. Table 13 shows that bottom performers are 23.6 percentage points less likely to switch to a harder major over a base of 40 percent in the control group. Simultaneously, they are 25.4 percentage points more likely to switch to an easier major and no one in the control group makes such change. Students in quartile 2 are 28.7 percentage points less likely to switch to a harder major relative to 42.9 percent of students in the control group that do so.

4.4 Effects on academic investments

Students in the bottom quartile who receive feedback take practice tests less often, study fewer hours, and perform worse in practice tests than similar students in the control group. There are no differences among students in other initial performance quartiles.

[10] I do not see the same in Table 13 which shows whether students switch to a harder and easier major from the midline survey to the actual exam registration. In this table, however, the sample is smaller because not all students responded the midline survey.

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.[11] Work focusing on debiasing students' beliefs conducts surveys but does not look at how the feedback they provide to students affects their investments while preparing for an admission exam assigning seats to Mexico City's public high schools (Bobba & Frisancho, 2016), or when enrolling in advanced placement (AP) courses in the US (Gonzalez, 2017). An exception is Azmat et al. (in press) 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 their dynamics. I can also find if feedback affects beliefs, perceived difficulty of practice tests, and confidence in gaining admission.

The first finding is that bottom performers in the treatment group are 5.7 percentage points less likely to take practice tests.This represents a reduction of about 6 percent relative to the base of students taking 95.2 percent of all practice tests in the control group (see Table 14). The breakup by week is in Figure 9. It shows that the vast majority of students in the sample take the practice tests, except in round 8 (practice test during Holy Week). Bottom performers are not consistently taking fewer practice tests but fail to show at higher rates than the control in several weeks. This results does not hold for other quartiles. On average, students take over 90 percent of the practice tests in the treatment or the control group.

In contrast with Azmat et al. (in press), I find that 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 asked for time spent working on problems and excluding class and practice test time. The case in which I find marginally significant but economically important effects is for bottom performers who study 2 fewer hours for math and 1.4 fewer hours for reading than control students who study 6.4 and 5.2 hours per week, on average, respectively. 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.

As expected from investing less time studying for the entrance exam, students receiving feedback obtain fewer correct practice test questions in math and reading (Table 14). Bottom performers have 1.7 and 1.3 fewer correct out of 40 questions in practice tests than students in the control group. Surprisingly, treated students in quartile 2 also have lower

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

performance in practice tests even though they did not study significantly less. They obtain 1.8 and 1.4 fewer correct questions than the control group in math and reading, respectively. One hypothesis that I am unable to test is that students in this quartile feel they are at the bottom of the top and get more anxious when they take practice tests. The dynamics of these two variables by quartile are in Figures 12 and 13.

4.5 Explaining selection into taking the entrance exam

One of the main findings of this paper is that students who receive feedback are less likely to take the exam particularly if they are at the top or bottom of the distribution. In this subsection I present evidence of a compositional effect in the pool of applicants that decide to take the exam.

The first piece of evidence is that bottom performers who decide to not take the exam are those whose practice tests scores were at the very bottom of the distribution. In Figure 14 even though the distribution of correct answers in math looks very similar for treated and control students and the p-value of a Kolmogorov Smirnov test is above 0.1, splitting the treatment group into those who take and who do not take the exam (right panel) shows that those who decide to not take the exam are disproportionately at the left tail of the distribution. Thus, relative to the control group in which everyone takes the exam, the relative performance feedback dissuades the worst performers from taking the exam who, in any case, have a very slim chance of gaining admission.

The evidence is less clear for top performers but suggests that those who decide to not take the exam are slightly better performers. Figure 14 shows that the distribution of math scores are similar - although with more mass on the right hand side for treated students - for treatment and control students. Once again, all control top performers took the exam so the righ panel of the Figure shows practice test math scores for treated students who took and did not take the exam. The density for those who did not take the exam is slightly shifted to the right, suggesting that students who decided not to take the exam are positively selected.

4.6 Do students' actions match their beliefs?

Even though poor performers do not seem to update their relative performance beliefs, they act like if they did because they invest less in practice tests and study time and take the entrance exam at lower rates than the control group. Top performers update in the lab task but are also less likely to take the exam. In this section I revisit these findings and propose some hypothesis to explain the discrepancies.

In Bayesian learning, individuals have beliefs about their ability that evolve according to signals that they receive from different sources[12]. 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 $\varepsilon_i \sim N(0, \sigma^2)$. 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 the ranking of their absolute score in the college majors they declare before taking the exam. In this sense, to obtain a slot at the university, what matters 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 behave according to Bayesian learning. However, this is not what they express when stating their beliefs in the lab elicitation task.[13] I discuss several hypothesis that may be behind this inconsistency. My design did not intend to disentangle different hypotheses but it is certainly an open area for future work.

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

[13] 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.

The first hypothesis involves individuals' 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 mistake. Indeed, lower ability students had a harder time understanding the task.

A second 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 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., 2011) that captures that the individuals care about their belief about how good they are. My experiment cannot provide direct evidence supporting this hypothesis 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).

The third 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 incentive. 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 2 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.

The fourth hypothesis is that the belief elicitation tasks do not do a very good job in eliciting the beliefs that people use to make decisions. 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, bottom 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. In fact, we saw that those who were performing worst were the ones not showing up to the exam. 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. Again two hypothesis could explain this behavior. First, they may be behaving as "workaholics" in the sense introduced by Müller and Schotter (2010). These are individuals who seem unable to stop working. In this case, they may think that they need more preparation so they can achieve their goal of gaining admission. For now, this hypothesis does not find support because they were not more likely to register for the September entrance exam. A second hypothesis is that they have better outside options. This hypothesis is being evaluated by conducting a survey on participants to better understand if they are beneficiaries of scholarships or are studying at other prestigious public universities.

5 Heterogeneity of relative performance feedback effects by gender

This section shows that, while women do not seem to change their level of effort, they react more strongly than men to receiving relative performance feedback in academic decicisions. Women in the treatment group contribute more than proportionally to the effects on taking the entrance exam. Men in the treatment group, on the other hand, are less likely to make decisions that are different from those made by the control group but lead the effects on lower academic investments documented in the previous section. The tables presenting results in this section contain treatment effects separately by gender, and follow the structure of the tables in the previous section.

Tables 15 and 16 show that women's beliefs are more responsive to receiving feedback than me'ns beliefs. In particular, women in the top quartile are 15 percentage points more likely to have correct priors in math and reading if they receive feedback from a base of between 36 and 39 percent of women in the control group, respectivelt. As I discussed previously, students in other quartiles do not seem to be reacting to feedback as their reported beliefs do not differ significantly from those of the control group. Across all quartiles, men are more accurate in their relative performance beliefs by a substantial amount and their beliefs are less responsive to receiving feedback.

Table 17 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. Women in the top quartile are 7.5 percentage points less likely to take the entrance exam and 7.8 percentage points more likely to not take the exam over two admission cycles if they receive feedback. Likewise, the effects for students in the bottom quartile are larger for women but indistinguishable from zero for both genders given the small sample size in this quartile.

Performance in the entrance exam, as explained earlier, does not vary by treatment status within gender. However, between gender, there are enormous differences in performance, especially in the two top quartiles. Men overperform women in math by 7 and 15 points in math in quartiles 1 and 2, respectively. The differences are smalle in the bottom two quartiles and women oputperform men in math in the bottom quartile. Men's advantage is less clear in reading, with scores that are 4 points above and 3 points below those of women in quartiles 1 and 2, respectively, and virtually equal in the two bottom quartiles. Overall, the overall scores that the university uses for admission are substantially higher for men in the top two quartiles. This is at odds with the results of Table 22 that show men and women having equal numbers of correct answers in practice tests in both subjects.

Given the important differences in scores between males and females, it is not surprising that admission rates are much lower for women as shown in Table 19. The differences are highest in the top two quartiles in which, for the control group, admission rates are 25 and 41 percent, for women and men respectively in quartile 1, and 6 and 26 percent in quartile 2. Despite the fact that treated women in the bottom quartile are less likely to switch to a harder major (Table 20), their admission rates are only slighlty higher than those of women in the control group (4.3 percent).

Ex ante, given that in general women's performance in practice tests does not differ from that of men, it makes sense that there are no differences in the cutoffs of majors that students of both genders choose.[14] In fact, Table 20 show that cutoff scores of majors chosen are very similar by gender. What is interesting is that the scores also do not decrease for students in lower quartiles even though I documented that the exam scores fall moniotonically by about 10 points with quartile.

Finally, the discouragement I found in the previous section reflected in the reduction of academic investments such as taking practice tests and study hours is primarily concentrated among male students. Table 21 shows that men in the bottom quartile are 6.7 points less likely to attend practice tests, and study math about half of the time male control students study per week. Another sign of discouragement could be that performance in practice tests is lower. The only case in which I fond differences by treatment status is among women in quartile 2. They have 3.1 and 2.7 correct questions in math and reading, respectively. This is likely explained by the fact that treated women 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". It is not hard to imagine that this may put extra pressure on them to try to be at the top and may harm their performance in practice tests.

6 Conclusion

In this paper, I design a field and lab-in-the-field experiments to understand how relative performance feedback affects students' beliefs, decisions and academic performance. I assemble a panel dataset from student surveys and use administrative data to reach four main findings. First, bottom performers become discouraged and invest less in preparing for the exam. Second, top- and bottom-performing students receiving feedback are less likely to take the college entrance exam they are preparing for in two consecutive admission cycles. Bottom performers are more likely to switch to easier majors than similar students in the control group. Third, male and female students react differently to feedback. While women do not decrease investments as men do, they are more likely to change their decision to take the entrance exam. Fourth, the beliefs students report in an incentive compatible elicitation

[14] Recall that students choose first and second option majors before they take the entrance exam so they base their decisions on information they have up to the moment of exam registration.

task do not match the beliefs revealed their real-life decision making.

Contrary to the traditional idea that "information can't hurt", I present evidence that providing relative performance information can discourage students at the bottom of the score distribution to try harder and in some cases 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 slots. On the other hand, it may not be ideal in other contexts or from a policy perspective seeking to reinforce effort in the long run rather than immediate achievements.

One limitation of this paper is that it studies a very specific population that is not representative of the average student in Colombia or other countries. Another limitation is that power may be limited to detect some effects that are economically meaningful but the sample sizes are too small to detect them as statistically significant. In future work I hope to address these shortcomings, as well as address other important questions such as what is the best way to provide feedback to poor performing students without discouraging them to exert effort.

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. Adding to the findings of recent work by Murphy and Weinhardt (2018) and Azmat et al. (in press), I show that providing relative performance feedback can in fact reduce student effort and investments in terms of study time, and taking exams and practice tests. In this sense, my findings should raise awarness on how students are being informed of their performance given that providing rank within a class is a widespread practice across the world. Perhaps the information per se is not the problem. Rather, the discouragement effect may arise from how information is delivered. I provided information in a very raw, straightforward way, just as schools do. Finding ways to inform students without discouraging them seems feasible, and something I am eager to explore as the immediate next step in my research agenda.

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. (in press). 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*.

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2013). Overconfidence and social signalling. *The Review of economic studies*, *80*(3), 949–983.

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.

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 childrens education: Experimental evidence from malawi. *American Economic Review (forthcoming)*.

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.

Englmaier, F. (2006). A brief survey on overconfidence. *Behavior Finance*.

Eriksson, T., Poulsen, A., & Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, *16*(6), 679–688.

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.

Gonzalez, N. (2017). *How learning about one's ability affects educational investments: Ev-*

idence 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.

Hoelzl, E., & Rustichini, A. (2005). Overconfident: Do you put your money on it? *The Economic Journal*, *115*(503), 305–318.

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).

Jalava, N., Joensen, J. S., & Pellas, E. (2015). Grades and rank: Impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, *115*, 161–196.

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

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

Kuhnen, C. M., & Tymula, A. (2012). Feedback, self-esteem, and performance in organizations. *Management Science*, *58*(1), 94–113.

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. (2011). *Managing self-confidence: Theory and experimental evidence* (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.

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.

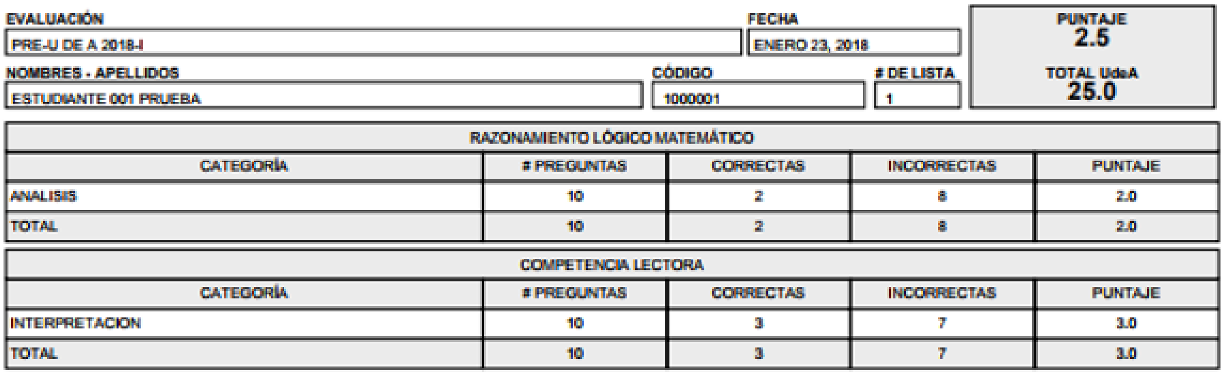
Schotter, A., & Trevino, I. (2014). Belief elicitation in the laboratory. *Annu. Rev. Econ.*, *6*(1), 103–128.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta psychologica*, *47*(2), 143–148.

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.



EVALUACIÓN PRE-U DE A 2018-I	FECHA ENERO 23, 2018		PUNTAJE 2.5
NOMBRES - APELLIDOS ESTUDIANTE 001 PRUEBA	CÓDIGO 1000001	# DE LISTA 1	TOTAL UdeA 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: Results report for control group

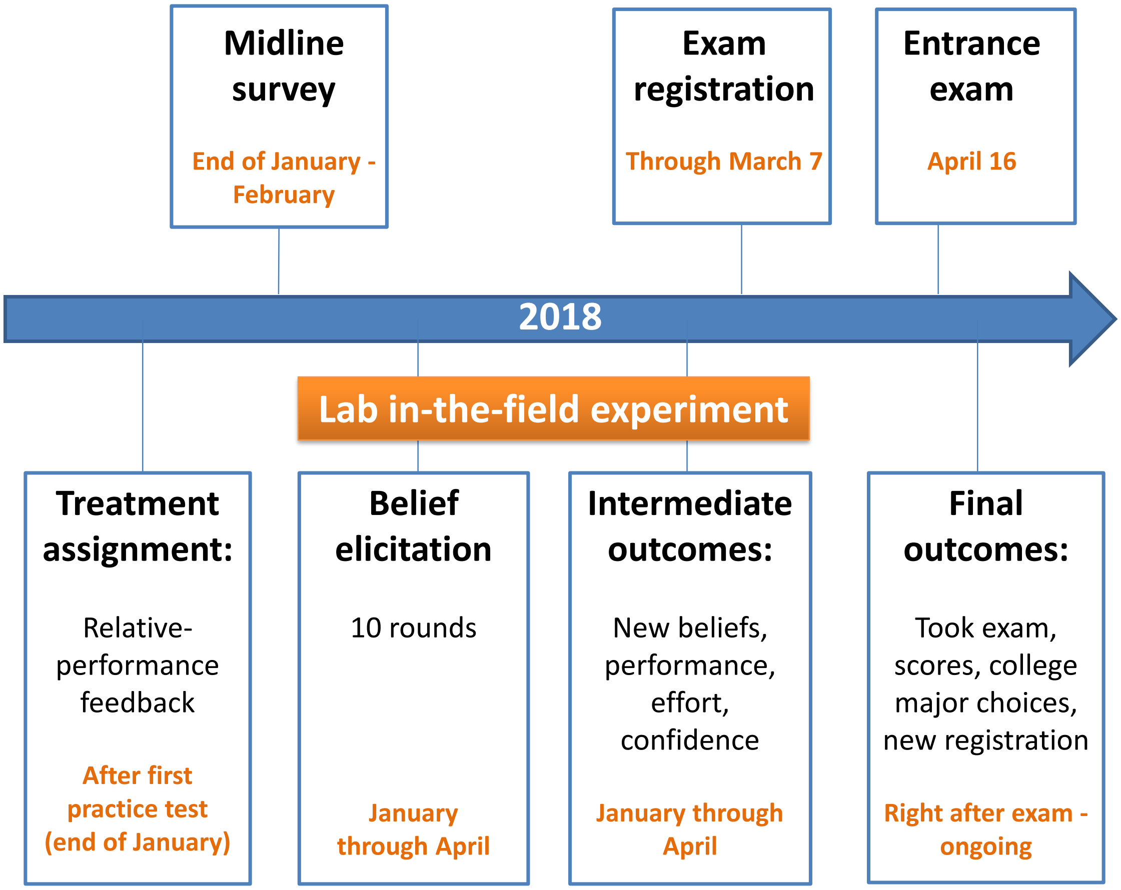

Midline survey
End of January - February
Exam registration
Through March 7
Entrance exam
April 16
2018
Lab in-the-field experiment
Treatment assignment:
Relative-performance feedback
After first practice test (end of January)
Belief elicitation
10 rounds
January through April
Intermediate outcomes:
New beliefs, performance, effort, confidence
January through April
Final outcomes:
Took exam, scores, college major choices, new registration
Right after exam - ongoing

Figure 2: Timeline

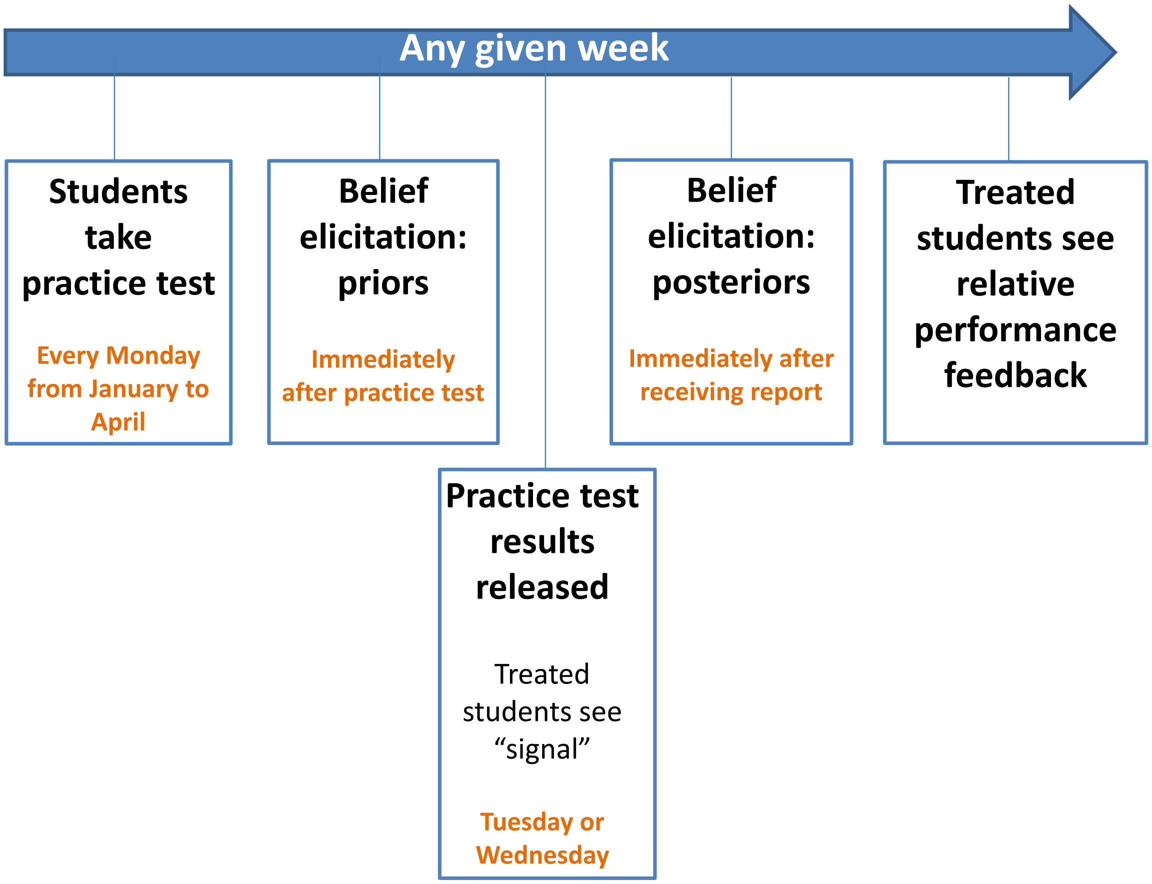

Any given week
Students take practice test
Every Monday from January to April
Belief elicitation: priors
Immediately after practice test
Practice test results released
Treated students see "signal"
Tuesday or Wednesday
Belief elicitation: posteriors
Immediately after receiving report
Treated students see relative performance feedback

Figure 3: Beliefs elicitation within rounds

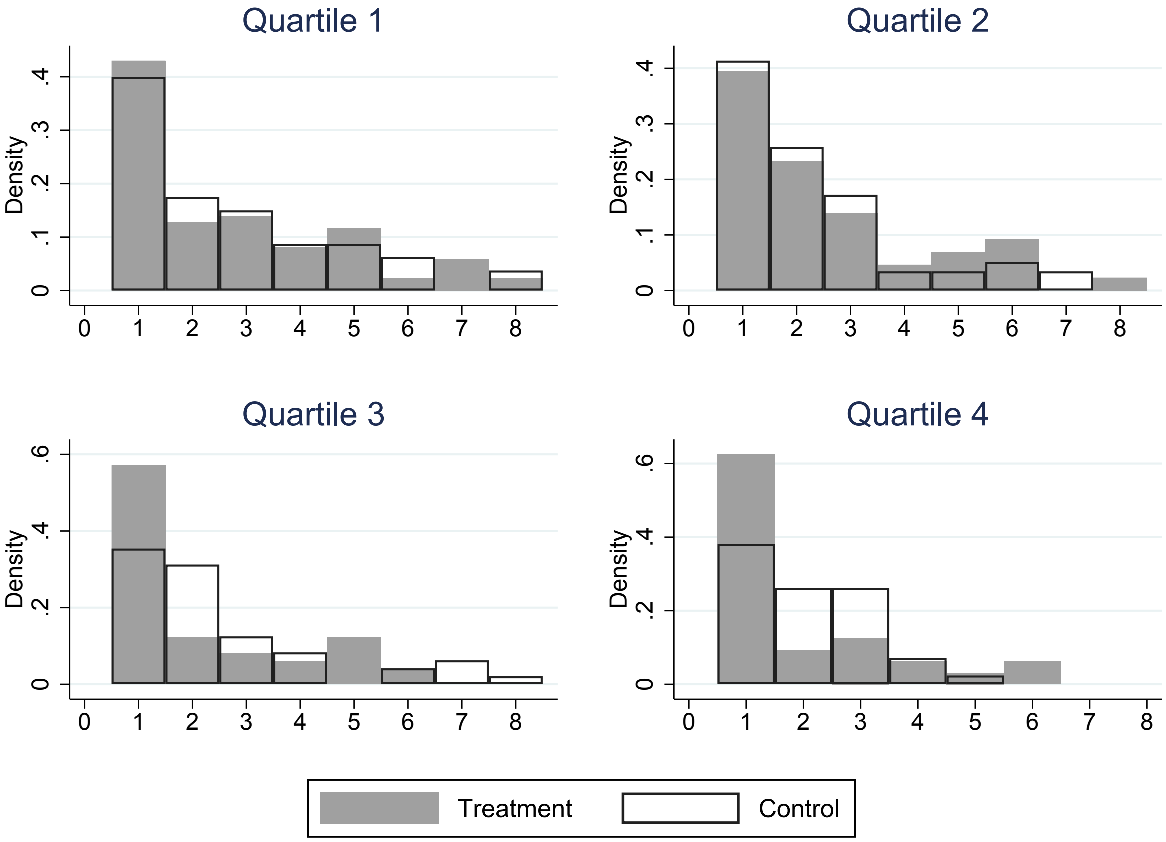


Figure 4: Number of times students check performance report by quartile

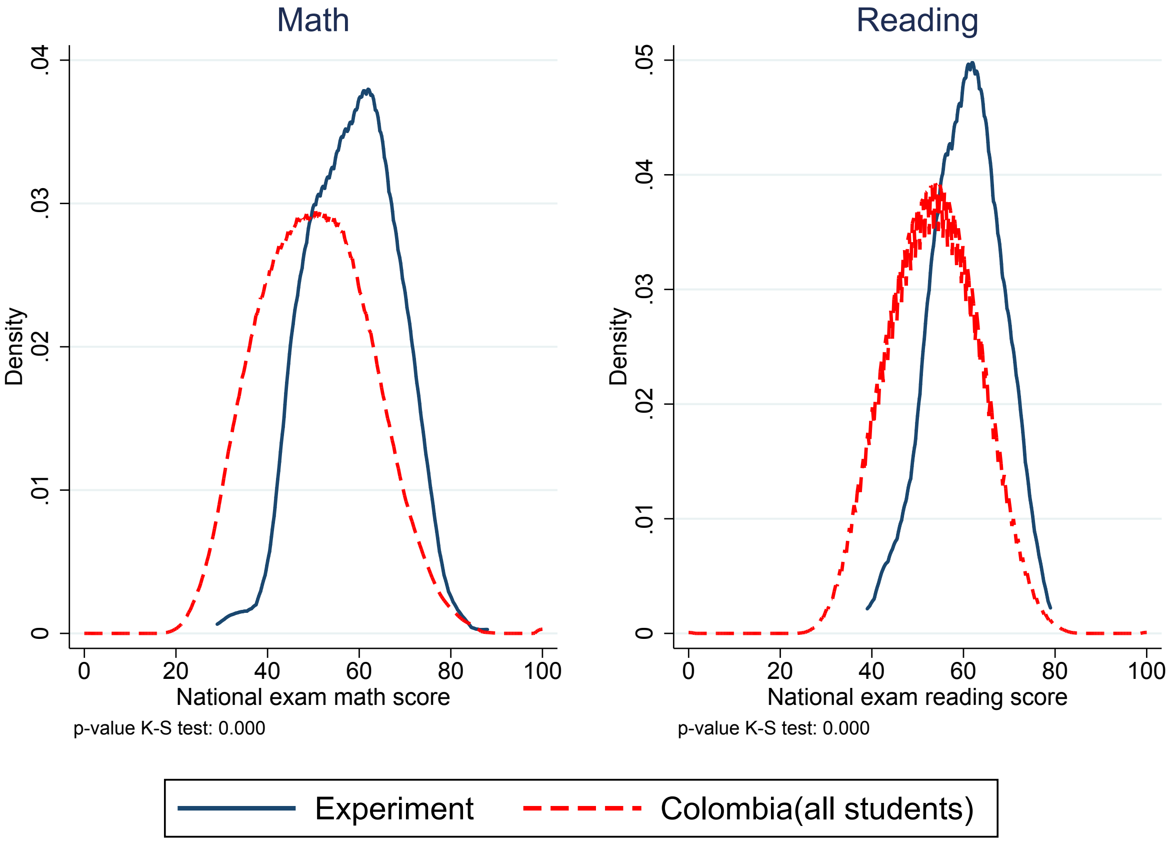


Figure 5: Positive selection of students in the sample relative all high-school graduates in Colombia

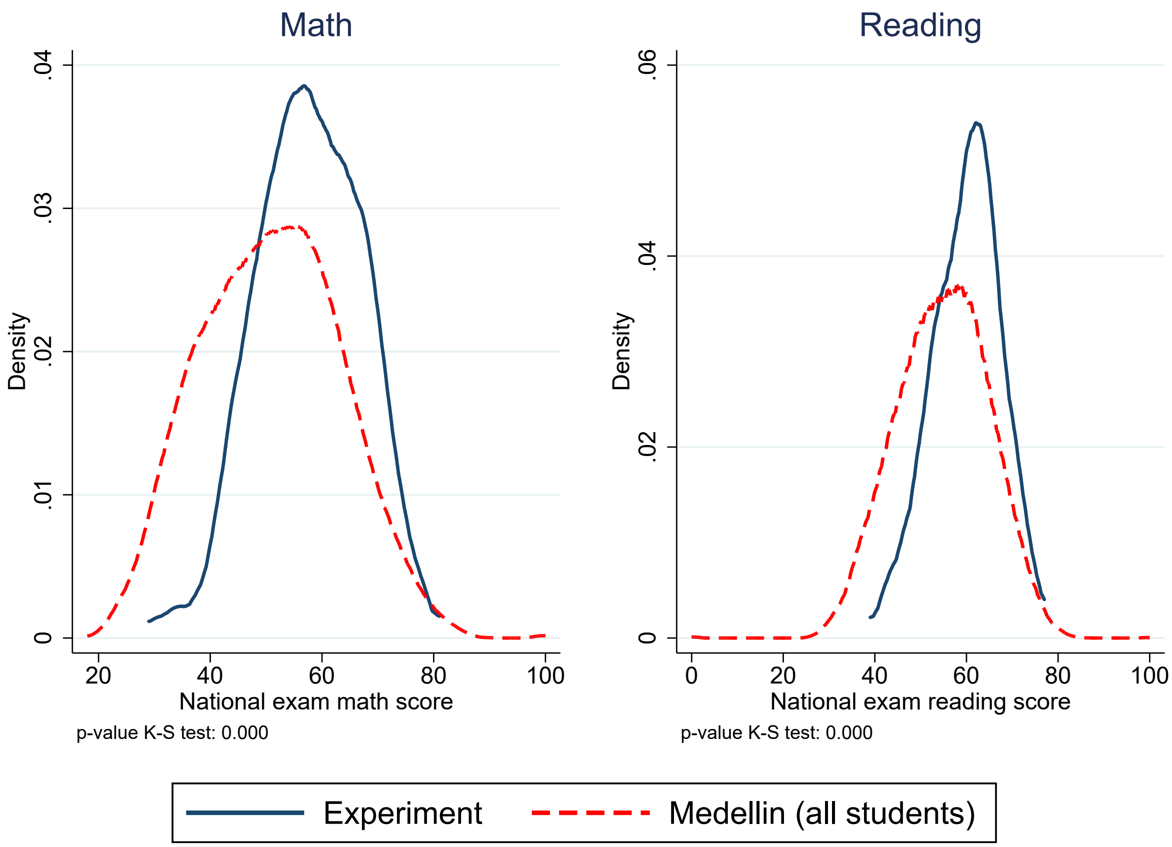


Figure 6: Positive selection of students in the sample relative all high-school graduates in Medellin

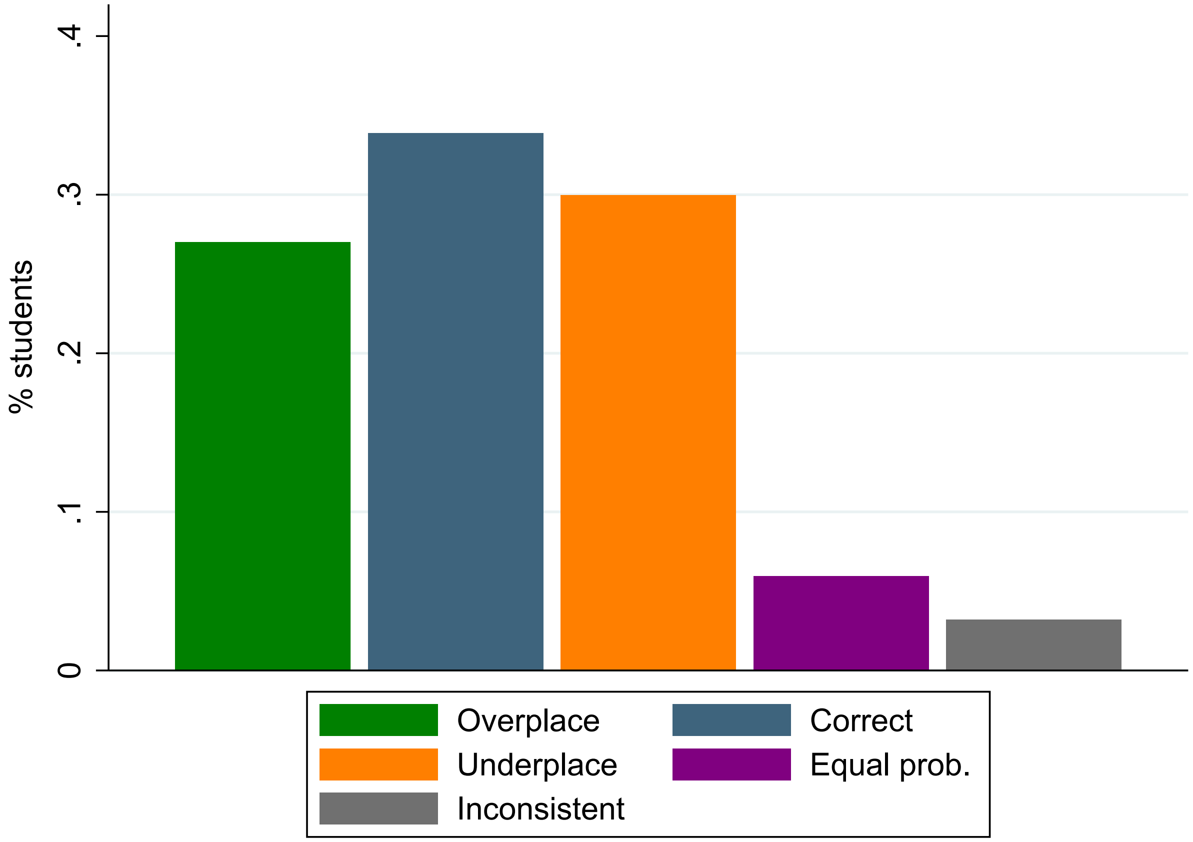


Figure 7: Classification of prior beliefs depending on quartile with highest probability assignment and actual quartile - all students

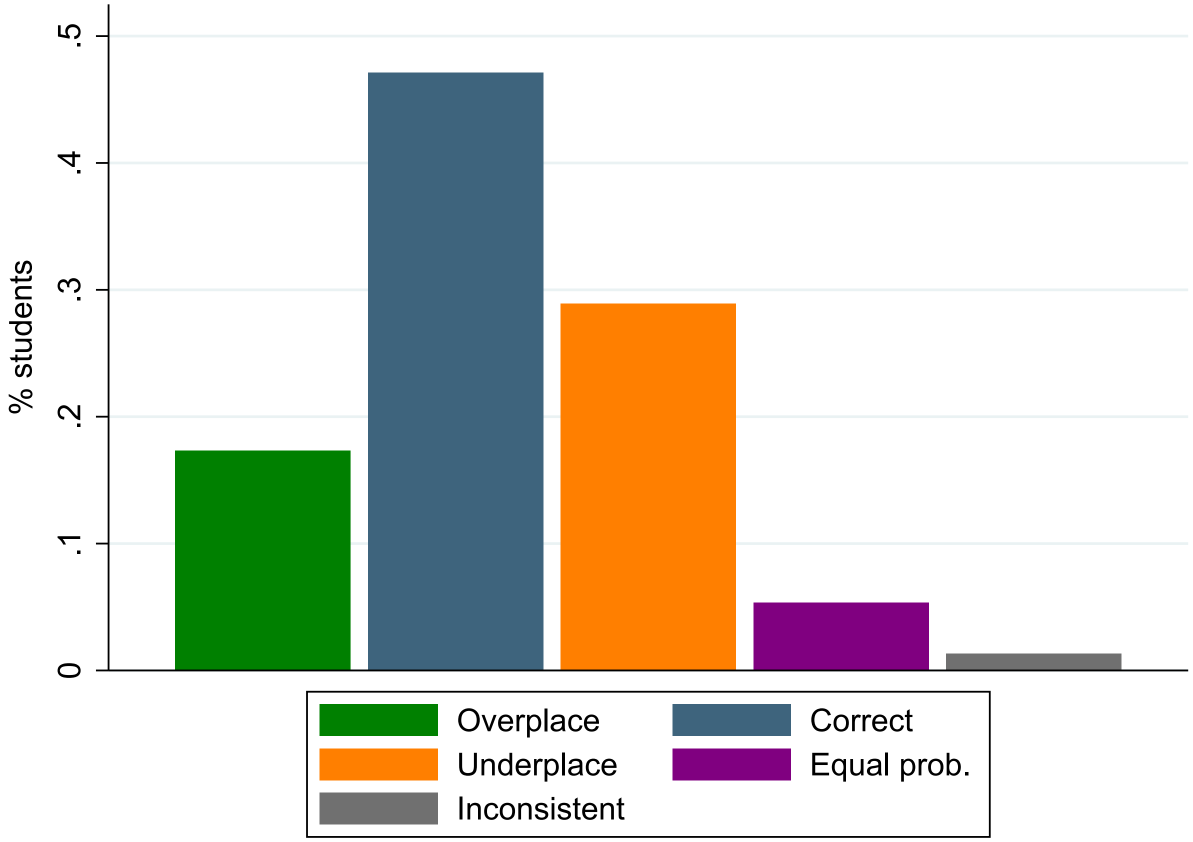


Figure 8: Classification of posterior beliefs depending on quartile with highest probability assignment and actual quartile - all students

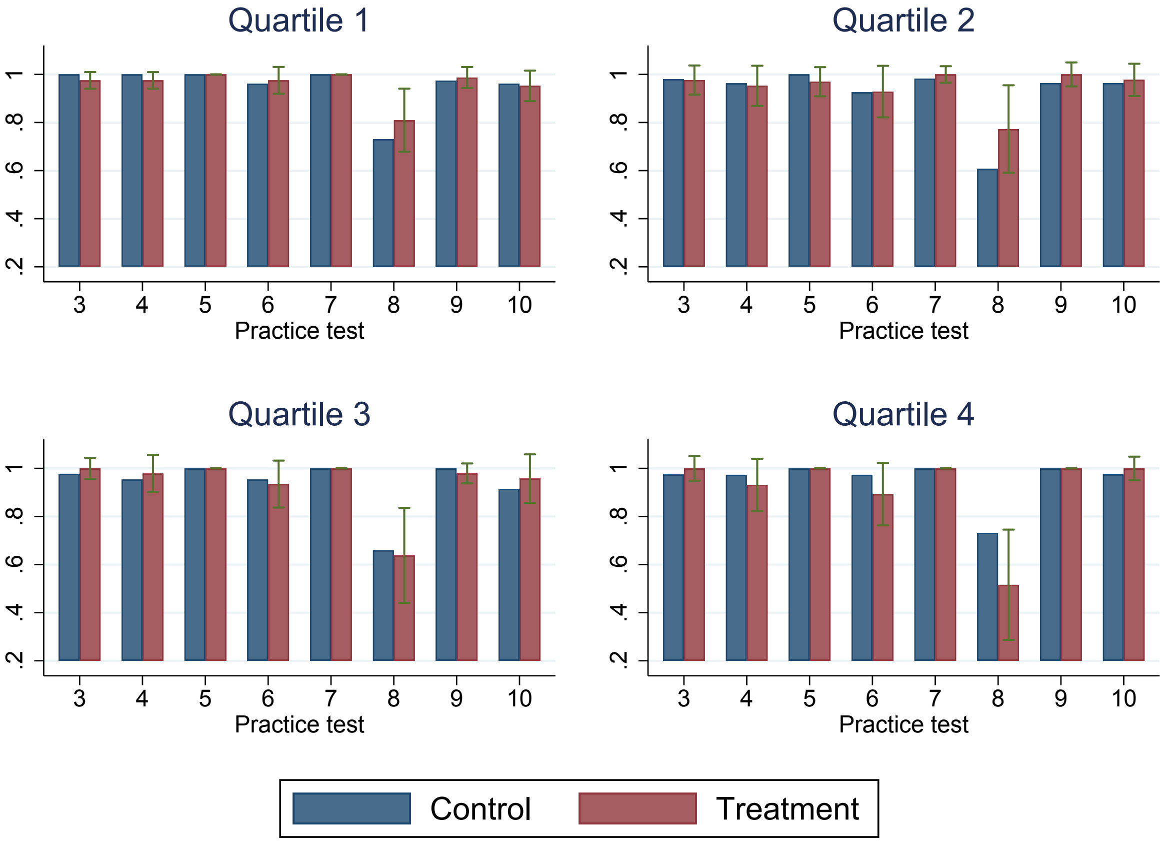


Figure 9: Fraction of students taking practice tests by week and quartile

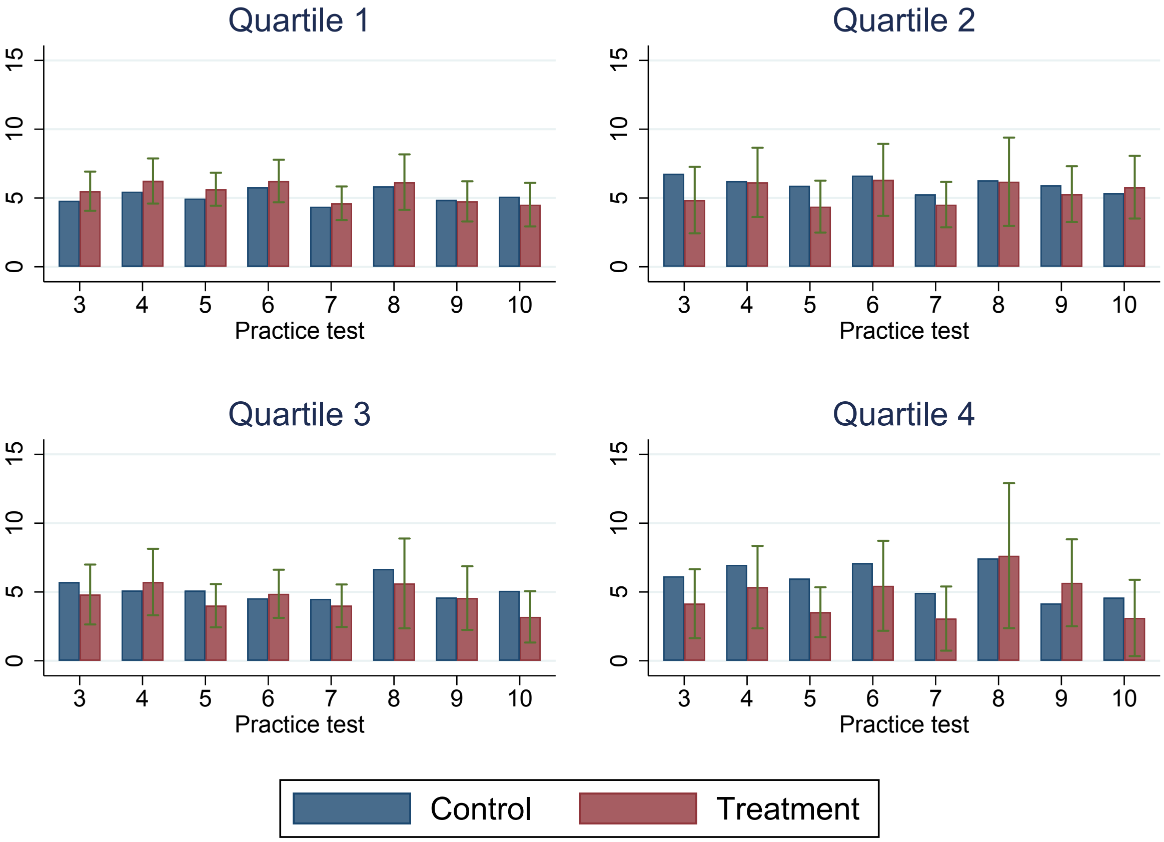


Figure 10: Average study hours in math by week and quartile

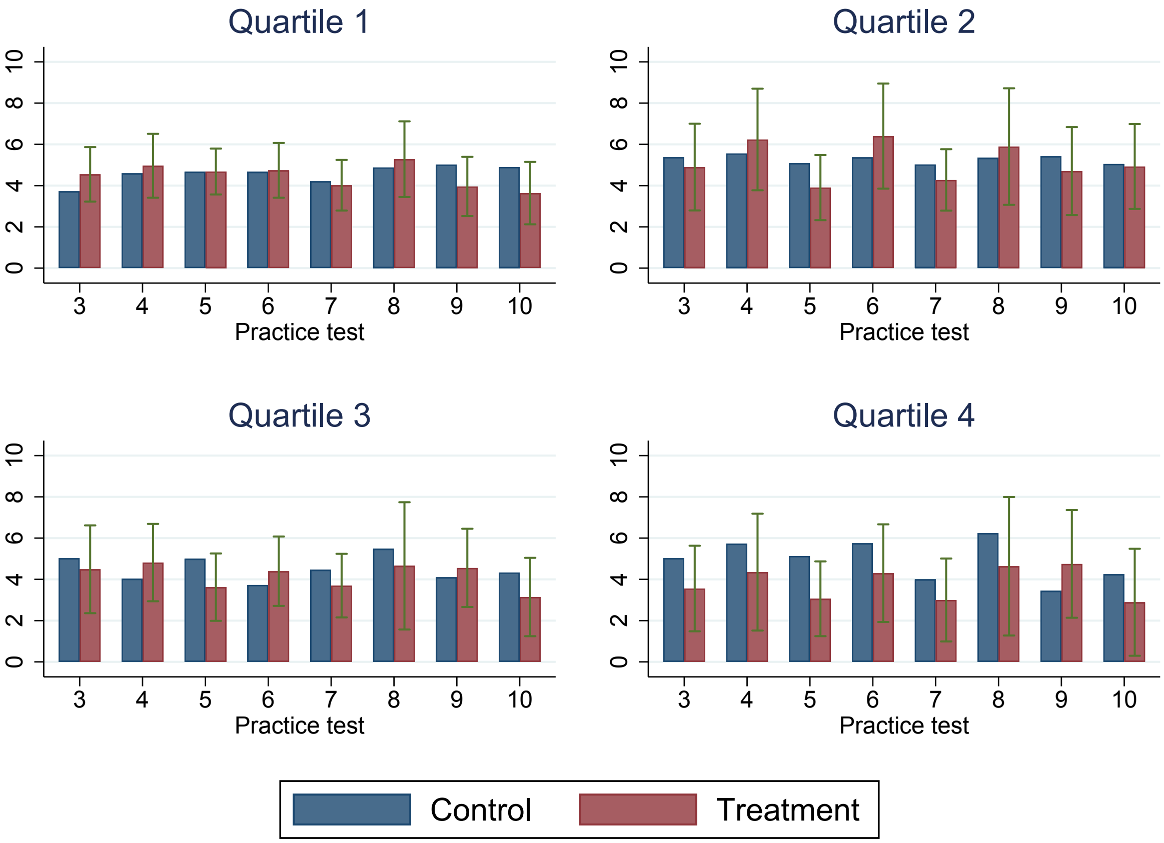


Figure 11: Average study hours in reading by week and quartile

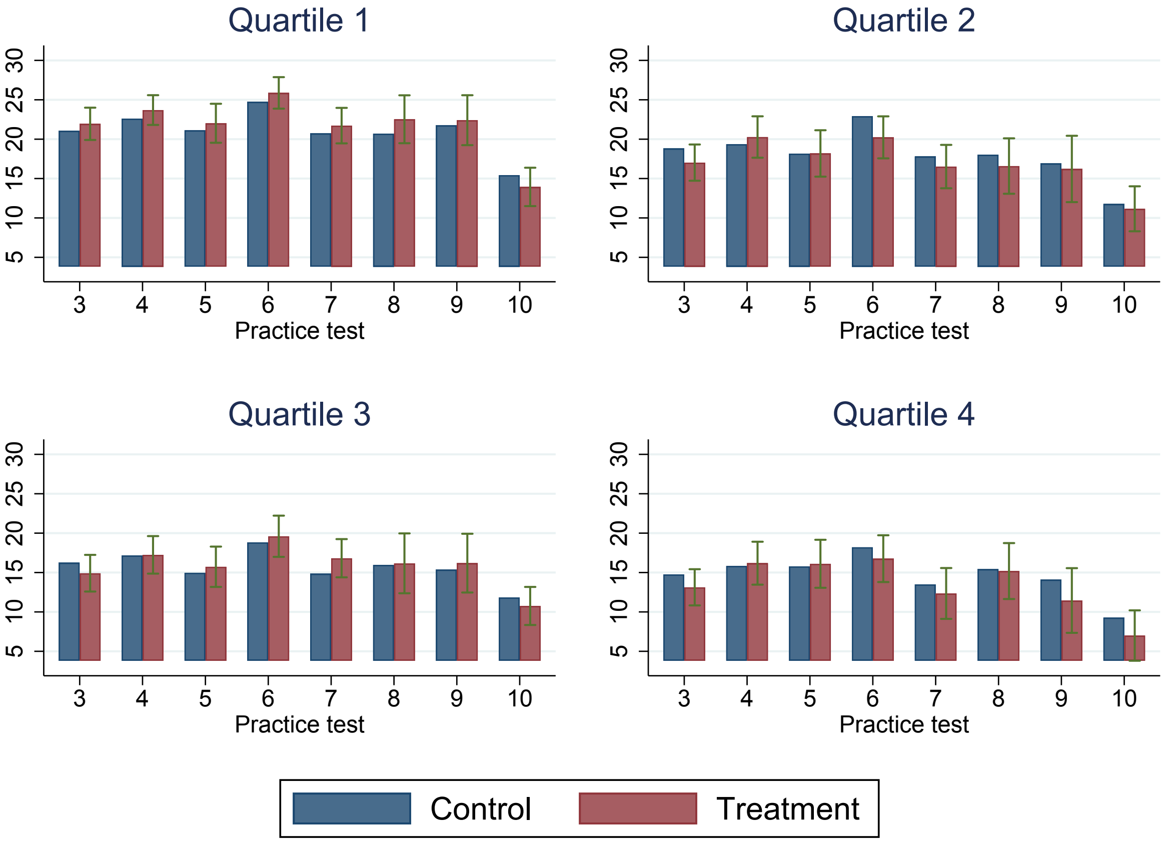


Figure 12: Practice tests: Number of correct answers in math by week and quartile

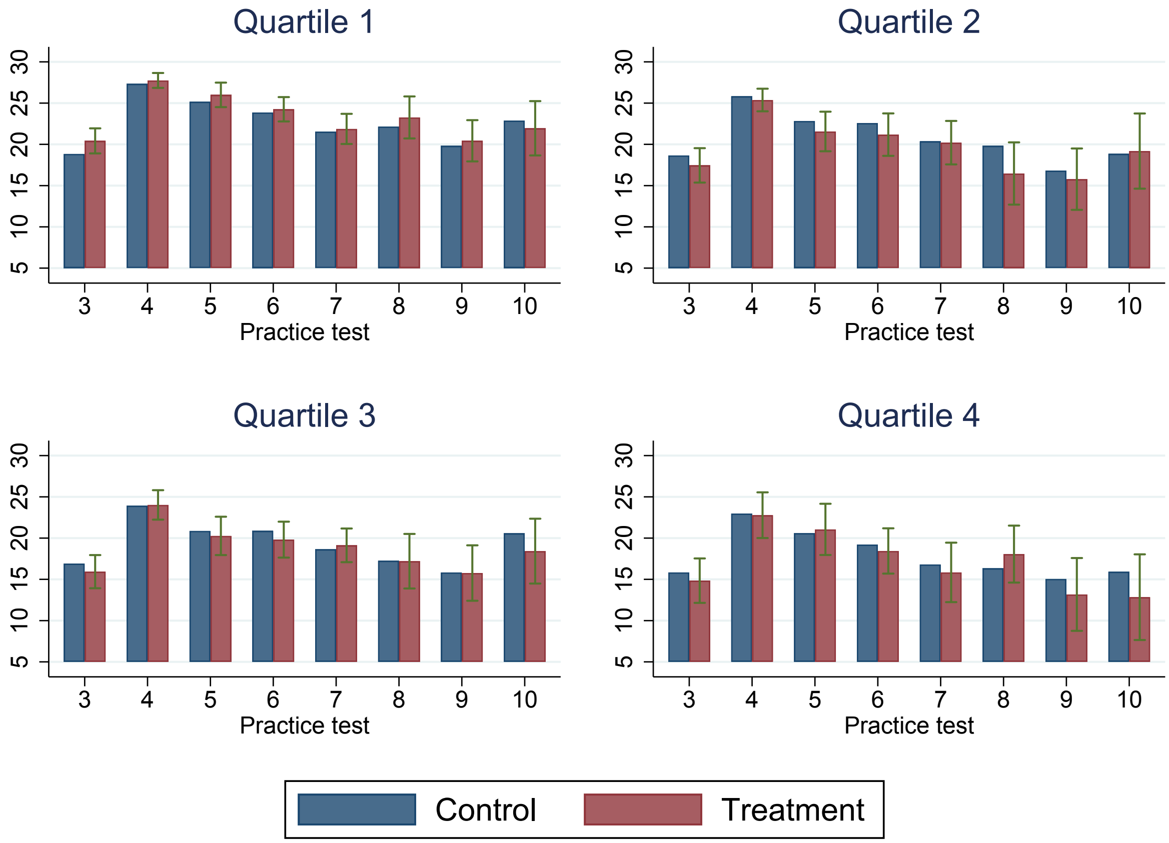


Figure 13: Practice tests: Number of correct answers in reading by week and quartile

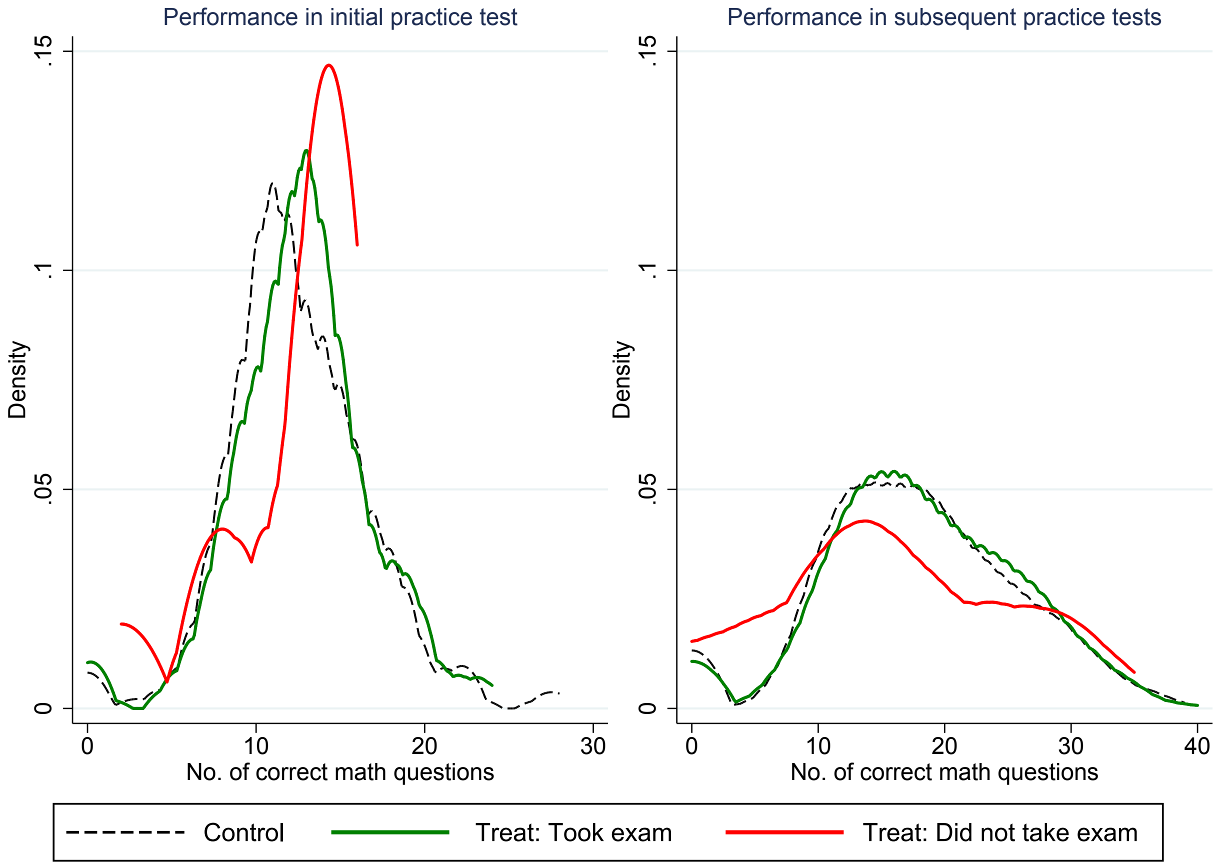


Figure 14: Bottom quartile: Number of correct answers in practice tests by treatment and by decision to take exam

Table 1: Balance of baseline characteristics

	Control	**Treatment**	**P-value (T-C)**	**No. obs**
Stratification variables				
Female	0.613	0.600	0.780	440
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	17.733	17.257	0.027	434
Single	0.973	0.976	0.787	433
Student	0.677	0.720	0.311	434
Residential strata	2.450	2.529	0.431	434
Urban	0.881	0.895	0.622	434
Academic variables				
Math no. correct (initial practice test)	11.579	11.811	0.553	439
Reading no. correct (initial practice test)	18.189	18.853	0.284	439
Avg. practice test score in classroom	38.067	38.143	0.762	440
Joint orthogonality test			0.2812	439

Notes: Robust standard errors. Each column contains the mean of the variable on the left-hand-side in the control and treatment groups. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 2: Balance table by quartile

	Q1 = top		**Q2**		**Q3**		**Q4 = bottom**	
	Control	Treat	Control	Treat	Control	Treat	Control	Treat
Female	0.652	0.562	0.604	0.592	0.652	0.683	0.527	0.563
Age	17.518	17.157	17.769	17.248	17.878	17.563	18.076	17.602
Single	0.974	0.975	1.000	1.000	1.000	0.981	0.951	1.000
Student	0.693	0.846**	0.763	0.667	0.649	0.714	0.793	0.760
Residential strata	2.627	2.866	2.485	2.336	2.319	2.540	2.608	2.304
Urban	0.910	0.892	0.891	0.888	0.877	0.938	0.876	0.896
Previously taken entrance exam	0.861	0.846	0.808	0.792	0.754	0.850	0.692	0.668
Math score (initial practice test)	3.915	3.842	3.061	3.108	2.763	2.751	2.023	1.992
Reading score (initial practice test)	6.383	6.555	5.189	5.242	4.171	4.286	2.821	2.444
Avg. practice test score in class	38.043	38.202	37.570	37.795	37.768	36.75*	36.190	36.435
AM course	0.925	0.874	0.902	0.928	0.820	0.922	0.863	0.914
PM course	0.303	0.361	0.382	0.416	0.446	0.468	0.459	0.480
Weekend course	0.009	0.008	0.013	0.000	0.008	0.015	0.009	0.012
Integrated UdeA - UNAL	0.007	0.02	0.021	0.027	0.051	0.012	0.043	0.058
Pre-medicine	0.166	0.191	0.145	0.138	0.107	0.122	0.199	0.101

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 3: 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 all participants	28.5%	25.0%	25.8%	20.7%	
Panel B. Students who checked at least one 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%	
Fraction of participants in quartile	56.8%	39.5%	36.7%	34.9%	
Panel C. Statistics or 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 4: Balnace of characteristics among attitors

	Control	**Treatment**	**P-value (T-C)**	**No. obs**
Stratification variables				
Female	0.553	0.575	0.592	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	17.682	17.667	0.953	568
Single	0.969	0.974	0.734	568
Student	0.822	0.834	0.747	569
Residential strata	2.618	2.581	0.681	569
Urban	0.907	0.919	0.643	569
Academic variables				
Math no. correct (initial practice test)	11.060	11.019	0.894	604
Reading no. correct (initial practice test)	17.461	17.252	0.676	604
Avg. practice test score in classroom	37.607	37.872	0.220	604
Joint orthogonality test			0.9572	551

Table 5: Persistence of initial quartile in reading

	Proportion of practice tests in reading quartile:			
	1 = top	2	3	4 = bottom
Panel A. Students in top quartile in initial practice test				
Treated	0.088**	-0.051*	-0.047*	0.010
	(0.043)	(0.027)	(0.024)	(0.019)
Constant	0.495***	0.256***	0.162***	0.087***
	(0.046)	(0.030)	(0.026)	(0.019)
Obs	1320	1320	1320	1320
No. students	166	166	166	166
Panel B. Students in quartile 2 in initial practice test				
Treated	-0.075	0.013	0.051	0.011
	(0.055)	(0.037)	(0.035)	(0.031)
Constant	0.320***	0.287***	0.234***	0.159***
	(0.057)	(0.039)	(0.035)	(0.037)
Obs	783	783	783	783
No. students	102	102	102	102
Panel C. Students in quartile 3 in initial practice test				
Treated	-0.024	0.027	-0.000	-0.003
	(0.041)	(0.040)	(0.041)	(0.038)
Constant	0.188***	0.236***	0.339***	0.237***
	(0.044)	(0.036)	(0.039)	(0.036)
Obs	743	743	743	743
No. students	97	97	97	97
Panel D. Students in bottom quartile in initial practice test				
Treated	-0.049	0.027	-0.013	0.035
	(0.045)	(0.043)	(0.039)	(0.058)
Constant	0.182***	0.284***	0.317***	0.217***
	(0.052)	(0.041)	(0.034)	(0.053)
Obs	551	551	551	551
No. students	75	75	75	75

Notes: Each column shows coefficients of a regression of a dummy indicating whether the student was in the quartile of the table header 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 are in the top quartile in about 50-59% of all subsequent practice tests. 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: Persistence of initial quartile in math

	Proportion of times in math quartile:			
	1 = top	2	3	4 = bottom
Panel A. Students in top quartile in initial practice test				
Treated	0.060	-0.021	-0.031	-0.009
	(0.055)	(0.035)	(0.025)	(0.021)
Constant	0.560***	0.231***	0.111***	0.098***
	(0.055)	(0.038)	(0.025)	(0.019)
Obs	1320	1320	1320	1320
No. students	166	166	166	166
Panel B. Students in quartile 2 in initial practice test				
Treated	-0.077	0.020	0.050	0.008
	(0.060)	(0.044)	(0.036)	(0.040)
Constant	0.393***	0.333***	0.147***	0.128***
	(0.064)	(0.044)	(0.031)	(0.044)
Obs	784	784	784	784
No. students	102	102	102	102
Panel C. Students in quartile 3 in initial practice test				
Treated	0.034	0.020	0.001	-0.055
	(0.057)	(0.040)	(0.040)	(0.046)
Constant	0.211***	0.310***	0.286***	0.194***
	(0.064)	(0.043)	(0.038)	(0.046)
Obs	743	743	743	743
No. students	97	97	97	97
Panel D. Students in bottom quartile in initial practice test				
Treated	-0.071	-0.038	0.055	0.054
	(0.047)	(0.055)	(0.047)	(0.068)
Constant	0.174***	0.335***	0.230***	0.261***
	(0.048)	(0.053)	(0.043)	(0.057)
Obs	552	552	552	552
No. students	75	75	75	75

Notes: Each column shows coefficients of a regression of a dummy indicating whether the student was in the quartile of the table header 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 are in the top quartile in about 56-62% of all subsequent practice tests. Standard errors are clustered at the individual level. *** significant at the 1% level. ** significant at the 5% level. * significant at the 10% level.

Table 7: Effect of feedback on prior beliefs - reading

	Correct	**Overplace**	**Underplace**
Panel A. Students in top quartile in initial practice test			
Treated	0.097**	-0.070**	-0.004
	(0.040)	(0.035)	(0.037)
Constant	0.409***	0.234***	0.257***
	(0.045)	(0.037)	(0.042)
Obs	1036	1036	1036
No. students	166	166	166
Panel B. Students in quartile 2 in initial practice test			
Treated	0.036	0.094**	-0.075
	(0.043)	(0.047)	(0.051)
Constant	0.268***	0.269***	0.277***
	(0.044)	(0.050)	(0.056)
Obs	583	583	583
No. students	100	100	100
Panel C. Students in quartile 3 in initial practice test			
Treated	0.016	-0.011	-0.026
	(0.047)	(0.048)	(0.036)
Constant	0.349***	0.365***	0.163***
	(0.043)	(0.045)	(0.040)
Obs	573	573	573
No. students	97	97	97
Panel D. Students in bottom quartile in initial practice test			
Treated	0.014	-0.108	0.037
	(0.057)	(0.072)	(0.057)
Constant	0.304***	0.345***	0.171***
	(0.045)	(0.062)	(0.054)
Obs	365	365	365
No. students	72	72	72

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 8: Effect of feedback on prior beliefs - math

	Correct	**Overplace**	**Underplace**
Panel A. Students in top quartile in initial practice test			
Treated	0.120***	-0.080**	-0.012
	(0.046)	(0.034)	(0.043)
Constant	0.354***	0.178***	0.360***
	(0.048)	(0.038)	(0.044)
Obs	1037	1037	1037
No. students	166	166	166
Panel B. Students in quartile 2 in initial practice test			
Treated	0.036	0.063	-0.018
	(0.057)	(0.049)	(0.059)
Constant	0.392***	0.164***	0.263***
	(0.057)	(0.048)	(0.065)
Obs	583	583	583
No. students	100	100	100
Panel C. Students in quartile 3 in initial practice test			
Treated	0.028	-0.051	-0.025
	(0.052)	(0.056)	(0.043)
Constant	0.357***	0.294***	0.216***
	(0.052)	(0.055)	(0.049)
Obs	573	573	573
No. students	97	97	97
Panel D. Students in bottom quartile in initial practice test			
Treated	0.017	-0.125*	0.007
	(0.062)	(0.071)	(0.051)
Constant	0.279***	0.403***	0.174***
	(0.051)	(0.074)	(0.049)
Obs	365	365	365
No. students	72	72	72

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 35-47% 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 9: Effect of feedback on posterior beliefs - reading

	Correct	**Overplace**	**Underplace**
Panel A. Students in top quartile in initial practice test			
Treated	0.137**	-0.040	-0.098*
	(0.056)	(0.035)	(0.050)
Constant	0.506***	0.111***	0.329***
	(0.066)	(0.042)	(0.061)
Obs	449	449	449
No. students	166	166	166
Panel B. Students in quartile 2 in initial practice test			
Treated	0.062	0.095	-0.043
	(0.070)	(0.062)	(0.063)
Constant	0.272***	0.260***	0.264***
	(0.075)	(0.073)	(0.069)
Obs	247	247	247
No. students	102	102	102
Panel C. Students in quartile 3 in initial practice test			
Treated	0.067	-0.020	-0.052
	(0.071)	(0.078)	(0.062)
Constant	0.383***	0.348***	0.183***
	(0.064)	(0.075)	(0.065)
Obs	228	228	228
No. students	97	97	97
Panel D. Students in bottom quartile in initial practice test			
Treated	-0.032	-0.113	-0.019
	(0.086)	(0.085)	(0.077)
Constant	0.335***	0.317***	0.274***
	(0.074)	(0.088)	(0.081)
Obs	152	152	152
No. students	75	75	75

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 posterior in 50.6-64.2% 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 feedback on posterior beliefs - math

	Correct	**Overplace**	**Underplace**
Panel A. Students in top quartile in initial practice test			
Treated	0.132**	0.014	-0.137***
	(0.057)	(0.036)	(0.051)
Constant	0.452***	0.118***	0.379***
	(0.061)	(0.042)	(0.054)
Obs	431	431	431
No. students	162	162	162
Panel B. Students in quartile 2 in initial practice test			
Treated	0.048	0.025	0.073
	(0.071)	(0.054)	(0.068)
Constant	0.434***	0.149**	0.198***
	(0.071)	(0.062)	(0.072)
Obs	236	236	236
No. students	97	97	97
Panel C. Students in quartile 3 in initial practice test			
Treated	0.110	-0.053	-0.076
	(0.074)	(0.068)	(0.072)
Constant	0.402***	0.237***	0.289***
	(0.079)	(0.070)	(0.098)
Obs	208	208	208
No. students	89	89	89
Panel D. Students in bottom quartile in initial practice test			
Treated	0.037	-0.132	-0.071
	(0.091)	(0.097)	(0.080)
Constant	0.421***	0.259**	0.276***
	(0.087)	(0.104)	(0.089)
Obs	147	147	147
No. students	73	73	73

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 posterior in 45-58% 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 feedback on taking entrance exam over two admission cycles

			ITT
	Did not take exam	**Never registered**	**Did not take exam**
Q1 = top	0.056**	0.059**	0.008
	(0.025)	(0.025)	(0.024)
Mean control	0.000	0.000	0.035
Q2	0.042	-0.000	-0.037
	(0.052)	(0.044)	(0.037)
Mean control	0.052	0.052	0.107
Q3	-0.016	-0.016	-0.041
	(0.024)	(0.024)	(0.025)
Mean control	0.021	0.021	0.062
Q4 = bottom	0.106*	0.104*	-0.025
	(0.057)	(0.056)	(0.036)
Mean control	0.000	0.000	0.091
N	438	438	985

Notes: Each point estimate is the treatment effect on the outcome in the column heading within the quartile of the initial practice test. Robust standard errors in parenthesis. For reference, the mean of the control group in the quartile is reported below the standar error. Controls in this 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 feedback on entrance exam scores and admission

	Math score	**Reading score**	**Total score**	**Admitted to first option**	**Admitted to second option**
Q1 = top	1.437	-3.049	-0.718	-0.067	0.024
	(3.173)	(2.642)	(2.317)	(0.072)	(0.031)
Mean control	71.046	73.831	72.329	0.309	0.025
Q2	0.267	-2.011	-0.788	0.119	-0.040
	(4.789)	(4.551)	(3.461)	(0.084)	(0.030)
Mean control	60.644	63.163	61.849	0.130	0.037
Q3	1.197	-7.914	-5.245	0.015	-0.009
	(4.899)	(4.966)	(4.208)	(0.050)	(0.039)
Mean control	50.553	53.538	53.319	0.043	0.043
Q4 = bottom	-0.163	3.859	1.900	0.019	-0.016
	(6.029)	(6.220)	(5.163)	(0.067)	(0.020)
Mean control	43.273	47.090	45.183	0.070	0.023

Notes: Each point estimate is the treatment effect on the outcome in the column heading within the quartile of the initial practice test. Robust standard errors in parenthesis. For reference, the mean of the control group in the quartile is reported below the standar error. Controls in this 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 feedback on first option major declared

	Switched to harder major	**Switched to easier major**	**Cutoff score first option**	**First option cutoff in top scores**
Q1 = top	-0.010	0.027	0.463	0.054
	(0.097)	(0.070)	(1.652)	(0.074)
Mean control	0.235	0.088	80.396	0.444
Q2	-0.275*	-0.054	-1.512	-0.033
	(0.147)	(0.087)	(2.138)	(0.091)
Mean control	0.429	0.048	79.484	0.426
Q3	0.132	-0.053	0.037	0.101
	(0.138)	(0.095)	(2.092)	(0.096)
Mean control	0.150	0.100	78.918	0.298
Q4 = bottom	-0.235	0.251*	-2.926	-0.077
	(0.192)	(0.131)	(2.371)	(0.110)
Mean control	0.400	0.000	79.702	0.395

Notes: Each point estimate is the treatment effect on the outcome in the column heading within the quartile of the initial practice test. Robust standard errors in parenthesis. For reference, the mean of the control group in the quartile is reported below the standar error. Controls in this 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 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	-1.591*	-1.290
	(0.019)	(0.856)	(0.792)	(0.859)	(0.793)
Mean control	0.926	6.179	5.348	18.782	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

Notes: Each point estimate is the treatment effect on the outcome in the column heading within the quartile of the initial practice test. Robust standard errors in parenthesis. For reference, the mean of the control group in the quartile is reported below the standar error. Controls in this 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 15: Effect of feedback on prior beliefs by gender - reading

	Correct		**Overplace**		**Underplace**	
	Female	Male	Female	Male	Female	Male
Panel A. Students in top quartile in initial practice test						
Treated	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)	
Panel B. Students in quartile 2 in initial practice test						
Treated	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)	
Panel C. Students in quartile 3 in initial practice test						
Treated	-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)	
Panel D. Students in bottom quartile in initial practice test						
Treated	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.

Table 16: Effect of feedback on prior beliefs by gender - math

	Correct		**Overplace**		**Underplace**	
	Female	Male	Female	Male	Female	Male
Panel A. Students in top quartile in initial practice test						
Treated	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)	
Panel B. Students in quartile 2 in initial practice test						
Treated	-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)	
Panel C. Students in quartile 3 in initial practice test						
Treated	-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)	
Panel D. Students in bottom quartile in initial practice test						
Treated	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.

Table 17: Effect of feedback on taking entrance exam over two admission cycles by gender

	Took exam Apr.		**Did not take Apr. & Did not take Sep.**		**Took exam Apr. & Did not take Sep.**	
	Female	Male	Female	Male	Female	Male
Panel A. Students in top quartile in initial practice test						
Treated	-0.075**	-0.030	0.078**	0.029	-0.093	0.054
	(0.037)	(0.029)	(0.037)	(0.028)	(0.098)	(0.121)
Mean control	1.000	1.000	0.000	0.000	0.731	0.690
DiD F vs. M	-0.046		0.050		-0.147	
	(0.046)		(0.046)		(0.155)	
Panel B. Students in quartile 2 in initial practice test						
Treated	-0.074	0.006	0.036	-0.068	0.049	-0.071
	(0.055)	(0.106)	(0.042)	(0.090)	(0.133)	(0.160)
Mean control	1.000	0.870	0.000	0.130	0.571	0.652
DiD F vs. M	-0.080		0.103		0.120	
	(0.119)		(0.099)		(0.208)	
Panel C. Students in quartile 3 in initial practice test						
Treated	0.018	0.021	-0.020	-0.017	0.036	0.025
	(0.034)	(0.014)	(0.034)	(0.014)	(0.129)	(0.174)
Mean control	0.968	1.000	0.032	0.000	0.516	0.647
DiD F vs. M	-0.003		-0.004		0.010	
	(0.037)		(0.036)		(0.218)	
Panel D. Students in bottom quartile in initial practice test						
Treated	-0.119	-0.089	0.124	0.089	-0.166	-0.185
	(0.079)	(0.071)	(0.079)	(0.069)	(0.164)	(0.163)
Mean control	1.000	1.000	0.000	0.000	0.565	0.800
DiD F vs. M	-0.030		0.035		0.019	
	(0.103)		(0.101)		(0.230)	

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 18: Effect of feedback on entrance exam scores by gender

	Math score		**Reading score**		**Total score**	
	Female	Male	Female	Male	Female	Male
Panel A. Students in top quartile in initial practice test						
Treated	4.326	-2.653	-0.184	-7.175*	2.312	-5.046
	(4.130)	(5.071)	(3.375)	(4.329)	(2.982)	(3.799)
Mean control	68.351	75.878	72.129	76.884	70.070	76.380
DiD F vs. M	6.980		6.991		7.358	
	(6.576)		(5.540)		(4.879)	
Panel B. Students in quartile 2 in initial practice test						
Treated	-1.472	2.566	-3.077	-0.083	-2.230	1.396
	(6.251)	(7.065)	(6.162)	(6.425)	(4.759)	(4.679)
Mean control	54.600	71.777	64.042	61.545	59.322	66.503
DiD F vs. M	-4.038		-2.994		-3.625	
	(9.451)		(8.789)		(6.669)	
Panel C. Students in quartile 3 in initial practice test						
Treated	-1.081	4.936	-7.601	-8.195	-6.773	-2.477
	(6.007)	(8.581)	(6.344)	(8.143)	(5.189)	(7.298)
Mean control	49.169	52.994	53.221	54.098	52.694	54.420
DiD F vs. M	-6.017		0.594		-4.296	
	(10.523)		(10.345)		(8.969)	
Panel D. Students in bottom quartile in initial practice test						
Treated	-7.425	8.013	2.557	5.294	-2.477	6.812
	(7.990)	(8.729)	(7.834)	(10.033)	(6.653)	(8.033)
Mean control	44.878	41.427	47.402	46.731	46.141	44.082
DiD F vs. M	-15.438		-2.737		-9.290	
	(11.584)		(12.721)		(10.331)	

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 19: Effect of feedback on admissions by gender

	Admitted to first option		**Admitted to second option**	
	Female	Male	Female	Male
Panel A. Students in top quartile in initial practice test				
Treated	0.023	-0.193	0.003	0.053
	(0.090)	(0.119)	(0.033)	(0.059)
Mean control	0.250	0.414	0.019	0.034
DiD F vs. M	0.215		-0.051	
	(0.148)		(0.067)	
Panel B. Students in quartile 2 in initial practice test				
Treated	0.177*	0.015	-0.031	-0.056
	(0.097)	(0.157)	(0.034)	(0.053)
Mean control	0.057	0.263	0.029	0.053
DiD F vs. M	0.162		0.024	
	(0.185)		(0.059)	
Panel C. Students in quartile 3 in initial practice test				
Treated	-0.016	0.072	-0.050	0.069
	(0.051)	(0.110)	(0.048)	(0.070)
Mean control	0.033	0.059	0.067	0.000
DiD F vs. M	-0.088		-0.118	
	(0.122)		(0.085)	
Panel D. Students in bottom quartile in initial practice test				
Treated	0.017	0.016	-0.039	0.009
	(0.086)	(0.119)	(0.034)	(0.015)
Mean control	0.043	0.100	0.043	0.000
DiD F vs. M	0.002		-0.048	
	(0.155)		(0.038)	

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 20: Effect of feedback on majors declared by gender

	Switched to harder major		**Switched to easier major**		**Cutoff score 1st option**		**First option cutoff in top scores**	
	Female	Male	Female	Male	Female	Male	Female	Male
Panel A. Students in top quartile in initial practice test								
Treated	0.059	-0.171	-0.048	0.162	-1.226	3.008	-0.035	0.184
	(0.120)	(0.180)	(0.087)	(0.105)	(2.019)	(2.797)	(0.096)	(0.117)
Mean control	0.130	0.455	0.130	0.000	81.414	78.571	0.481	0.379
DiD F vs. M	0.230		-0.210		-4.233		-0.219	
	(0.219)		(0.132)		(3.428)		(0.151)	
Panel B. Students in quartile 2 in initial practice test								
Treated	-0.315	-0.207	0.050	-0.229*	-2.019	-0.608	-0.000	-0.085
	(0.192)	(0.253)	(0.098)	(0.131)	(2.945)	(3.023)	(0.119)	(0.140)
Mean control	0.385	0.500	0.000	0.125	80.085	78.377	0.429	0.421
DiD F vs. M	-0.108		0.279*		-1.411		0.085	
	(0.322)		(0.144)		(4.223)		(0.185)	
Panel C. Students in quartile 3 in initial practice test								
Treated	0.117	0.132	-0.088	-0.025	-0.775	1.798	0.136	0.049
	(0.181)	(0.228)	(0.144)	(0.044)	(2.637)	(3.578)	(0.121)	(0.162)
Mean control	0.167	0.125	0.167	0.000	79.674	77.584	0.300	0.294
DiD F vs. M	-0.015		-0.063		-2.574		0.087	
	(0.298)		(0.146)		(4.497)		(0.203)	
Panel D. Students in bottom quartile in initial practice test								
Treated	-0.535**	0.158	0.195	0.396	-0.970	-5.272	0.088	-0.280*
	(0.221)	(0.329)	(0.137)	(0.296)	(3.253)	(3.519)	(0.147)	(0.148)
Mean control	0.625	0.143	0.000	0.000	79.819	79.567	0.391	0.400
DiD F vs. M	-0.694*		-0.201		4.302		0.368*	
	(0.401)		(0.325)		(4.885)		(0.205)	

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 21: Effect of feedback on academic investments by gender

	Takes practice tests		**Math study hours**		**Reading study hours**	
	Female	Male	Female	Male	Female	Male
Panel A. Students in top quartile in initial practice test						
Treated	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)	
Panel B. Students in quartile 2 in initial practice test						
Treated	-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)	
Panel C. Students in quartile 3 in initial practice test						
Treated	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)	
Panel D. Students in bottom quartile in initial practice test						
Treated	-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 22: Effect of feedback on performance in practice tests by gender

	Math correct answers		**Reading correct answers**	
	Female	Male	Female	Male
Panel A. Students in top quartile in initial practice test				
Treated	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)	
Panel B. Students in quartile 2 in initial practice test				
Treated	-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)	
Panel C. Students in quartile 3 in initial practice test				
Treated	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)	
Panel D. Students in bottom quartile in initial practice test				
Treated	-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.

A Behavioral theories explaining biases in beliefs

My design allows to test different theories that have been posited to explain why individuals do not update like a Bayesian agent when provided a "signal" about their ability. The analysis I present is similar to Burks et al. (2013), who conduct and test different overconfidence theories using IQ and numeracy tests among trainee truck drivers. I use as a task the weekly practice tests that the test preparation institute provides to students. By design, I exclude explanations of biases related to social signalling because the belief elicitation and feedback provision are completely private and this is made explicit to students.

My setting has at least two main advantages relative to Burks et al. (2013). First, the task is relevant to the context that I am studying. The information students obtain by taking practice tests and the feedback I provide is useful for the college entrance exam they are preparing for. IQ and numeracy tests may not be that useful during of after the truck driving training. In this sense, the task in Burks et al. (2013) is more related to a task typically used in the lab with low stakes and external validity. Second, I collect multiple rounds of data as opposed to a single elicitaton.

I first consider how far students are from the Bayesan benchmark by computing the posteriors a Bayesian agent would have based on the students' reported priors. The basic question that papers in this literature try to answer is whether overconfidence is the result of biased information processing regarding own skills (Mobius et al., 2011; Ertac, 2011; Eil & Rao, 2011; Grossman & Owens, 2012; Berlin & Dargnies, 2016). To test this, the authors elicit performance beliefs, then provide and informative signal (e.g., whether the subject's score is at the top or bottom 20 percent of scores of participants in the same session), and then re-elicit beliefs. The main findings in this literature is that people update far from the Bayesian benchmark, that is, are conservative, and update more if the signal places them at the top rather than at the bottom of the distribution, a phenomenon known as asymmetry. Interestingly, when the updating task is self-relevant, the distance to the Bayesian benchmark is bigger than when the task is not related to their ego (Ertac, 2011).

Figure 15 shows how much students update their reading beliefs at the posterior stage relative to the Bayesian benchmark across all lab-in-the-field rounds (results for math beliefs are similar). Relative updating is calculated as average of the ratio of how many tokens students assign to each quartile over what a Bayesian with the same priors would assing to the same quartiles. Recall that the posteriors are collected when students receive the performance report a few days after the practice test and that control students only receive absolute scores while treated students receive a signal indicating whether their score was above or below the median. The left panel in the figure shows how much control students update relative to a Bayesian whe their scores were below or above the median. The right panel shows the same but in this case the students actually see the signal indicating whether they are below or above the median. The figures show 83 percent confidence intervals based on regressions that cluster the standard errors at the individual level. The confidence intervals allow to test whether the height of the two bars is statistically different from each other as opposed to a 95 percent confidence interval, which tests whether the height is different from zero.

I find support for the findings of conservatism and asymmetry outside of the lab environment with a real-stakes task. First, students in the control group update between 52 and 70 percent of what a Bayesian would update. In the treatment group this is larger, as expected, with students updating between 60 and 76 percent relative to the Bayesian benchmark. Updating among these students is larger relative to what other papers find in the lab. For example, subjects in Mobius et al. (2011) update about 35 percent of what a Bayesian would update.[15] Second, students update significantly more when their scores are above the median than when they are below, which is consistent with asymmetry.

Conservatism holds across all quartiles while asymmetry varies by quartile. Figure 16 shows the percentage of updating split by quartile in the initial practice test for treated students only. The Figure shows that students in the top quartile (Q1) update most, and students in quartiles 2 and 4 update the least at around 60 percent of what a Bayesian would update. Asymmetry is not clear in these two quartiles as students update about the same when receiving a below or above median signal. Asymmetric updating is more clear in quartiles 1 and 3, although only statistically different in quartile 3.

Having found evidence that individuals do not update like Bayesians, I turn to test wether the theory proposed by Köszegi (2006) holds in my data. This is an ego-utility model that predicts that individuals will be overconfident because if they initially obtain signals that they are of the high type, they will stop collecting more signals because they my revise they belief down by error. Similarly, if the initial signals they receive tell them that they are of the low-type, they will keep looking for signals under the expectation that they will revise their belief up.

I do not find evidence that the Köszegi (2006) model holds in the data.[16] Figure 17 shows what fraction of the performance reports students check by the quartile to which they assign the highest number of tokens in reading. The left panel shows students in the treatment group and the right panel shoes treated students. For the Köszegi (2006) model to be true we

[15] This could be related to differences in the measure of updating. Mobius et al. (2011) compute logit beliefs while I compute the ration relative to the Bayesian posteriors.

[16] Burks et al. (2013) also do not find evidence for this theory in the truck trainees data.

would see that students who think they are in the bottom quartile check more performance reports while students at the top check fewer. We see that students in the treatment group behave in the opposite way as was found by Burks et al. (2013).

Finally, I present evidence that the data conform with confirmatory bias in Figure 18. Rabin and Schrag (1999) proposes 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. The figure shows, for reading, how much individuals update relative to a Bayesian in four scenarios. In the left panel, the left bar shows how much students update when they are above the median and that confirms their belief. The right bar shows a disconfirming signal, that is, they though 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 though they were above. In both cases students update more when no news rather than when disconfirming signals are received as the model predicts.

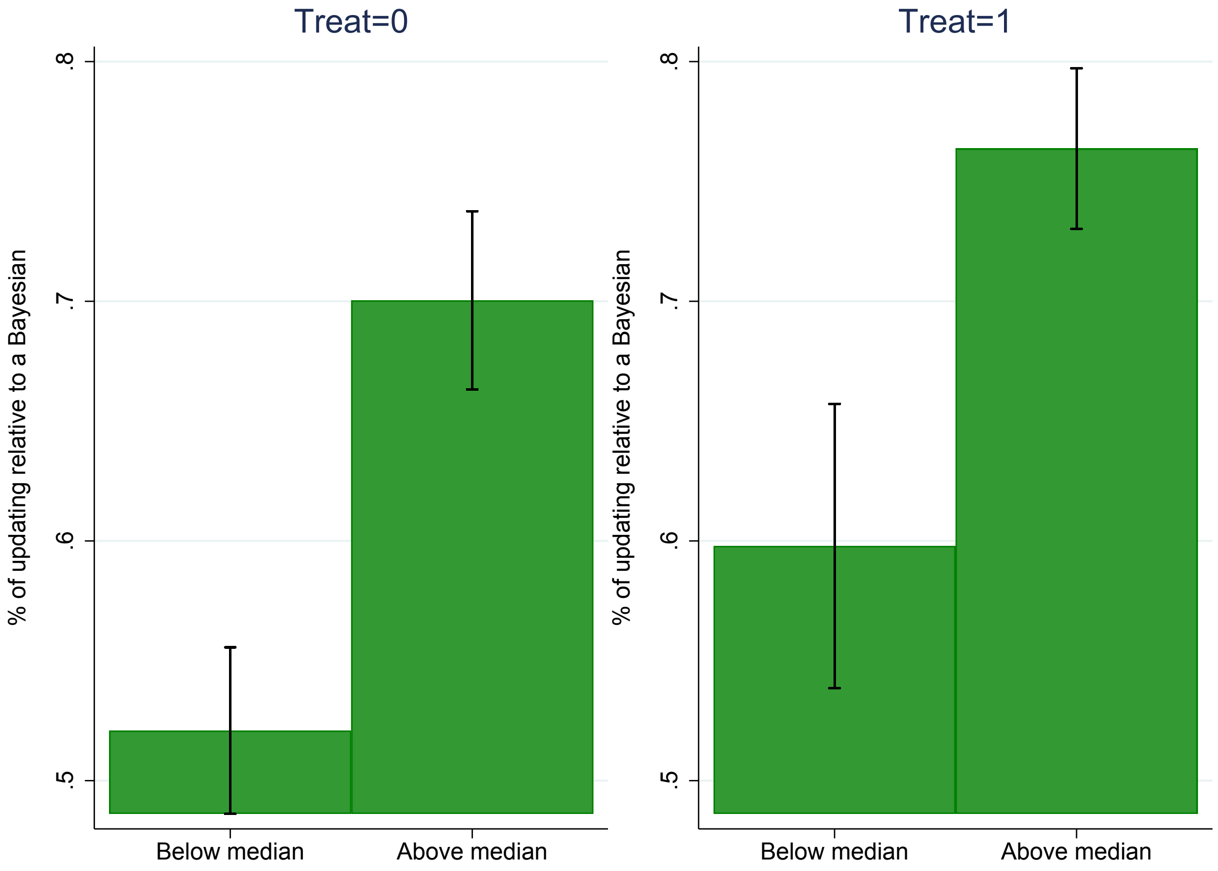


Figure 15: Updating after receiving performance report relative to the Bayesian benchmark - reading

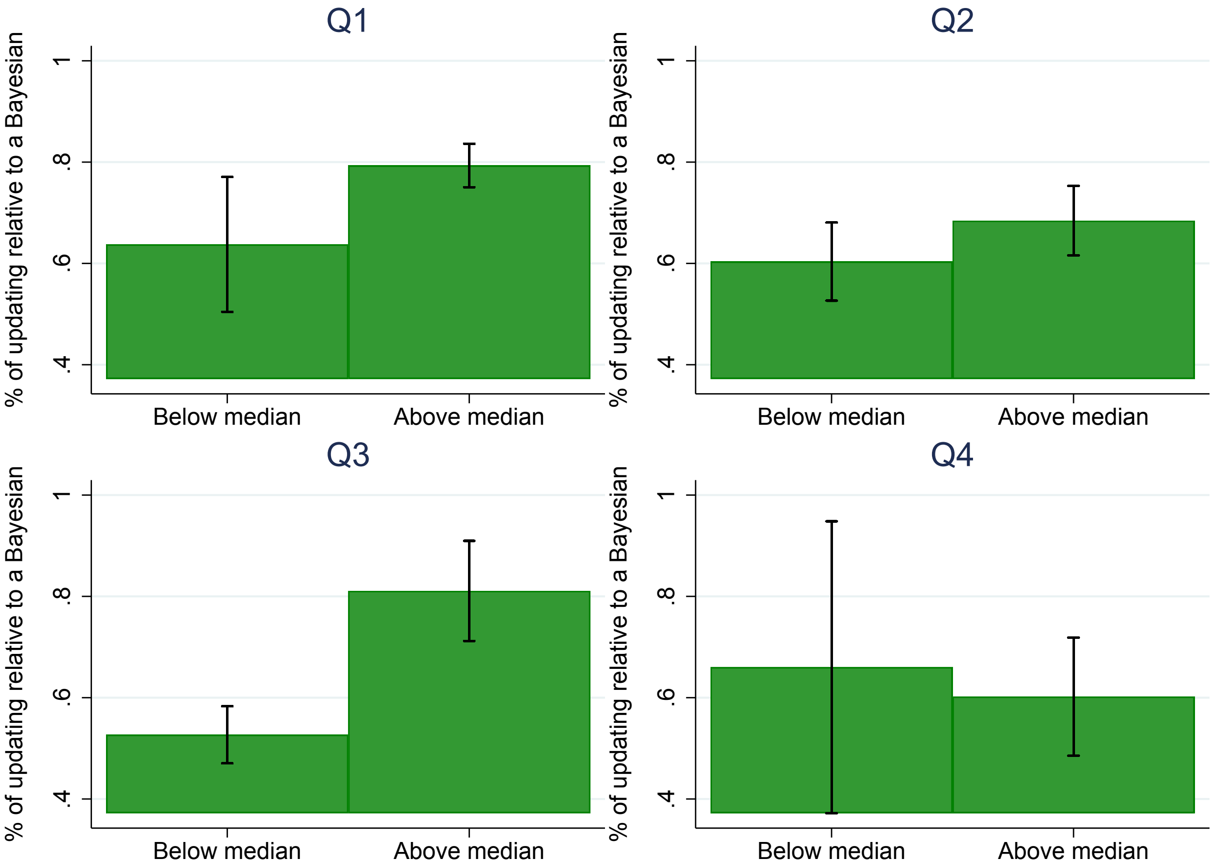


Figure 16: Updating after receiving performance report relative to the Bayesian benchmark - reading

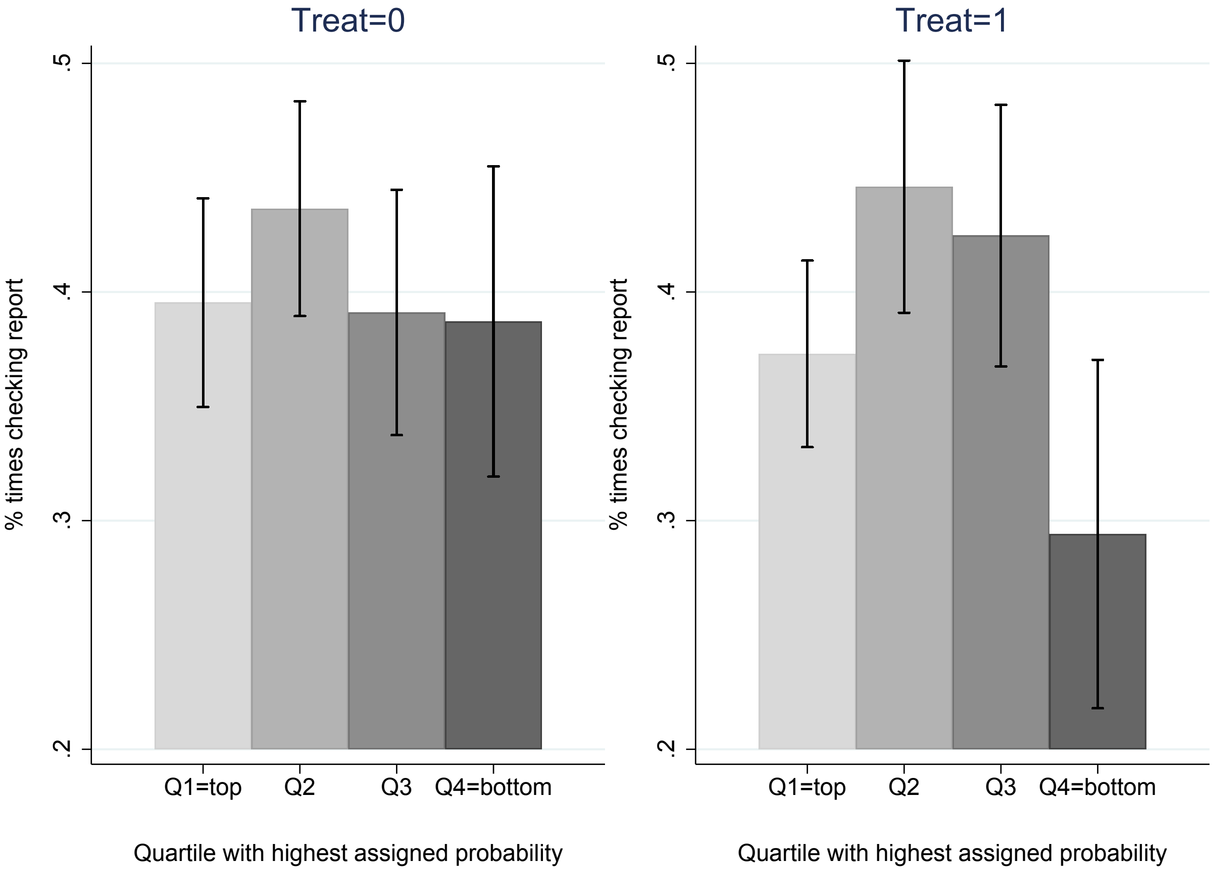


Figure 17: Information search by quartile with highest assigned probability - reading

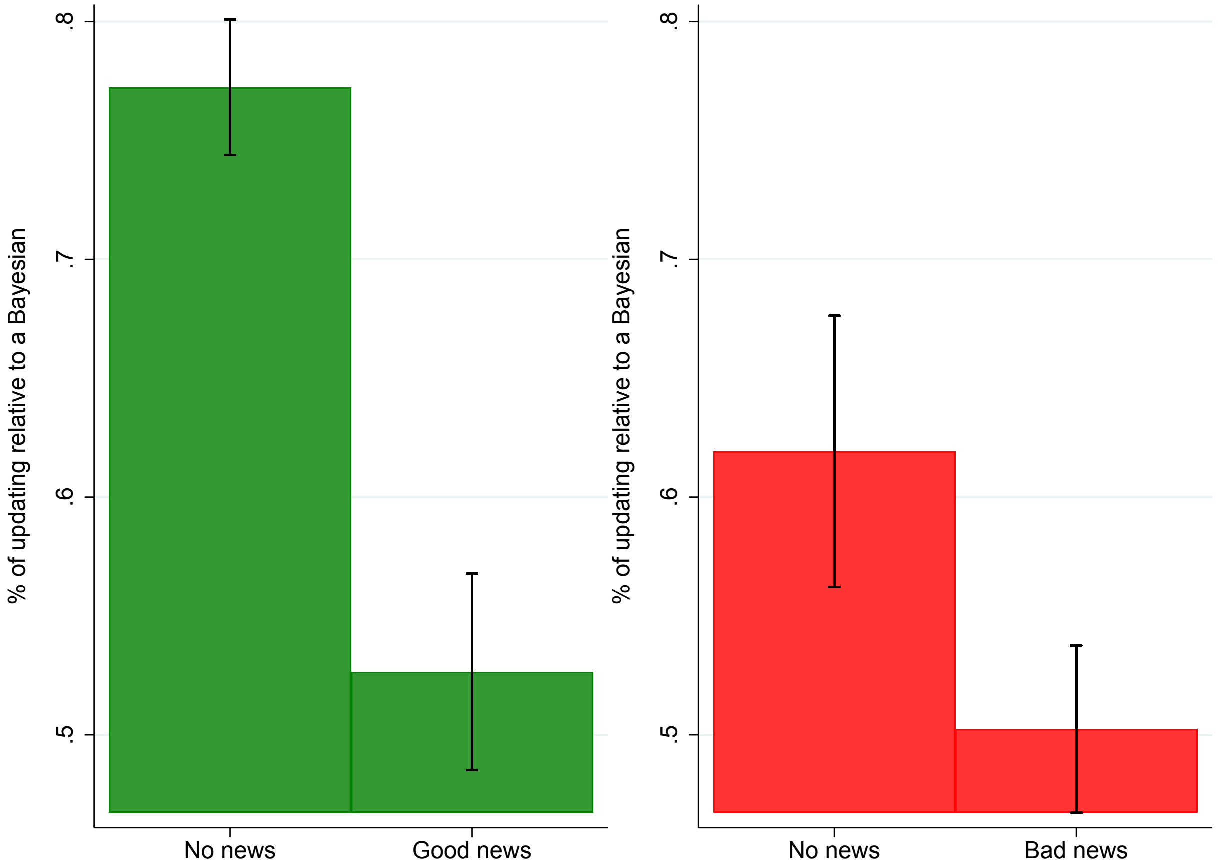


Figure 18: Confirmatory bias in reading