FACULTAD DE CIENCIAS EMPRESARIALES Y ECONOMIA

Serie de documentos de trabajo del Departamento de Economía / Department of Economics Working Papers Series

Gender Differences to Relative Performance Feedback: A Field Experiment in Education

José María Cabrera
Universidad de Montevideo

Alejandro Cid
Universidad de Montevideo

February 14th, 2017

Gender Differences to Relative Performance Feedback:

A Field Experiment in Education

José María Cabrera* and Alejandro Cid**

February 14th, 2017

PRELIMINARY

Abstract

Individuals care about both their absolute performance and their performance relative to others. For example, workers satisfaction is affected not only by their nominal wage but also by the comparison of their salaries relative to colleagues. We analyze the effect of providing relative performance feedback using a field experiment with university students. Untreated students misplace themselves in the grade distribution. Poor performing students over report their placement (they say that they have a better position in the classroom ranking than they actually have). On the other hand, good students (especially women) under place themselves: they report that they don't perform as well as they actually do. We experimentally change the information that treated students have, so they know exactly how they perform relative to their peers. We find that the information feedback has asymmetric effects for men and women. Treated men report higher satisfaction with their GPA while treated women report less satisfaction, regardless of their position in the grade distribution. We also show that this non-monetary incentive caused a decrease in women academic performance. Two possible channels may explain our results: women may shy away from competition and they face an increasing marginal cost of effort. More information is not always beneficial for everybody.

Keywords: ranking; field experiment; overconfidence; education

* Erasmus School of Economics and Universidad de Montevideo (jmcabrera@um.edu.uy).

** Universidad de Montevideo (acid@um.edu.uy).

We thank seminar participants at Universidad de Montevideo, Erasmus School of Economics and Encuentro SEU 2016 for helpful comments.

I. Introduction

Researchers in Social Sciences have long been interested in the possibility that individuals care about both their absolute performance and their performance relative to others. This issue is present in salaries comparisons at the labor market (Card et al., 2012), and in the grade ranking in education (Azmat et al., 2016; Bursztyn, 2015; Tran and Zeckhauser, 2009).

By a randomized control experiment in the field, we study the effect of providing 1,048 undergraduate students with information of their relative position in the distribution of grades. Our main outcomes are students' satisfaction and educational outcomes (after one and two years). Treated students received feedback on their exact placement within their peers: an ordinal ranking. A treated student could learn, for example, that his GPA places him in the 9^{th} position out of 120 classmates, information which he didn´t have before. We study how males and females response to the competitive incentives that the ranking creates.

A first result is that untreated students misplace themselves in the grade distribution. Poor performing students over-place: they report a better position in the ranking that their actual performance. On the other hand, students in the upper part of the grade distribution under-place: good students (especially women) tend report that they perform worse than they actually do). Treated students report a more accurate position in the ranking. The treatment gave them information which they didn't have before. Our next step is to see the impact of the new information on satisfaction and academic performance.

We find asymmetric gender responses to the information on the personal position in the ranking. While treated men increased their reported satisfaction, female students in the treatment group report lower satisfaction. Moreover, treated women seem to have decreased their academic performance. They score less in their exams (especially in the short run), they take less exams and approve less courses. More information is not always beneficial for everybody.

The remainder of the paper is organized as follows. Section II states the conceptual framework. Section III describes the intervention, the experimental design, and our data collection. Section IV presents our main empirical results. Section V concludes. Supplementary results are gathered in the Appendix.

II. Conceptual Framework

Satisfaction and Ranking

This paper builds on previous papers that have empirically examined the relationship between relative position and satisfaction. Frey and Stutzer (2002) provide an excellent review of this literature. A first reason why information on peers' rewards may affect utility is individuals care directly about their relative rewards. Luttmer (2005) investigates whether individuals feel worse off when others around them earn more. He found that, controlling for an individual's own income, higher earnings of neighbors are associated with lower levels of self-reported happiness. Card et al. (2012) have documented the effect of peer salaries on job satisfaction. By an experimental design applied in the University of California (they randomly disclosed peers' salaries), they find that job satisfaction depends on relative pay comparisons.

Also, people may react to new information on peer rewards even if they do not care directly about relative position. In particular, it is possible that students have no direct concern over peer position in the grade distribution, but rationally use this information to update their future pay prospects. Relative position in the grade distribution may provide a signal about future wages.

Gender and Ranking

Theory suggests that heterogeneous effects by gender may be found if information about relative position in the grade distribution is disclosed to the students. Thanks to advances in the psychology and experimental literatures, we now have a much more concrete sense of psychological factors that appear to systematically differ between men and women. While there is an abundance of laboratory studies regarding each of these psychological factors, there has not been to date only a very limited amount of research on the relevance of these factors in real environment outcomes.

Bertrand (2011) reviews the evidence regarding gender differences in risk preferences and in attitudes toward competition. Croson and Gneezy (2009) and Eckel and Grossman (2008) conclude that women are more risk averse than men. Also, recent experimental papers argue that women are underrepresented in competitive environment because they prefer to stay away from such environments. This risk and competitive aversion may prevent women to disclose the relative position in the grade distribution at the education centre.

Overconfidence and Ranking

In the words of Moore and Healy (2008), "Overconfidence can have serious consequences. Researchers have offered overconfidence as an explanation for wars, strikes, litigation, entrepreneurial failures …", and -we could add- educational failures. In their article, they analyse different types of overconfidence. One of them is of particular interest for the present research: the overplacement of one's performance relative to others, that is, people that believes themselves to be better than others. Benoît and Dubra (2011) discuss the rationality behind these empirical regularities.

People often have imperfect information about their own performances, abilities, or chance of success. And they may have even worse information about others. When performance is high, will the student underestimate their own performance? Or will she underestimate others even so? And what happens when the performance is low? Thus, as far as we know, there's no piece of research that evaluate, by a field experiment, the heterogeneity by performance of self-estimation relative to others' in college education. We try to contribute with some findings also in this field.

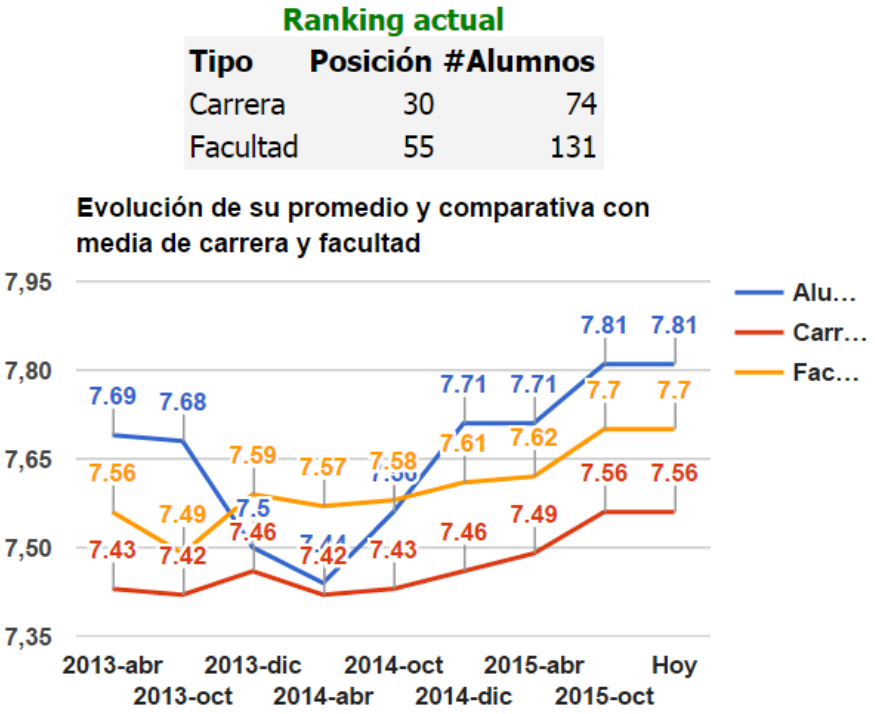
III. Data and Experimental Design

The experiment

In mid-2013 we decided to conduct an experiment to measure the reactions of students to the availability of information of their position in the grade distribution: an ordinal ranking. It was a tool being developed by IT team at the University. We proposed to run a pilot to test the impact on the students and we got involved in the development of the treatment. We focused on three Schools at Universidad de Montevideo: Economics, Engineering, and Law. The evaluation was performed using a randomized control trial. Treated students could start using a new platform in the intranet were they would see their ranking relative to their peers. Control students would not see this new information. Treated and control students could access the official transcript of grades, as they did before the experiment was launched. Appendix 1 shows the personal transcript of grades. It includes all the courses taken by the student and the grade achieved in each one. On the bottom right of the transcript the student can see their grade point average (GPA), but with this information they don't know where they are placed in the distribution of grades, relative to their peers.

The treatment consists in placing that GPA in the context of their peers. Treated students could access the new tool with the ranking information on their relative performance. Figure 1 shows a real example of the treatment.

Figure 1



The information treatment was designed to be visual and easy to understand. It includes a comparison with their peers about the current GPA at the cohort+major level (average size of 37 students) and at the cohort+school level (average size of 84 students). It also includes figures with the evolution of the ranking of their GPA across the semesters. Students could learn if they are improving or getting worse over the semesters (they already knew if their absolute GPA was improving or not, but not how they were evolving relative to their peers). In the case depicted in the example of Figure 1, the student has a current ranking of 30 out of 74 students in his cohort+major. He also learns that he is placed in the 55th position out of 131 students in his cohort+school. We also provided information of the evolution of his GPA over the years, relative to his peers; and

the evolution of the position in the ranking. In Appendix 2 we show two other real cases of treated students who had different evolutions over the semesters. The ranking was recalculated and updated on a daily basis. We think that with the access to this new information tool a student could get a very informative picture of their relative performance.

We used several channels to make students aware of the launching of this new tool and try to increase awareness and access. We placed banners in the personal intranet, which students access almost daily for course materials and administrative tasks. We also sent an email informing about the new tool (Appendix 3). This email was also sent to control students but differing only in the last line (a link to the ranking for treated students). The reason to send a placebo email to control students was to disentangle the effect of receiving an email from the university (which may lead to an higher use of the existing intranet resources: announcements, dates of exams, etc) from the effect of the treatment itself. With this placebo email we are sure that the effect comes from the use of the ranking tool and not by a higher access to the university intranet, or other effects that may arise upon receiving an email from the university staff.

Randomization

We have 1,048 students participating in this field experiment. The treatment group consists of 529 students (50.5%) and the control group of 519 students (49.5%). We wanted to have balance across several important dimensions (quality of students, sex, cohort, etc). For doing this, we constructed 300 random assignments of students to treated and control groups, were balance was achieved. Among this 300 assignments which we were comfortable with, we selected one with a randomly generated number. Table 1 shows the balancing condition among several pre-treatment characteristics. In the control group there is a 44% of women, they have taken an average of 25.55 courses at the beginning of the experiment, approving on average 21.32 courses, which lead them to 138.39 credits. The cumulative GPA is 7.54 (in a 1 to 12 scale). There is a small but statistically significant difference in the number of degrees a student is attending. Treated students attend slightly more degrees than control students. This was loophole in computer system, which we didn't know beforehand and was explained to us by the IT team afterwards. If a student was registered in two majors and he was placed in the Treated group in one major (i.e. economics), the he could also access his ranking in the second major (i.e. accountancy). Nonetheless, very few students register for two majors:

9% in the treated group and 5% in the control group. In the regressions we will control for this unbalanced pre-treatment characteristic. We also construct three variables which we knew beforehand that were linked to academic performance (and supposedly also to satisfaction): the top three high schools where the 29% of the students come from (27% in the control and 30% in the treated group), the proportion of students who come from Montevideo (the capital city of the country), and the proportion of students with a scholarship larger than 20% of the tuition fee. We also balanced in the cohort (year when the student entered the university). We have students from cohorts from 2008 to 2014. Since the experiment started in 2014, students in cohort 2014 are freshmen. They are the 21% of the sample (and for them we don`t have cumulative GPA information). On the other end of the cohort distribution we have students from the 2008 cohort (we have excluded previous generations from the experiment). These older students represent only the 4% of the sample and are students who have lagged behind (they are starting their seventh year at the university and should have graduated if they had done their major on track). We have also balanced major and by decile of the GPA distribution, so for example there are as many good students in the control as in the treatment arm of the experiment[1].

[1] This descriptives are omitted for the sake of brevity, but are available upon request.

Table 1 –Descriptive Statistics

	(1)	(2)	(3)	(4)	(5)	(6)
	Treated	Control	difference (1-2)	Standard Error	p-value	Obs.
Student characteristics						
1 *Female*	0.47	0.44	0.03	-0.03	0.40	1048
2 *Courses*	26.50	25.55	0.95	-1.12	0.40	1048
3 *Cumulative GPA*	7.50	7.54	-0.04	-0.10	0.70	834
4 *Approved courses*	21.79	21.32	0.47	-0.99	0.63	1048
5 *Credits earned*	142.00	138.39	3.61	-6.32	0.57	1048
6 *Number of degrees*	1.09	1.05	0.04	-0.02	0.02 **	1048
7 *School of origin*	0.30	0.27	0.03	-0.03	0.24	1048
8 *Montevideo*	0.67	0.65	0.02	-0.03	0.42	1048
9 *Large Scholarship*	0.23	0.24	-0.01	-0.03	0.70	1048
10 *Cohort 2008*	0.04	0.04	0.00	-0.01	0.83	1048
11 *Cohort 2009*	0.05	0.05	0.00	-0.01	0.94	1048
12 *Cohort 2010*	0.13	0.12	0.01	-0.02	0.66	1048
13 *Cohort 2011*	0.20	0.20	0.00	-0.02	0.94	1048
14 *Cohort 2012*	0.20	0.20	0.00	-0.02	1.00	1048
15 *Cohort 2013*	0.17	0.16	0.01	-0.02	0.55	1048
16 *Cohort 2014*	0.20	0.22	-0.02	-0.03	0.44	1048

The difference in means is calculated with an OLS regression with robust standard errors. *** p<0.01, ** p<0.05, * p<0.1. From the 1048 students, there are 529 (50.5%) in the treated group and 519 (49.5%) in the control group. Balance was also performed by schools, degrees and place in the distribution of grades at the major and school level. These results are omited from this table to ease the display of the main results.

Data Collection

We have two main outcomes: satisfaction with GPA and academic performance. Satisfaction data comes from a short term web survey to treated and control students, implemented 12 days after the ranking tool was launched. Academic performance is obtained from administrative records (grades, credits achieved, etc) in the longer run: one and two years after treatment.

The short term survey was answered by 861 students out of 1,048 participants (response rate of 82%). This high response rate was achieved by strongly publicizing the survey and by temporarily blocking the access to some intranet resources for students who did not start answering the survey (this blocking policy is a standard practice for

student surveys at the university for surveys, for example for the evaluation of courses and lecturers and other surveys[2]). It is important to note that the survey was in no way directly related to the ranking experiment. It was presented as a "satisfaction survey" and was sent from the Bedele´s office.

There was no statistically significant difference in the response rate by treatment status (83.7% in the Treated group and 80.5% in the Control group). Moreover, there were no differences by gender (which will be the main explanatory variable in the analysis). Appendix 4 shows that, as one might expect, the survey was answered in a greater proportion by better (more responsible) students: those with higher cumulative GPA and who are placed in higher deciles of the GPA distribution.

Although attrition in the survey was -fortunately- not correlated with treatment status nor gender, we also have to check that the balance is still respected in other characteristics for the students who answered the survey. Attrition may have the same proportion between treated and control groups, but for different reasons. It could happen that attrition is balanced but the remaining students have different characteristics. Appendix 5 shows that, even after attrition, balance is respected. So we can proceed with the analysis of satisfaction outcomes using the random variation in treatment status generated by the experiment.

Finally, attrition is not a concern in the administrative database (grades, exams, dropouts, etc), since we have the academic outcomes for all the 1,048 students in the experiment.

Were students really treated?

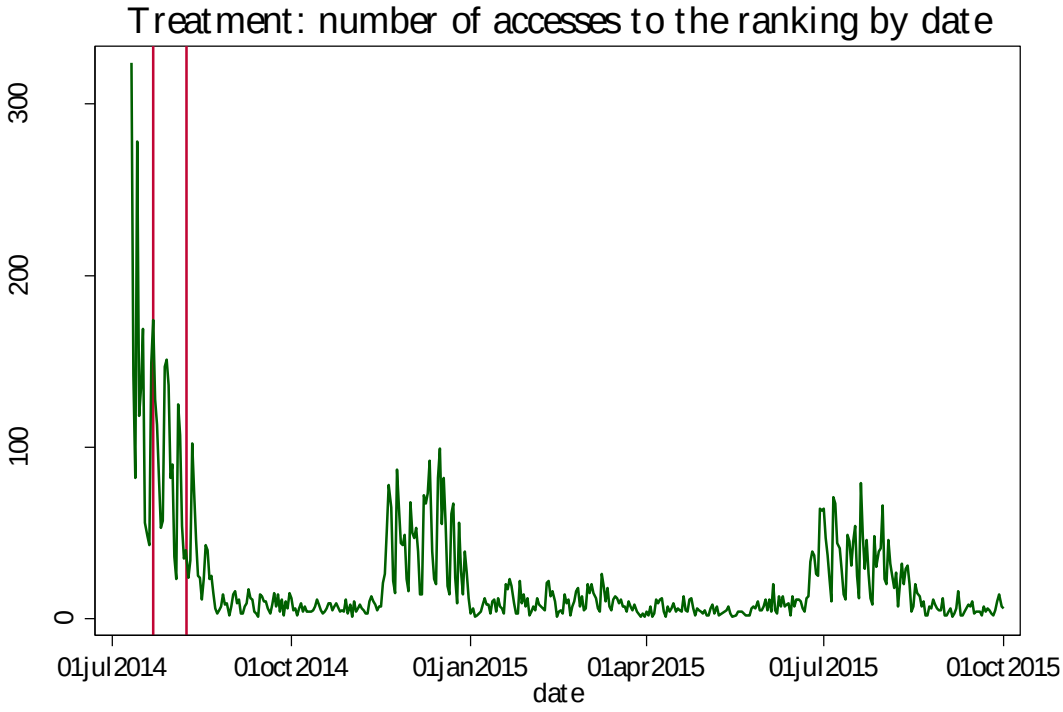
Figure 2 displays information about the number of accesses to the ranking system. The busiest day was when we launched the system for treated students: emails were sent, banners were placed in the individual intranet, and the button to access the ranking system was activated. There were 326 accesses on July 11th, 2014. The pattern, with peaks and valleys, is given by the low access during weekends and the surge on Mondays (when students can use the PCs at the university or at their work places). The pattern of access to the system is very similar to the one we observe of accesses to the official

[2] Our blocking policy for the intranet resources was more benign than other blockings from administrative staff. We implemented the block mainly as a clear notice that there was a survey expected to be answered (since some students don´t read emails). If a student didn´t want to answer this survey, he could opt-out and access his intranet web site with no further delay.

transcript: it is very pronounced during the main exams periods (December and July). During the rest of the academic year students access the ranking, but with less intensity. The reason is that when the exams periods ends, no more grades are added to the official transcript, so the ranking doesn´t change, and incentives to access the system decrease.

The two vertical lines show the time window for the survey about satisfaction with the grades (the fists outcome measure). Appendix 6 shows a zoom to that period for greater detail.

Figure 2



The system was designed to keep a record of the exact use that each student gave to the ranking information. For example, in the period from July 11th 2014 to August 31st 2015, students saw their ranking a total of 9,625 times.

The average student entered the system 18.2 times (min = 0, max = 430). From a total of 529 students in the treatment group, 508 accessed the ranking at least one time before September 2015[3]. Compliance with the assignment to treatment was high. From

[3] Other access statistics, available upon request, show that there are no differences between men and women and that students with a higher GPA accessed more times. Students closer to graduation (with more credits earned) used the system fewer times. Until September 2016 we have registered 14,298 accesses to the system. During the two years of experiment the use of the system slightly reduced since some students were graduating and therefore left the University.

529 students in the treatment group, 508 entered the system. And from 519 students in the control group, 518 didn´t enter the ranking (an exception was made by the IT team for one student in the control group who requested a special access). So in the main results tables we will show intention to treat effects from reduced form models, using the initial assignment to treatment. Instrumental variables estimates, using randomization as an instrument for actual treatment, are very similar to ITT, given the high take-up rate, and are available upon request.

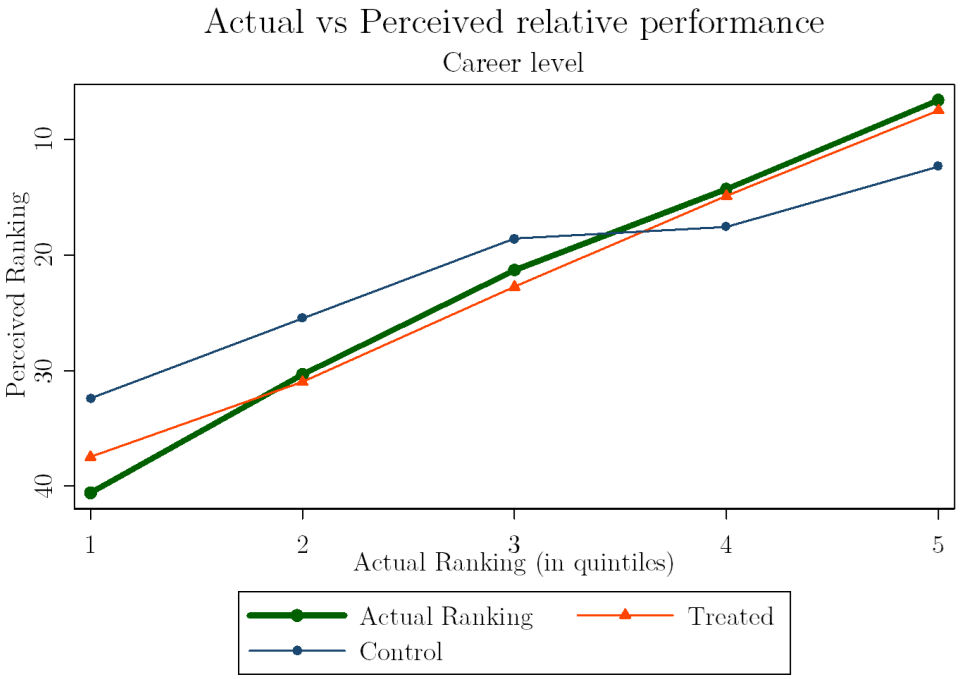
So the new tool with the ranking information was highly used by treated students. Our new step is to show that the ranking provided students with new information.

Did students change their perceptions?

In the previous section we show that the ranking was used by students. Now we will show that treated students have a much more accurate picture of their real placement in the distribution of grades. That is, treatment was successful to increase a students awareness of his relative performance. A major challenge to the experiment would have been that students already had an accurate picture not only of their current GPA but also about their relative performance. Indeed, since cohorts at the university are small, a student may have good information about the academic performance of his peers, and thus of his relative performance. Moreover, once the ranking system was implemented, control students may have gotten some information on their relative performance from treated friends. For example, a student from the control group may not know if he is placed in the top 10% of the cohort. But if he has a friend with a similar GPA in the treatment arm of the experiment, when his friend knows his ranking and communicates this information to him, this student in the control group would know if he is also in the top 10%. So *prior accurate knowledge* and *contamination* could be two major challenges to the experiment.

In the Satisfaction Survey we included a question to test if the ranking provides new information to students. We asked treated and control students what was their placement in the distribution of grades (ranking), being #1 the student with the higher GPA in the cohort (and major or school). Figure 1 plots the actual (objective) ranking versus the perceived (subjective) ranking.

Figure 3



There are three plotted lines. The first one is the actual ranking (in quintiles) plotted against the actual ranking (in individual positions); it is like a 45º line. The other two lines are one for control students and other for treated students (regardless of whether they accessed the system or not). A student located at the right end of the x-axis is a very good student. If he places himself correctly in the ranking, he should report being among the top 10 students (y-axis). We find that treated students report their actual ranking very accurately. They report a placement in the ranking very similar to the actual one. If they had not received the information treatment they would have behaved like students in the control group. And control students misplace themselves in the distribution of grades. It is very interesting to note that high achieving control students (at the right of the figure) under-place themselves: they report a perceived position in the ranking that is bellow how they actually perform. On the other hand, underperforming control students (at the left of the figure) over-place themselves in the distribution of grades: they report a perceived relative performance higher than the actual one.

We propose three hypothesis for control students miss-placing in the ranking:

(i) *statistical inference problem and selection*: students select into groups of friends and then observe grades of their closer peers to infer the entire distribution of grades.

(ii) *regression towards the mean*: a student knows that he has performed well in an exam, but doesn´t know if it is because he is good (and thus above his peers) or because the test was easy (and other students also performed well), so he reduces his expected placement in the distribution of grades.

(iii) *cognitive biases*: students may have the necessary information, but fail to use it correctly.

We have data that allows us to test the first hypothesis. In the satisfaction survey we have directly asked students to name their best friends in their cohort. We have 785 students who answered this question and provided information on 3.389 friends. This data allows us to construct networks of friends. We are able to show that peer groups are not formed randomly. Good students are friends of other good students. Figure 7 in the appendix shows that a student GPA is correlated with his friends GPA: one more point of the friends grades is associated with an increase of 0.77 in a student grades (t-value=13.9). This means that there is a strong selection process. Now we will look at the inference problem. If a student doesn´t know his placement in the grades distribution, he must infer it from the comparison with his closets peers. Since peers are positively selected, a good student will have high performing peers, so, on average, he will be placed lower in the ranking of his peers than if he had a random group of peers (representative of the whole distribution of grades). In the data, a student in the 5^{th} quintile of the grades distribution (a good student) has friends with a higher GPA (8.3) than the mean of the whole cohort (7.8), so if he constructs his perceived ranking with the information at hand, he will think that the cohort is better performing than it really is, so he will under-place in the ranking given the information he gets from his peers.

Figure 3 is one way to summarize individual answers to show the impact of the treatment on knowledge. In Appendix 8 we provide detailed figures with data at the individual level, for control and treated students. Individual answers to the survey show that underperforming control students overplace themselves (they are above the 45º line), and also have more dispersion that good ones. In the figure for treated students we see that they are located in a greater proportion on the 45º line, especially good students.

Finally, Table 2 shows more evidence that the ranking system provided *new information* to treated students, relative to control ones. For each student, we calculate the difference (in absolute value) between the reported (perceived) ranking and the actual (objective) one.

Table 2 – "First Stage": Changes in perceptions.

	(1)	(2)	(3)	(4)
	Absolute Difference		Exact placement	
	Major level	School level	Major level	School level
Treated	-3.405***	-8.766***	0.297***	0.195***
	(0.568)	(1.407)	(0.031)	(0.024)
Cohort Size	0.122***	0.118***	-0.002***	0
	(0.013)	(0.01)	(0.001)	(0)
Grades distribution	-0.421***	-1.413***	0.028***	0.031***
	(0.114)	(0.301)	(0.005)	(0.005)
Constant	6.806***	22.303***	0	0
	(2.241)	(4.979)	(0.063)	(0.076)
Observations	599	608	599	608

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Absolute difference is the absolute value of the decleared-actual position in the ranking. ***Exact placement*** is a dummy variable equal to one if the students reports his exact placement in the ranking. Cohort size is the number of students in the major (col 1 & 3) or in the school (col 2 & 4). Grades distribution is decile of the student in the distribution of grades. Controls include cohort dummy variables.

Column 1 shows that a treated student places himself 3.4 positions more accurately than a student from the control group (who misplaces himself by 6.8 positions)[4]. The other coefficients show that the larger the cohort size, the larger the error a student makes when reporting his place in the ranking, and the higher GPA a student has, the more accurate he reports his placement (a smaller difference). Column 2 shows that this difference is larger at the school level: control students misplace by 22 places, while treated students reduce misplacement by 9 positions. Moreover, while only 8% of control students report their *exact* objective placement in the grade distribution at the major level, 37% of treated students report their exact position (those numbers are 2% and 22% at the school level)[5]. Appendix 9 shows another picture of the accuracy of treated students.

We call these results "First Stage" since for the impact of the experiment on the outcomes (satisfaction and academic performance) to be credible, we need that the experiment has effectively changed the information that treated students have, relative to

[4] Recall form previous sections that the average cohort size at the major level is 32 students, and at the school level is 131 students.

[5] Since the ranking is updated every day, the exact position may have changed between the moment a student saw it, the moment when he answered the survey, and when we extracted the "actual" ranking from the system. This means that the reported error for treated students may be larger than what it really is.

untreated students. Taking this into account, in the design of the experiment we included questions to be able to show that there is a First Stage in the experiment[6].

IV. Results

After explaining the intervention and showing that it was effective to changed the relative performance information that treated students have, we now proceed to show the causal impact of the experiment on satisfaction and academic outcomes. We will focus our analysis on the different responses between male and women.

Satisfaction

Satisfaction is the main focus of the online survey. We have two measures of satisfaction which were placed directly as the first and second question of the survey. The first measure, graded in a 1 to 5 points scale, is placed inside a box which also has placebo questions (i.e. satisfaction with the location of the university campus which shouldn´t change with the ranking treatment) [7]. The second question about satisfaction is in a specific module, more specific than the first measure. In this more detailed question answers are reported in a 1 to 10 points scale[8]. In this section we also included anchoring vignettes. If two students report a satisfaction of, say, eight, then we don´t really know if both answers are comparable. Maybe each student has a different internal scale where an eight means different satisfaction levels. For example, a high ability student may value less a grade of 10 than a low ability student. Moreover, a bad student mainly knows students with low grades (as we have shown with the information of the network of friends he knows only part of the distribution of grades) and may answer with a different scale than a good student who has high performing friends[9]. Anchoring vignettes may be a solution to these problems since they offer fixed and objective situations to be evaluated. So, we include vignettes to anchor the (subjective) satisfaction valuations. These vignettes show four different students trajectories: a top performing student

[6] We refer to First Stage not in a instrumental variables 2SLS sense, but as a necessary first step prior to looking for impacts in other variables.
[7] The exact wording of the first question is: "Currently, on a scale from 1 to 5, where 1 is "Very Unsatisfied" and 5 is "Very Satisfied", indicate how you feel with: (…) your cumulative GPA".
[8] We show the exact wording of the second question in Appendix 10.
[9] A similar argument is explained in Ravallion (xxx) regarding income distributions. If a poor person answers about his subjective welfare, he knows the distribution of wealth from 0 to M, and a rich person knows from M to 1, so when they answer they are not using the same scale. The rich citizen would say that he is worse than what he really is, while the poor will report a higher position that his real one (since he really doesn´t know how a rich person lives).

(*Guille*, with a GPA which placed him in the top 10% of the distribution), two students in the middle of the ranking (*Jose* and *Fer*, in percentiles 40 and 60), and a student with a GPA from the bottom 10% of the distribution (*Fran*)[10]. These hypothetical situations also had different expected time for graduation. Respondents had to evaluate these objective situations of four hypothetical students with the same 1 to 10 points satisfaction scale used to report their own subjective satisfaction[11]. The vignette for the hypothetical good student (*Guille*) is consistently evaluated as better than any of the other situations with lower GPA; while the bad student vignette (*Fran*) is evaluated worse than the other situations (Appendix 11). Moreover, we find evidence of heterogeneity in the reporting of satisfaction scales. Good and bad students report different satisfaction with the four hypothetical students. All the four hypothetical situations have a downward slope: good students give fewer points to a given GPA than low performing students. For example, satisfaction reported by students placed in the bottom quintile of the ranking with the situation of a bad student (*Fran*) is higher (6.1 points) than the evaluation of good students for the same vignette (4.4 points).

With the first measure of satisfaction (Table 3, Col 1), we find that treated men increase satisfaction by 0.13 points. The interaction term shows that the treatment effect depends on gender. Indeed, *treated*woman* has a coefficient of -0.29 (0.11) which means that treated female students decrease their satisfaction with their GPA when they were exposed to the ranking. The placebo regression (Column 2) shows no results of the treatment on the satisfaction with the university location.

[10] As it is customary when using anchoring vignettes, we used gender neutral nicknames (in Spanish unisex names are extremely rare and it was difficult to deliver gender-specific questionnaires).
[11] Other assumptions are vignette equivalence (a vignette should bring to each participant the same image or picture of the hypothetical situation that is being depicted) and response consistency (that the same process by which a student evaluates his own subjective GPA is used to evaluate the GPA in the vignettes).

Table 3 – Satisfaction with GPA

	(1)	(2)	(3)
	First measure	Placebo	Second measasure
*Treated*Woman*	-0.294***	0.041	-0.540**
	(0.106)	(0.126)	(0.222)
Treated	0.133*	-0.021	0.158
	(0.069)	(0.08)	(0.145)
Woman	0.144*	-0.205**	0.316*
	(0.078)	(0.095)	(0.167)
Constant	2.647***	4.622***	3.443***
	(0.166)	(0.191)	(0.626)
Observations	859	859	857

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.
First measure of satisfaction is constructed in a 1-5 scale. ***Placebo*** is a measure of satisfaction not affected by the treatment (satisfaction with the location of the University). ***Second measure*** of satisfaction is constructed in a 1-10 scale and includes a full set of dummy variables with information on 4 anchoring vignettes. All models include pre-treatment controls.

Column 3 in Table 3 shows the results of our most accurate measure of satisfaction (correcting for vignettes evaluations). These results also show that treated women report a significant decrease in their satisfaction with GPA as a result of being exposed to the ranking treatment. The point estimate of -0.54 is a quarter of a standard deviation in the reported satisfaction. Results in Table 3 also shows that female students have a higher satisfaction with a given GPA (even after controlling for pre-treatment GPA). The drop in satisfaction is larger than the difference in satisfaction for women relative to men.

Academic outcomes

We obtain the academic outcomes from the administrative records of the University. Therefore, we also have information for those students who didn`t answer the satisfaction survey. We report results after one and two years of treatment (Table 4 and Table 5, respectively). Columns 1 to 3 show results at the exam level, and columns 4 to 7 at the student level.

Table 4 - Academic Outcomes after one year of treatment.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Exam grade	Took the exam	Passed the exam	GPA	Approved courses	Credits gained	Dropout
Panel A: Reduced Form							
*Treated*Woman*	-0.251**	-0.031*	-0.022	-0.183*	-1.261*	-5.886	0.087***
	(0.127)	(0.017)	(0.017)	(0.101)	(0.677)	(3.988)	(0.033)
Treated	0.149*	0.01	0.012	0.063	0.453	1.746	-0.016
	(0.086)	(0.012)	(0.012)	(0.065)	(0.44)	(2.613)	(0.021)
Woman	0.274***	0.021*	0.01	0.277***	0.376	2.824	-0.013
	(0.096)	(0.012)	(0.013)	(0.078)	(0.534)	(3.074)	(0.022)
Constant	6.011***	0.776***	0.700***	5.753***	11.304***	62.325***	0.130**
	(0.186)	(0.026)	(0.025)	(0.173)	(1.058)	(6.166)	(0.058)
Observations	15,040	17,510	15,040	1,046	1,046	1,044	1,046
Panel B: First Stage							
*Treated*Woman*	0.923***	0.920***	0.923***	0.957***	0.957***	0.957***	0.957***
	(0.01)	(0.01)	(0.01)	(0.013)	(0.013)	(0.013)	(0.013)
Observations	15,040	17,510	15,040	1,046	1,046	1,044	1,046
Panel C: Second Stage							
*Treated*Woman*	-0.274**	-0.034*	-0.024	-0.191*	-1.318*	-6.154	0.091***
	(0.138)	(0.019)	(0.019)	(0.104)	(0.703)	(4.135)	(0.035)
Treated	0.164*	0.011	0.013	0.066	0.475	1.83	-0.016
	(0.094)	(0.013)	(0.013)	(0.067)	(0.454)	(2.698)	(0.022)
Woman	0.275***	0.021*	0.01	0.277***	0.38	2.834	-0.012
	(0.096)	(0.012)	(0.013)	(0.078)	(0.531)	(3.055)	(0.022)
Constant	6.008***	0.775***	0.700***	5.752***	11.297***	62.295***	0.130**
	(0.186)	(0.026)	(0.025)	(0.172)	(1.051)	(6.121)	(0.057)
Observations	15,040	17,510	15,040	1,046	1,046	1,044	1,046

*** p<0.01, ** p<0.05, * p<0.1

Controls: variables used in the randomization. Standard errors clustered at the student level.

Treatment started in July 2014. Columns (1) to (3) show exam level data between August 2014 and August 2015. Columns (4) to (7) show student level data for November 2015 (16 months after the begining of the treatment).

In the instrumental variables estimation, there are two endogenous variables ("saw_ranking" and "saw_ranking*woman") which are instrumented with "treated" and "treated*woman". For brevity, we show the results for just one endogenous variable ("saw_ranking*woman").

We find that treated women decreased their academic performance: they took less exams and got worse grades. Moreover, after one year they had a lower cumulative GPA, less approved courses and earned credits, and a higher dropout rate. The coefficient for *woman* shows that on average a female student performs better at the university than a male student. To put the magnitude of some of the results in context, the estimated (negative) impact on exam grades for treated women is almost equal to the gender difference in performance (control women score 0.274 points higher than control men,

and a treated woman decreases his score by 0.251 points). The estimated impact on GPA (Col 4) is a 66% of the gander GPA gap. The IV results (Panel C) are very similar to the corresponding reduced form estimates (Panel A), since the first stage coefficient is almost one (Panel B). Finally, Appendix 12 shows the results of regressing the treatment on pre-treatment outcomes. We can think of this exercise as a placebo test to suggest that the statistically significant effects from columns 1 to 3 in Table 4 are not due to a large sample size (even after clustering standard errors at the student level). There is no impact of the (placebo) treatment on pre-treatment individual exams grades. These results can also be seen as a pre-treatment balance check. Recall that we had balanced the cumulative GPA at the student level (n=834) and not on pre-treatment *individual exams grades* (n=31,694).

Table 5 shows the impact after two years of treatment. The negative effects on treated women are still visible. Moreover, the negative impact hasn't diluted with time. Taken together, tables 4 and 5 show that women have decreased their academic performance after treatment in several dimensions. The ranking feedback seems to have been harmful to them.

Table 5 - Academic Outcomes after two years of treatment.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Exam grade	Took the exam	Passed the exam	GPA	Approved courses	Credits gained	Dropout
Panel A: Reduced Form							
*Treated*Woman*	-0.149	-0.045**	-0.028	-0.195*	-2.546**	-13.873**	0.059
	(0.153)	(0.021)	(0.02)	(0.101)	(1.136)	(6.965)	(0.037)
Treated	0.062	0.016	0.024*	0.062	0.579	2.41	0.001
	(0.1)	(0.013)	(0.014)	(0.064)	(0.737)	(4.495)	(0.024)
Woman	0.366***	0.018	0.037**	0.286***	0.673	5.111	0.018
	(0.112)	(0.015)	(0.015)	(0.079)	(0.863)	(5.29)	(0.026)
Constant	6.560***	0.817***	0.763***	5.857***	18.511***	93.315***	0.147**
	(0.221)	(0.027)	(0.029)	(0.173)	(1.727)	(10.565)	(0.066)
Observations	8,819	9,983	8,819	1,045	1,045	1,044	1,045
Panel B: First Stage							
*Treated*Woman*	0.930***	0.932***	0.930***	0.875***	0.875***	0.875***	0.875***
	(0.016)	(0.016)	(0.016)	(0.023)	(0.023)	(0.023)	(0.023)
Observations	8,819	9,983	8,819	1,045	1,045	1,044	1,045
Panel C: Second Stage							
*Treated*Woman*	-0.17	-0.051**	-0.033	-0.238*	-3.050**	-16.445*	0.067
	(0.174)	(0.024)	(0.023)	(0.125)	(1.409)	(8.631)	(0.046)
Treated	0.077	0.02	0.029*	0.089	0.83	3.483	0
	(0.124)	(0.016)	(0.017)	(0.089)	(1.017)	(6.21)	(0.033)
Woman	0.384***	0.023	0.042**	0.319***	1.05	6.994	0.012
	(0.125)	(0.017)	(0.017)	(0.093)	(1.021)	(6.238)	(0.031)
Constant	6.548***	0.814***	0.760***	5.837***	18.299***	92.340***	0.149**
	(0.224)	(0.027)	(0.029)	(0.175)	(1.761)	(10.764)	(0.066)
Observations	8,819	9,983	8,819	1,045	1,045	1,044	1,045

*** p<0.01, ** p<0.05, * p<0.1

Controls: variables used in the randomization. Standard errors clustered at the student level.

Treatment started in July 2014. Columns (1) to (3) show exam level data between September 1st, 2015 and the August 31st, 2016. Columns (4) to (7) show student level data by October 12th 2016 (two years after the begining of the treatment).

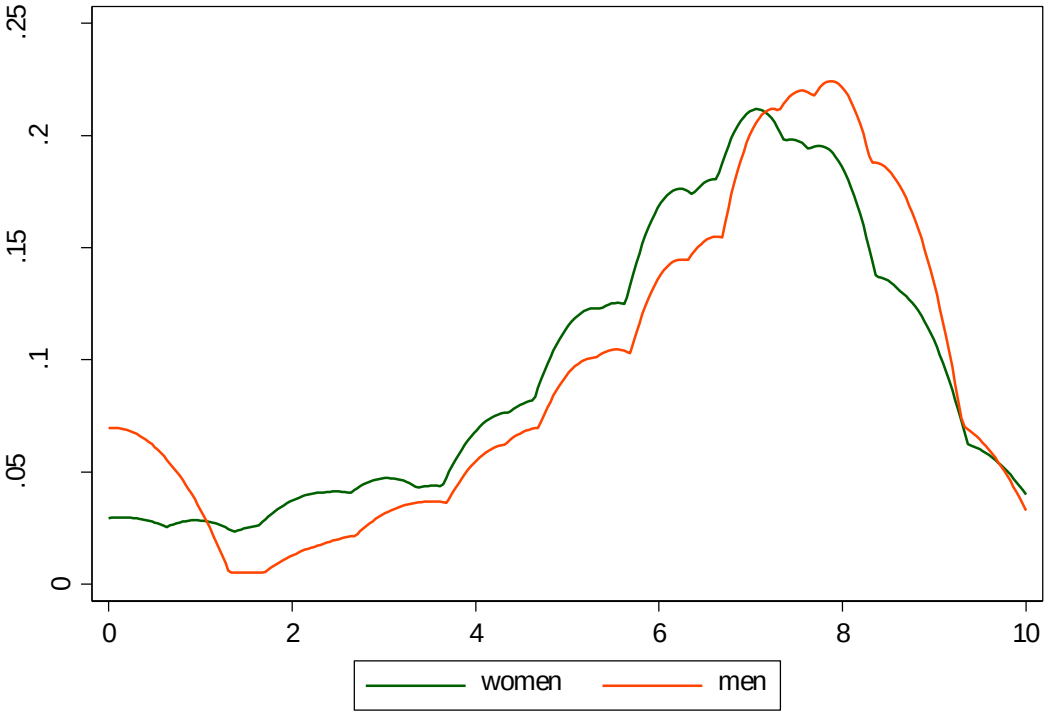
In the instrumental variables estimation, there are two endogenous variables ("saw_ranking" and "saw_ranking*woman") which are instrumented with "treated" and "treated*woman". For brevity, we show the results for just one endogenous variable ("saw_ranking*woman").

Possible channels

We will discuss to possible channels by which treatment had a negative impact of female students. We want to answer *why questions.* Not only what has happened with students in the experiment (outcomes) but also why those effects happened (channels). There is

suggestive evidence for two explanations: (i) a different willingness to compete for men and women, and (ii) an increasing marginal cost of effort.

Competitiveness. When we started the project we had the hypothesis that women were less competitive than men. Since the ranking includes a component of competition, we may expect that women will underperform relative to men after the introduction of the ranking. To test this hypothesis we have included a question on how competitive a student declares to be. Women report being 10% less competitive than men (coef = -0.107, s.e. 0.035).



	Competitiveness	
	(1)	(2)
Woman	-0.072**	-0.099***
	(0.033)	(0.035)
Constant	0.382***	0.432***
	(0.023)	(0.101)
Controls	NO	YES
Observations	843	841

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. ***Competitive*** is a dummy variable equal to one if a student reports a 7 or more in a 0-10 competitiveness scale.

*Increasing marginal cost of effort***.** To climb in the rank a student needs to put in more effort (hours of study), or if other students are improving, to keep the position in the ranking a student should study more. Since female students study more, it is more costly for them to improve the ranking.

V. Conclusions

We report the results of a field experiment on relative performance feedback. We find that students misplace themselves in the grade distribution. Our treatment increased the information they had. The ranking treatment has a competitive feature which may not have been beneficial form female students. We find asymmetric gender responses to the information on the personal position in the ranking. While treated men increased their reported satisfaction, female students in the treatment group report lower satisfaction. Moreover, treated women seem to have decreased their academic performance. They score less in their exams (especially in the short run), they take less exams and approve less courses.

The use of non-monetary incentives with a component of competition should be carefully assessed. In our case, before scaling the project for every student, based on the results from our study, the university administration decided to: (i) change the name of "ranking" to "academic trajectory" in order to reduce the competitiveness element of the new took; (ii) make the access to the system optional (an opt-in design); and (iii) exclude freshmen from this information, to let them settle in the university, and offer the relative performance feedback from year two onwards.

References

Azmat, G., Bagues, M., Cabrales, A., & Iriberri, N. (2016). What You Don't Know... Can't Hurt You? A Field Experiment on Relative Performance Feedback in Higher Education. IZA Discussion Paper No. 9853.

Benoît, J., & Dubra, J. (2011). Apparent Overconfidence. *Econometrica, 79*(5): 1591-1625.

Bertrand, M. (2011). New Perspectives on Gender. Handbook of Labor Economics, Volume 4B, Card, D., & Ashenfelter, O. Eds.

Bursztyn L., J. R. (2015) How does peer pressure affect educational Investments? *Quarterly Journal of Economics, 130*(3), 1329-1367.

Card, D., Mas, A., Moretti, E., & Saez, E. (2012). Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *American Economic Review, 102*(6), 2981-3003.

Croson, R., & Gneezy, U. (2009). "Gender Differences in Preferences." *Journal of Economic Literature*, 47(2): 448-74.

Eckel, C. C., & Grosssman, P. J. (2008). Men, Women and Risk Aversion: Experimental Evidence. In: Charles R. Plott & Vernon L. Smith (ed.), 2008. Handbook of Experimental Economics Results, Amsterdam: Elsevier.

Luttmer, E. (2005). Neighbors as Negatives: Relative Earnings and Well-Being. *The Quarterly Journal of Economics, 120*(3): 963-100.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. Psychological Review, 115(2): 502-517

Tran, A. & Zeckhauser, R. (2009). Rank as an incentive. HKS Faculty Research Working Paper Series RWP09-019.

ONLINE APPENDICES

Not intended for publication

Appendix 1 – Transcript of grades

Código	Materia	Año cursada	Sem.	Notas		Notas y fechas de exámenes				Creds.
				Final	Curso	Per. 1	Per. 2	Per. 3	Per.Extr.	
01.Obligatorias generales =CP										
00102988	Algebra	2013	I	6	6	NA 3	6			7,5
00102987	Cálculo Básico	2013	I	7	5	NA 5	8			6
00100139	Introducción a la Contabilidad	2013	I	6	7	6				9
00100227	Introducción a la Economía	2013	I	7	7	6				6
00100213	Principios de administración	2013	I	7	7	7				6
00101641	Comunicación Profesional	2013	II	7	6	8				3
00102989	Cálculo	2013	II	7	4	NP	NA 4	10		9
00100216	Derecho privado I	2013	II	6	6	6				6
00100232	Macroeconomía I	2013	II	6	4	8				6
00100050	Antropología	2014	I	8	6	10				6
00100229	Microeconomía I	2014	I	8	10	7				6
00100739	Probabilidad	2014	I	NP	5	NP	NP			7,5
00101642	Etica profesional I	2014	II	9	9	9				4,5
00100735	Matemática financiera	2014	II	10	11	8				6
00100739	Probabilidad	2014	II	8	9	7				7,5
00100322	Finanzas de la empresa I	2015	II	8	9	NP	NA 5	8		8
00103626	Pasantía Social	2015	II	12		12				1,5
02.Obligatorias específicas =CP										
00100145	Contabilidad básica	2013	II	6	5	6				9
00100148	Contabilidad intermedia	2014	I	8	9	7				9
00100217	Derecho privado II	2014	I	7	7	8				6
00100152	Contabilidad de costos	2014	II	10	9	NA 5	11			9
00100215	Derecho laboral	2014	II	6	5	6				6
00100154	Contabilidad avanzada	2015	I	7	8	6				12
00100151	Contabilidad de gestión	2015	I	11	10	11				6
00100203	Costos para la toma de decisiones	2015	I	8	8	8				6
00100220	Legislación tributaria	2015	I	10	9	10				6
00100204	Control interno	2015	II	6	7	6				3
00100159	Teoría contable	2015	II	9	8	10				9
00100223	Técnica tributaria I	2015	II	8	5	NA 4	10			8
00100133	Auditoría I	2016	I	8	8	8				9
00100318	Banca y bolsa	2016	I	8	8	9				6
00100324	Finanzas de la empresa II	2016	I	8	8	8				8
00100225	Técnica tributaria II	2016	I	7	7	6				8
03.Electivas ciencias sociales										
00100751	Historia contemporánea	2013	II	10	10	10				4,5
08.Electivas generales =CP										
00100466	Marketing I	2013	II	9	7	10				6
00101728	Informática Intermedia	2014	I	11	11	11				4,5
00102128	Preparation FCE I	2015	I	9	9	9				6
00102129	Preparation FCE II	2015	II	9	9	9				6
00101732	Contabilidad Sector Agrop. e Industrial	2016	I	10	10		10			4,5
00104627	Procesos de negocio con SAP	2016	I	12	12	12				4,5
19.Prácticas profesionales = FCE										
00100791	Práctica Profesional I	2014	I	8		8				10
Total de créditos obtenidos: 264										
Promedio de Aprobaciones: 8,2 **Promedio General: 7,7**										

Appendix 2 – Ranking treatment - examples

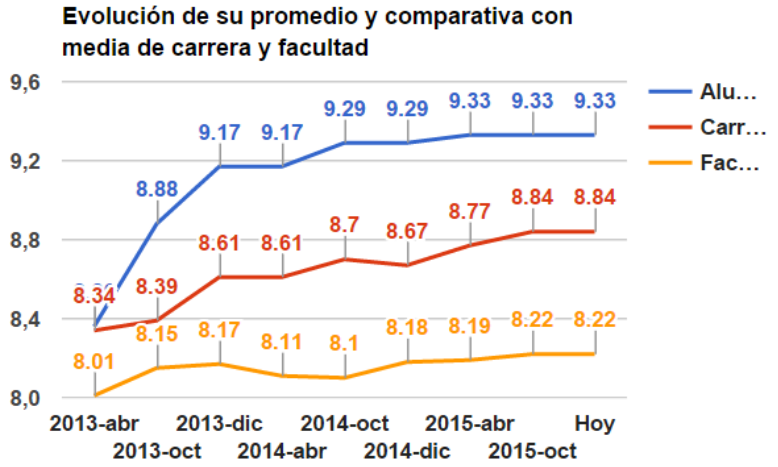
Example 1

¿Qué es el ranking?

Licenciatura en Economía

Ranking actual

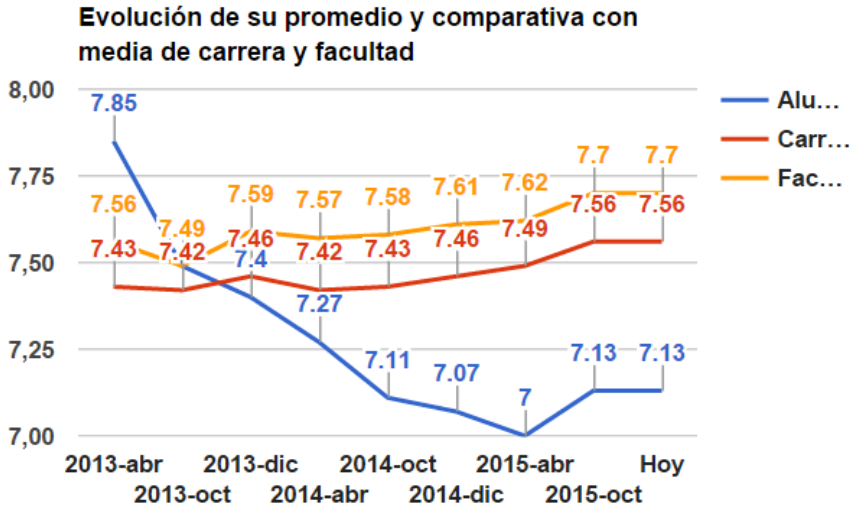
Tipo	Posición	#Alumnos
Carrera	10	20
Facultad	35	144



Example 2

Ranking actual

Tipo	Posición	#Alumnos
Carrera	49	74
Facultad	93	131



Appendix 3 – Email sent to treatment and control group

Estimados alumnos:

Te recomendamos planificar bien tus exámenes y el semestre que viene. Para eso ten en cuenta:

- las fechas de exámenes (disponibles en xxx),
- el período de modificaciones (xxx)
- tu escolaridad y la grilla de avance académico (xxx)
- *También tienes disponible a partir de hoy un ranking de tu desempeño en la facultad.*

Saludos,

XX

Notes:

1. The email to treated students included the text in *italics*, letting them know that they had available from that day the ranking of their academic performance at the Faculty.
2. The xxx replace the links to specific intranet web pages.

Appendix 4 – Characteristics of the students who answered the satisfaction survey.

Variable	Completed survey	Didn´t complete	Difference
Treated	0.51	0.46	0.05
	(0.02)	(0.04)	0.04
Female	0.45	0.43	0.03
	(0.02)	(0.04)	0.04
Courses	26.35	24.45	1.90
	(0.60)	(1.47)	1.58
Cummulative GPA	7.58	7.23	0.35
	(0.05)	(0.13)	0.14 **
Approved courses	21.85	20.11	1.74
	(0.53)	(1.32)	1.42
Credits earned	142.93	126.91	16.02
	(3.41)	(8.12)	8.79 *
Number of degrees	1.07	1.06	0.01
	(0.01)	(0.02)	0.02
School of origin	0.29	0.27	0.03
	(0.02)	(0.03)	0.04
Montevideo	0.65	0.72	-0.07
	(0.02)	(0.03)	0.04 *
Large Scholarship	0.24	0.22	0.02
	(0.01)	(0.03)	0.03
Cohort	2011.81	2011.86	-0.05
	(0.06)	(0.14)	0.15
Decile grades_career	4.95	4.20	0.75
	(0.10)	(0.21)	0.23 ***
Decile grades_school	5.10	4.42	0.68
	(0.10)	(0.22)	0.24 ***
Observations	861	187	

Difference in means in baseline characteristics between students that completed the survey and those who didn´t.

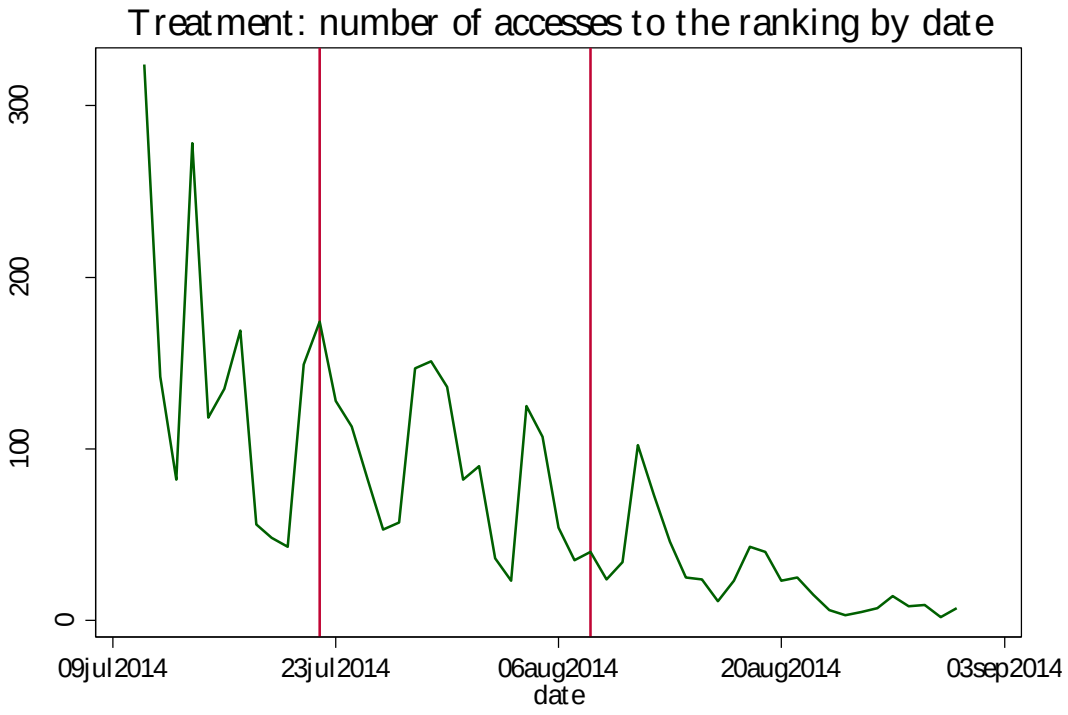
*** p< 0.01, ** p< 0.05, * p< 0.1. Standard errors in parenthesis.

Appendix 5 – Balance in survey response (after attrition).

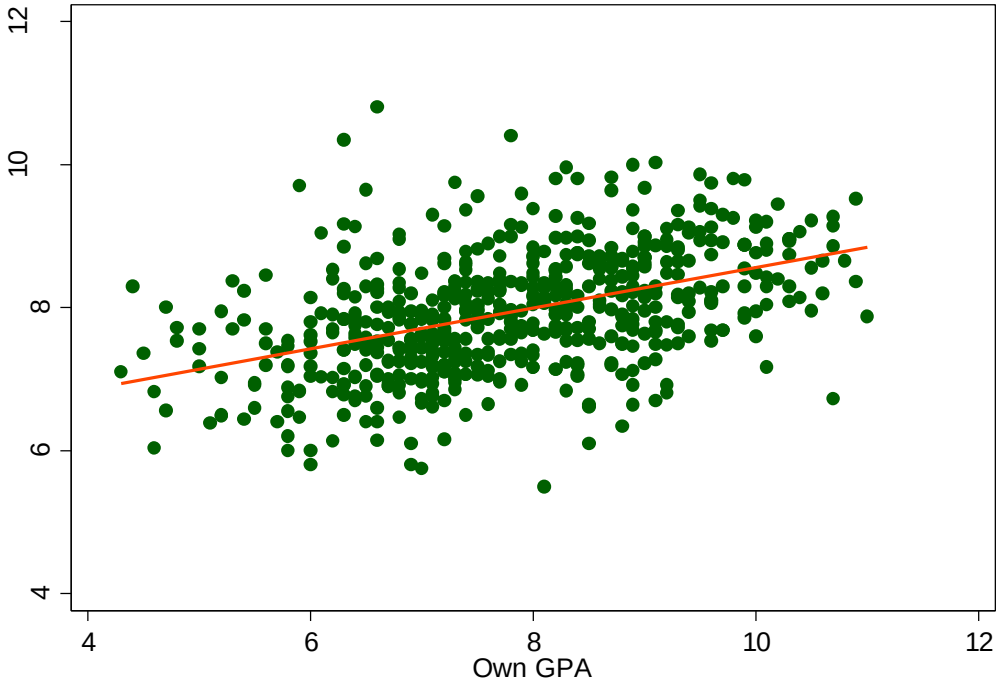
	(1)	(2)	(3)	(4)	(5)	(6)
	Treated	Control	difference (1-2)	Standard Error	p-value	Obs.
Student characteristics						
1 *Female*	0.46	0.44	0.02	-0.03	0.65	861
2 *Courses*	26.44	26.30	0.14	-1.21	0.91	861
3 *Cumulative GPA*	7.55	7.61	-0.06	-0.10	0.57	698
4 *Approved courses*	21.77	21.98	-0.21	-1.06	0.84	861
5 *Credits earned*	142.72	143.48	-0.76	-6.83	0.91	861
6 *Number of degrees*	1.09	1.06	0.03	-0.02	0.10	861
7 *School of origin*	0.31	0.28	0.03	-0.03	0.34	861
8 *Montevideo*	0.66	0.64	0.02	-0.03	0.53	861
9 *Large Scholarship*	0.25	0.24	0.01	-0.03	0.82	861
10 *Cohort 2008*	0.04	0.04	0.00	-0.01	0.87	861
11 *Cohort 2009*	0.06	0.05	0.01	-0.01	0.66	861
12 *Cohort 2010*	0.13	0.13	0.00	-0.02	0.98	861
13 *Cohort 2011*	0.21	0.22	-0.01	-0.03	0.66	861
14 *Cohort 2012*	0.21	0.21	0.00	-0.03	0.99	861
15 *Cohort 2013*	0.19	0.16	0.03	-0.03	0.26	861
16 *Cohort 2014*	0.18	0.20	-0.02	-0.03	0.45	861

The difference in means is calculated with an OLS regression with robust standard errors. *** p<0.01, ** p<0.05, * p<0.1. From the 861 students, there are 443 (51.5%) in the treated group and 418 (48.5%) in the control group. Balance was also performed by schools, degrees and place in the distribution of grades at the major and school level. These results are omited from this table to ease the display of the main results.

Appendix 6 – Access to the Ranking System before and after the time of the satisfaction survey.

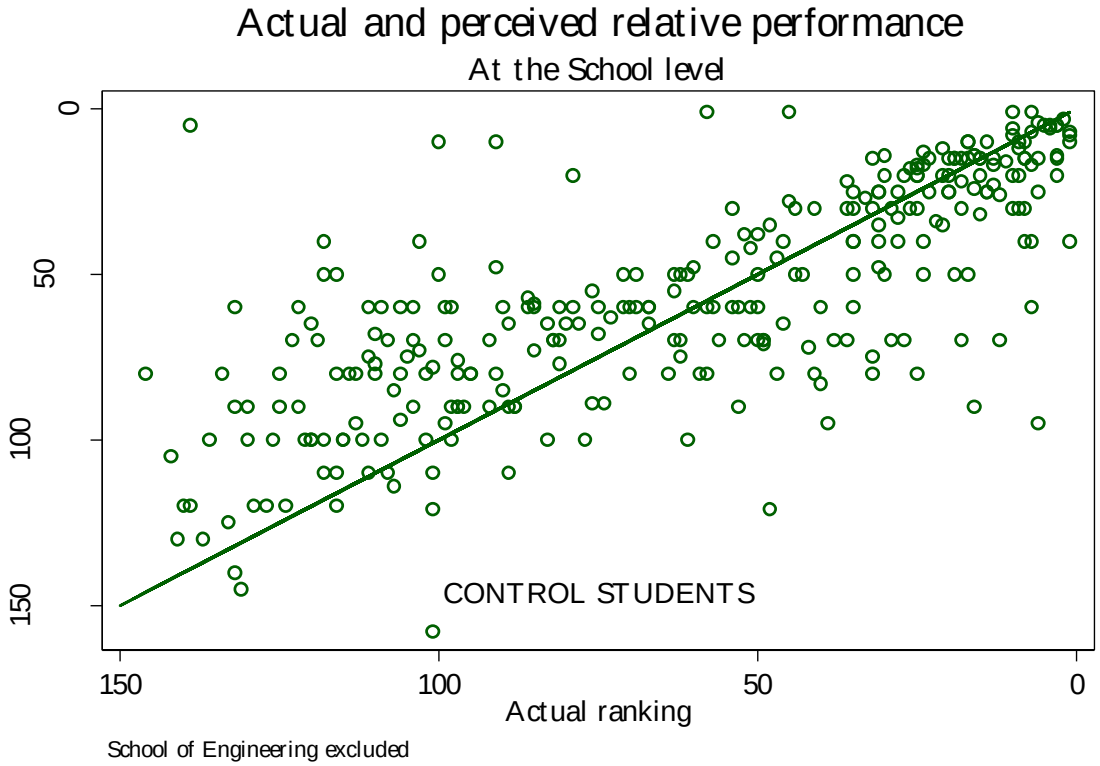


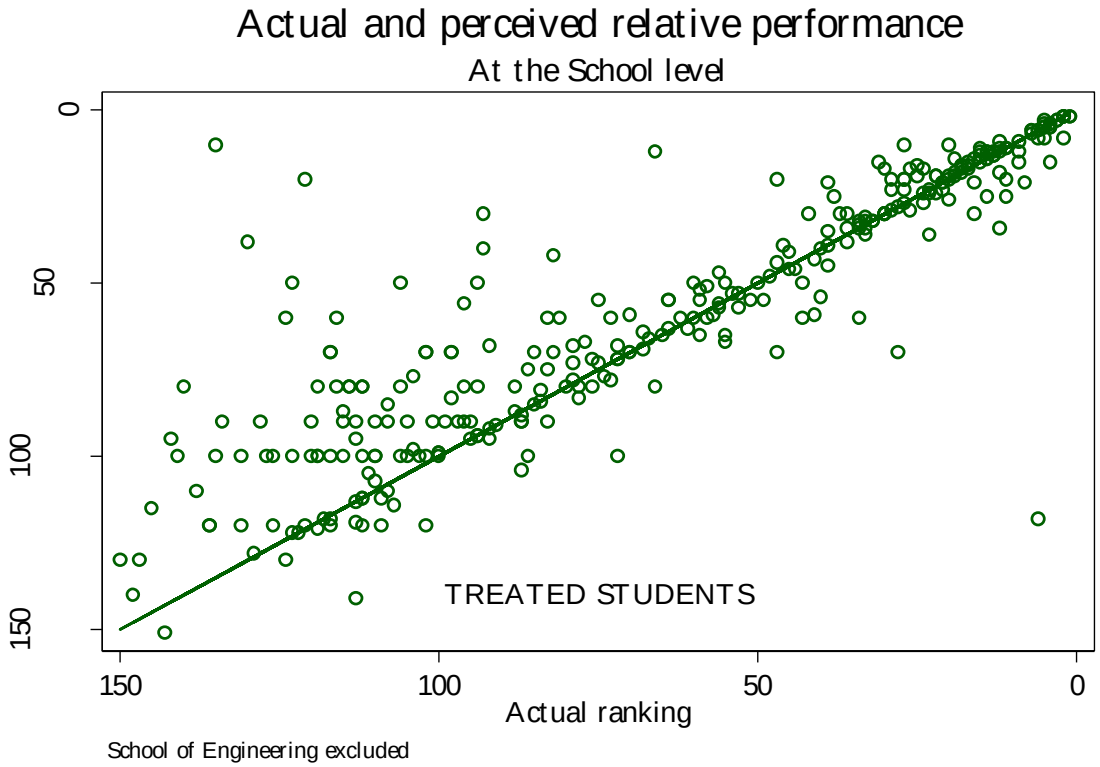
Appendix 7 – Selection of friends. Own GPA vs Friends GPA



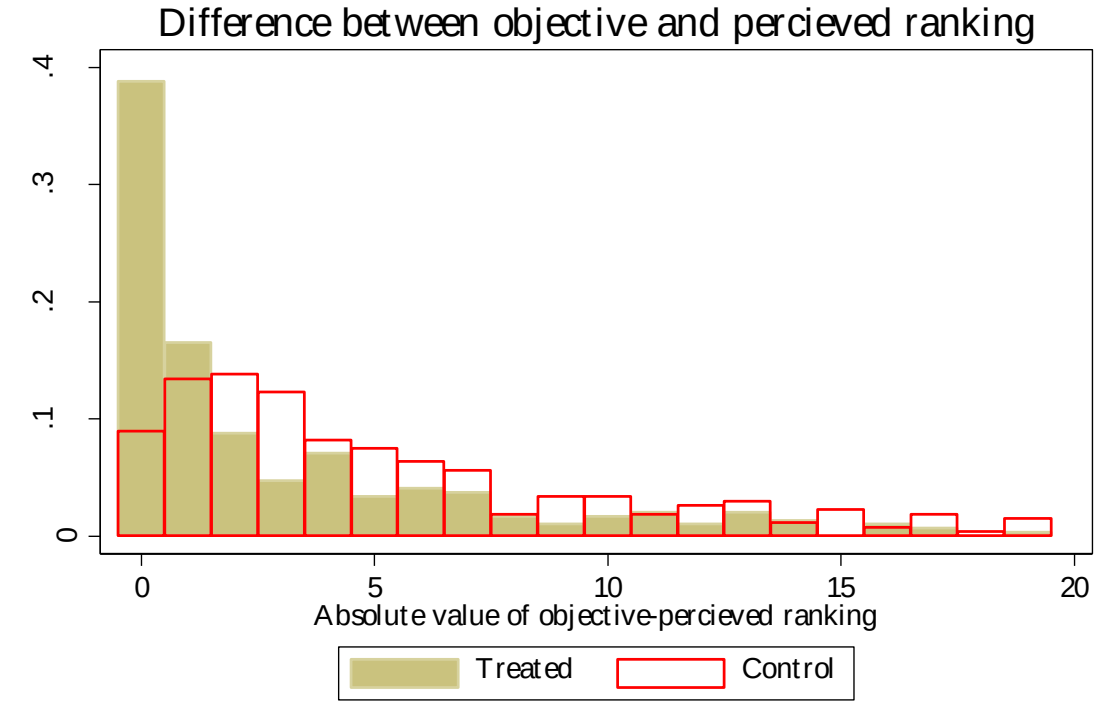
One more point in friends GPA is associated with an increase of 0.77 points in Own GPA

Appendix 8 – Individual answers to perceived ranking question.

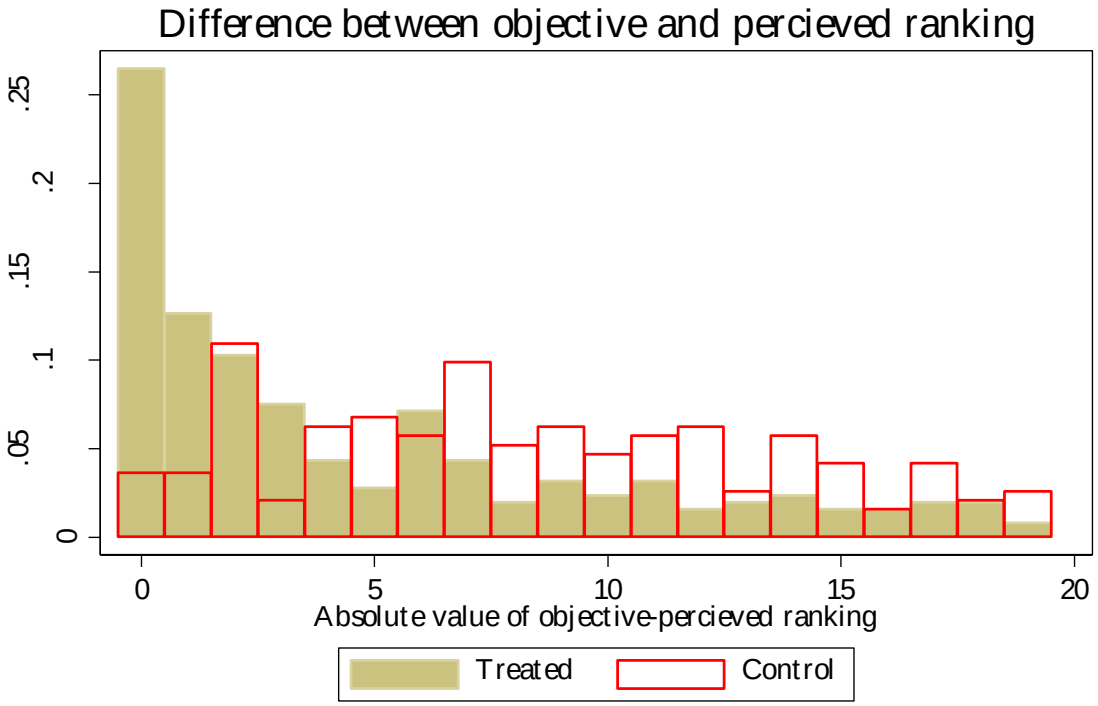




Appendix 9 – Accuracy of students report of their placement in the grade distribution



Zoomed at absolute difference < 20
Major level

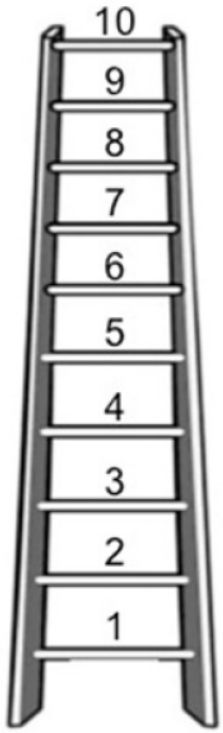


Zoomed at absolute difference < 20
School level

Appendix 10 – Second question about satisfaction.

ENCUESTA DE ASPIRACIONES Y SATISFACCIÓN CON LA CARRERA

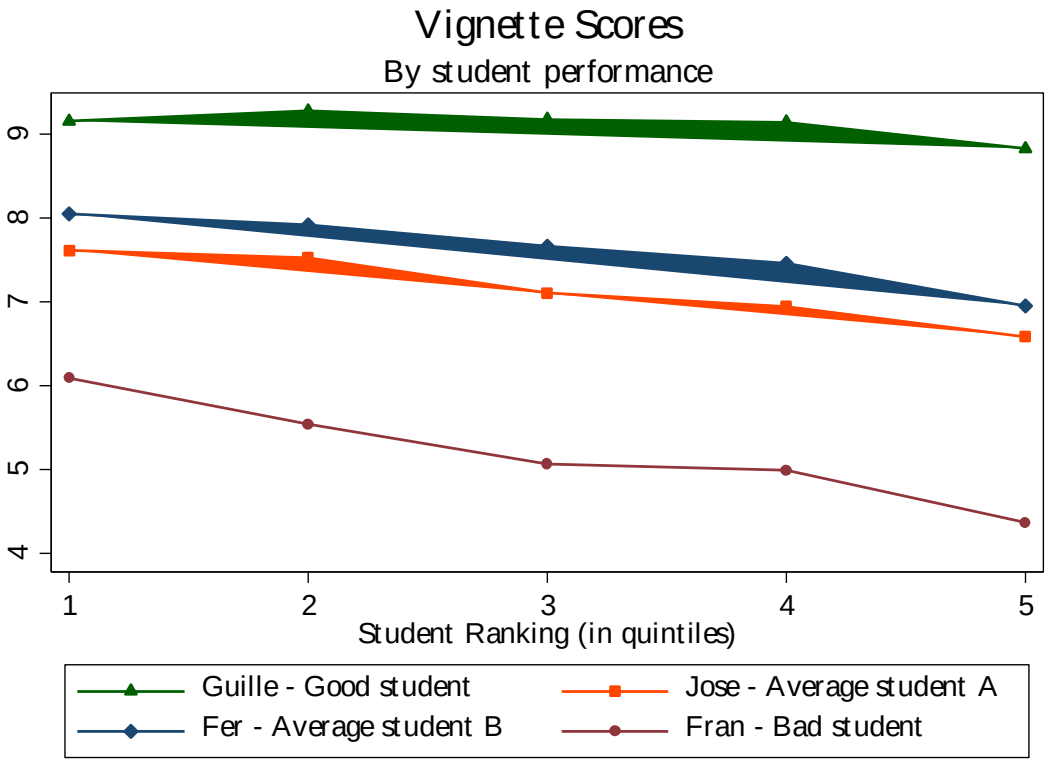
Imagina una escalera con escalones numerados desde el 1 abajo hasta el 10 arriba. La parte superior de la escalera representa la mejor vida académica para tí y la parte inferior la peor vida académica para tí.



Completamente satisfecho

Totalmente insatisfecho

Appendix 11 - Punctuation of the vignettes



Appendix 12 – **Pre-treatment (placebo) impact**

	(1)	(2)	(3)
	Exam grade	Took the exam	Passed the exam
*Treated*Woman*	-0.041	-0.01	0.005
	(0.063)	(0.014)	(0.009)
Treated	-0.015	-0.002	-0.004
	(0.04)	(0.009)	(0.007)
Woman	0.053	0.003	0.001
	(0.047)	(0.01)	(0.007)
Constant	6.250***	0.792***	0.708***
	(0.202)	(0.032)	(0.022)
Observations	31,694	37,462	31,694

*** p<0.01, ** p<0.05, * p<0.1

Controls: variables used in the randomization. Standard errors clustered at the student level.

Columns (1) to (3) show exam level data between July 2003 and August 2014. Treatment started in July 2014.