THE ECONOMIC JOURNAL

The Economic Journal, **131** (*November*), 3145–3181 https://doi.org/10.1093/ej/ueab043
Advance Access Publication Date: 6 May 2021

RELATIVE PERFORMANCE FEEDBACK IN EDUCATION: EVIDENCE FROM A RANDOMISED CONTROLLED TRIAL*

L.I. Dobrescu, M. Faravelli, R. Megalokonomou and A. Motta

In a one-year randomised controlled trial involving thousands of university students, we provide real-time private feedback on relative performance in a semester-long online assignment. Within this set-up, our experimental design cleanly identifies the behavioural response to rank incentives (i.e., incentives stemming from an inherent preference for high rank). We find that rank incentives boost performance in the related course assignment, but also the average course exams grade by 0.21 SDs. These beneficial effects remain sizeable across all quantiles and extend beyond the intervention period. Furthermore, rank feedback stimulates social learning, i.e., rank incentives make students engage more in peer interactions.

We live in a world obsessed with social comparisons. From sport competitions and school ratings to the number of likes on a Facebook post or views of a YouTube video, we constantly witness society's fixation with relative performance. Social comparisons are also encouraged within organisations: allowing individuals to compare their performances has the potential to increase their productivity in educational (Azmat and Iriberri, 2010; Tran and Zeckhauser, 2012; Katreniakova, 2014; Brade *et al.*, 2020), labour (Mas and Moretti, 2009; Blanes i Vidal and Nossol, 2011) and public goods provision settings (Chen *et al.*, 2010).

Social comparisons, however, are not always a silver bullet. If individuals discover that others' ability is lower than initially thought, they might decide to put less effort into the task at hand (Azmat *et al.*, 2019). The same drop in effort can occur due to demoralisation effects brought by a lower-than-expected rank (Barankay, 2011).[1] Overall, comparing the existing studies is particularly challenging because of the numerous confounding factors: the behavioural response to incentives stemming from an inherent preference for high rank (henceforth *rank incentives*)[2] is potentially compounded with financial and signalling aspects, learning and experimentation processes, multi-tasking considerations, peer pressure and changes in beliefs about future compensation schemes and relative ability. This underlying complexity could explain the mixed

* Corresponding author: L.I. Dobrescu, School of Economics, UNSW Business School, University of New South Wales, Kensington 2052, Australia. Email: dobrescu@unsw.edu.au

This paper was received on 19 June 2019 and accepted on 1 May 2021. The Editor was Gilat Levy.

The authors were granted an exemption to publish their data because access to the data is restricted. However, the authors provided the Journal with temporary access to the data, which allowed the Journal to run their codes. The codes are available on the Journal website. The data and codes were checked for their ability to reproduce the results presented in the paper.

A previous version of this paper was circulated under the title 'Rank Incentives and Social Learning: Evidence from a Randomized Controlled Trial'. We are grateful to Raphael Brade, Zachary Breig, David Byrne, Dimitri Christelis, Luke Chu, Francesco Fallucchi, Miguel Fonseca, Lata Gangadharan, Benjamin Hansen, Oliver Himmler, Robert Jackle, Boon Han Ko, Umair Khalil, Andreas Leibbrandt, Daniele Nosenzo, Maria Recalde, Tom Wilkening, Haishan Yuan, the editor and four anonymous referees, as well as seminar and conference participants at several institutions, for helpful discussions and suggestions. This work was approved by the Human Ethics Research Committee at University of New South Wales (HREC201700142) and the University of Queensland (HREC2017001402).

[1] There are also a considerable number of laboratory experiments on the effect of feedback on relative position—see, for instance, Hannan *et al.* (2008), Eriksson *et al.* (2009), Kuhnen and Tymula (2012), Charness *et al.* (2013), Gerhards and Siemer (2014), Azmat and Iriberri (2016), Gill *et al.* (2019), among others.

[2] See Barankay (2012) and Tran and Zeckhauser (2012).

results in the literature, with rank feedback being productivity-enhancing in certain contexts but not others (Bandiera *et al.*, 2013; Bursztyn and Jensen, 2015; Blader *et al.*, 2016).[3]

In this paper, we report results from a one-year randomised controlled trial (RCT) in the education sector, involving university students who received real-time, private feedback on their relative performance in a semester-long computerised (online) assignment. Our contribution is multi-fold. First, our experimental design allows to cleanly identify the behavioural response to rank incentives per se. Second, we are the first to provide evidence that rank incentives can be effective in higher education.[4] Indeed, we find that feedback has a positive impact on students' performances not just in the online assignment on which they were ranked, but also in all invigilated exams taken in the (intervention) course and across the entire grades' distribution; this impact is long-lasting, with positive spillovers to other courses beyond the intervention period. Third, we are the first to explore the virtues of real-time feedback; as we will discuss below, this feature may be at the basis of the success of our implementation. In fact, the success of our RCT does suggest that results may be very sensitive to apparently innocuous design features.[5] Our fourth contribution is to examine some of these feedback characteristics by analysing a number of alternative variations in the way it is provided. Last, but not least, we uncover the mechanism through which rank feedback translates into academic performance. Our findings suggest that relative performance feedback makes students engage more in *social learning*.[6,7]

Tackling our research questions is challenging because it requires detailed information on how individuals *prepare* for a task. To achieve this, we first develop a semester-long online assignment with a leaderboard system. The assignment is a collection of practice tests (i.e., sets of exercises) tackling a series of simple problems. The leaderboard score is the product of the completion rate (i.e., the number of completed exercises over the total number of exercises available) and the success rate (i.e., the number of successful attempts to solve exercises over the total number of attempts made).[8] This structure ensures that students remain clustered around a similar score as they progress in the assignment (completion rate), while also allowing for enough variation (success rate). All students are free to engage with the assignment at any time during the semester

[3] From a welfare perspective, the provision of relative performance feedback does not seem to increase stress levels (Katreniakova, 2014), but it can affect satisfaction (happiness and dominance levels) in either direction (Azmat and Iriberri, 2016).

[4] The interventions in Tran and Zeckhauser (2012) and Brade *et al.* (2020) also involve university students. However, (*i*) Tran and Zeckhauser (2012) study the impact of rank incentives on the score attained in a standardised, externally administered TOEIC test, while (*ii*) Brade *et al.* (2020) provide feedback on accumulated course credits; thus, one cannot disentangle rank incentives from the other incentives associated with the tangible benefit of signalling ability or readiness for the job market.

[5] The potential for real-time feedback to address the bias toward what is salient and immediately visible has also been documented by Jessoe and Rapson (2014) and Tiefenbeck *et al.* (2016) for energy consumption. The design of real-time feedback also relates to the emerging area of dynamic information management (Ely, 2017).

[6] From a theoretical perspective, social learning has been proven to emerge when individuals are allowed to observe each other's outcomes. Observing these outcomes, for example, can help one select a more effective technology (Wolitzky, 2018). In our context, students may question whether they are approaching the task (i.e., studying) in the most effective manner: ranking not only conveys information about one's relative ability and effort levels, but it also reveals how efficient one's productive technology is.

[7] Engaging in social learning is only one of the potential strategies that treated students may have utilised. Other strategies, such as studying more by themselves, would not be captured in our dataset and might well be the channels that generate the increase in performance. It is also possible that engaging in social learning (and the associated increase in message board contributions) is both instrumental and a signalling behaviour, i.e., the treatment affects the contributions via a boost in self-confidence. We thank two anonymous referees for suggesting these points.

[8] Students are provided with random values for each exercise and have to take decisions that lead them to achieve certain economic goals. Once a decision is taken, it is automatically stored on the server. Students receive feedback after every decision: if it is correct, they are allowed to proceed to the next one; if it is suboptimal, they are asked to review their choice.

and receive the same type of information about their absolute assignment performance. On top of this, those in the treatment group are also shown their score *rank* for one minute any time this varies, either upward or downward.

A consequence of this design is that students can act right after inspecting their relative performance and, therefore, they can affect their assignment rank almost instantly. Hence, any initial boost of effort due to students experimenting with the effort level required to change their rank—and its transitory impact on performance—is unlikely to affect the overall performance by the end of the semester. We also note that the leaderboard score had no bearings on the overall course grade, as 20% of it depended exclusively on a student's completion rate at the end of the semester and not on how well they performed (either in terms of success rate or in terms of leaderboard score). This precludes any possible financial or signalling considerations. Our set-up is also characterised by the private nature of feedback. This feature allows us to rule out confounding factors that may otherwise concur to generate the results—such as status seeking behaviour or attitudes driven by parents' pressure.[9] Finally, to verify whether treated students perform better at the expense of other academic activities (i.e., address the multi-tasking argument), we collect information on their performance in all the other courses taken in the same semester. All this considered, our treatment effects are likely to cleanly capture *rank incentives* as in Barankay (2012).[10]

To identify the impact of this type of ranking, we exploit the tests required to assess the notification system when the online assignment was first deployed to students. The RCT took place in the first semester of 2016 and involved the students enrolled in a large Principles of Microeconomics course at a major, research intensive, selective Australian university. As part of these tests, students were randomly divided into five groups: one control (receiving no information on their relative performance) and four treatments (featuring the leaderboard system described above and three slight variations of it). Students were not aware of these tests and thus did not know they were being treated. Instructors were not involved in these tests either and did not know to which groups their students were assigned.

We find that treated students perform better in the online assignment, ending up, on average, 62 (out of 1,093) positions higher in the final ranking than students in the control group. Such improvement is robust and of similar magnitude across all quantiles. What really matters, though, is that this direct effect on the assignment also translates into better performance in all course exams taken over the semester. Our results show that providing feedback increases the average grade across all invigilated exams over the semester (i.e., two mid-terms and a final exam) by 0.21 SDs. These effects remain positive, robust and sizeable across all quantiles, with no further heterogeneity by gender, age, international status or field of study.[11]

Our findings so far prompt the question of whether feedback provision in a course assessment can have a spillover impact on the other courses taken in the same semester. And can we identify any long-lasting effects of our intervention on academic performance in the next semester? To

[9] For an overview on providing private versus public ranking feedback, see also Tran and Zeckhauser (2012) and Gerhards and Siemer (2014; 2016).

[10] One additional concern is that treated students might feel more monitored or looked after as a consequence of receiving relative performance feedback. However, during the online assessment, both control and treatment students receive detailed feedback and summary statics about every single action they take. Strictly from this perspective, all students would be likely to feel monitored or that they receive sustained attention. While all this feedback pertains to their absolute performance, it is common knowledge that it could be used to construct relative performance indicators as well in the background. We thank an anonymous referee for raising this point.

[11] Exceptions involve females and international students with beneficial effects in the final exam, and 19+ students whose Week 10 and final exams were positively impacted.

answer these questions, we use the adjusted gross point average (GPA) in Semester 1, 2016,[12] as well as GPA in Semester 2, 2016, and re-run our analysis. We find that our intervention does not change performance in the other courses taken in the same semester, with the corresponding GPA remaining unaffected by feedback provision across all treatments. Estimates also reveal, however, an increase in academic performance for the next semester by 0.17 SDs. Because our intervention had no effect on the number of courses taken (or passed), we conclude that not only did treated students perform better at the expense of other academic activities, but also that the beneficial effects of the intervention are long-lived.

These results are noticeable for four reasons. First, providing feedback on a drill that has no bearings on students' grades not only has a positive effect on students' performance in that drill, but also on their grades throughout the semester. Second, this effect translates into a performance boost which is equivalent to being taught by a teacher 2 SDs above the average or, alternatively, in a class 20% smaller—both extremely costly interventions to implement and support.[13] Third, our findings indicate that relative performance feedback can benefit students across the entire grade distribution. Fourth, the persistence of the effect several months after the intervention is remarkable and points towards ranking being a successful and effectively costless tool to implement in order to boost performance, not only in the short-term but also over time.

Finally, we turn to our findings pertaining to the underlying mechanism. Following Azmat and Iriberri (2010), we present two theoretical models, one based on competitive preferences and the other on self-perception. Despite being quite simple, the competitive preferences model manages to underpin most of our empirical results. The intuition is that a more precise signal regarding the average performance in the course makes the competitive component of the utility function more salient, triggering a bigger behavioural response that induces an increase in effort.

Can we actually observe this increase in effort? Fortunately, we possess two such measures. First, we observe the time that they spend on the online assignment. In our case, the time spent appears to be the same across all treatment and control groups. However, effort takes different forms and shapes. Our second source of data is related to students' activity as recorded on two separate discussion boards, one internal and one external to the official course website. Both boards are used by students to ask questions related to the course content and materials. Among the students who were engaged in the discussion boards, we find that those exposed to their rank post about 53% more often than their control counterparts. However, we find no effect at the extensive margin, suggesting that our intervention made active students even more active, rather than encouraging more students to contribute to these public forums. Taking the number of posts as an indication of course engagement, we interpret this as evidence that students in the treatment group interacted more with their peers. This suggests that *social learning* is one of the channels at play here.

A secondary contribution of this paper is to provide the first evidence of the effectiveness of real-time, relative performance feedback and to show that results on younger, pre-college aged students do carry on to higher education.[14]

Using this new type of continuous feedback raises a number of interesting issues. What is the optimal design? How frequent should the feedback be? What type of feedback works best? In this paper, we attempt to offer a few preliminary answers to these questions by breaking down the main features of the leaderboard system to study what elements are most effective in

[12] Adjusted GPA is derived from all the courses taken in Semester 1, 2016, except for the intervention course.
[13] See Angrist and Lavy (1999), Krueger (1999), Hanushek *et al.* (2005) and Chetty *et al.* (2014).
[14] See Azmat and Iriberri (2010), Tran and Zeckhauser (2012) and Katreniakova (2014).

eliciting effort. Specifically, we exploit the RCT to deconstruct the leaderboard based on two basic components: feedback type and duration. Students in the second treated group were *constantly provided with their rank* instead of being exposed to it for just one minute when it varied. By comparing this treatment ('Non-stop') with our principal treatment ('Main'), we can study the impact of feedback duration on students' behaviour. In the third treatment ('Positive') students were shown their rank for one minute *if and only if this varied upward*. In the fourth treatment ('Negative'), they were shown their rank for one minute *if and only if this varied downward*.[15] We use these last two treatments to examine how performance is affected by the type of news that the feedback conveys.

Taken separately, all treatments have positive point estimates relative to control, with magnitudes decreasing in the following order in the case of average exam grade: Main, Non-stop, Negative and Positive. A similar trend exists for course engagement, namely: Main, Non-stop, Negative and Positive. That said, only students in our Main treatment significantly outperform control students in terms of both course exam grades and social engagement. These results are consistent with our theoretical model's predictions. Based on insights from neuroscience, the model proposes that our rank indicator is a bottom-up type of stimulus—i.e., the rank signal will be at its most accurate in the Main treatment (a pop-up stimulus) compared to Non-stop (static stimulus), Positive and Negative (less frequent and informative by construction). This is in line with the notion that feedback provision is very sensitive not only to the *type* of information disclosed, but also to the *way* in which it is disclosed (Wedel and Pieters, 2012). In fact, pooling all four feedback treatments and comparing them with the control group yields significant effects only for the second mid-term exam, although estimates are positive for all assessments and, hence, suggestive of a beneficial rank feedback impact across the board.

The paper proceeds as follows. The next section describes the RCT and its context, while Section 2 derives the theoretical predictions. We then move to present the data and our empirical approach, followed by a discussion of our findings and of the mechanism behind them. Finally, we conclude and discuss future work.

1. Randomised Controlled Trial

1.1. *Environment*

The RCT took place in Semester 1, 2016, among the students of a large Principles of Microeconomics course at a research intensive, selective Australian university.[16] The course is taught every semester to more than 1,000 students over 13 teaching weeks. Each week, students attend a two-hour live lecture and a one-hour tutorial; neither classes are compulsory and attendance is not recorded. Lectures are delivered by academic staff, while tutorials are taught by teaching assistants (also known as 'tutors'). All lectures and tutorials take place in the same campus. In our case, all lectures throughout the semester are taught by the same lecturer, who is also the sole course coordinator. As for tutorials, students are randomly assigned to these classes at the beginning of the semester and cannot switch between them during the term. Each tutorial includes, on average, 24 students and effectively consists of solving exercises and discussing course materials, both activities guided by a tutor. All instructors (i.e., lecturer and tutors) use the

[15] To the best of our knowledge, only a few papers so far have studied how productivity is affected by changing the likelihood of rank feedback, the reference group used and the informativeness of feedback (Kuhnen and Tymula, 2012).

[16] Similar to the United States, the Australian academic year includes an autumn and a spring semester. The *autumn* semester—and the Australian academic year—however, starts in March, while the *spring* semester starts in July.

same teaching material, including textbook, course notes and slides, and tutorial exercises (with standardised solutions provided by the course coordinator). Finally, there are two discussion boards associated with the course, one internal and one external to the official course website. Both discussion boards are accessible by all students and all instructors and are used to post comments and ask (or answer) any course-related questions.

From 2016 onwards, educational software was adopted as part of the course material. This software provides students with access to an extensive database of exercises and links to the (electronic) course textbook. The textbook covers all the topics traditionally taught in a standard Principles of Microeconomics course, from the principle of comparative advantage through to externalities and public goods. Exercises, on the other hand, are grouped into several sets, each focusing on a different economic topic and each corresponding to a different textbook chapter. These sets are released progressively throughout the semester, keeping track of the issues discussed in class. Students are required to master them in a certain order, but upon completing them, they can go through them at will, in any order and at any time. Correctly solving all available exercise sets (i.e., fully completing the online assignment) by the end of Week 13 is worth 20% of the overall course grade. In case of partial completion, students receive a proportion (approximated to the first decimal) of the 20% that is equivalent to their completion percentage.

Besides the semester-long online assignment, the course assessment structure also included (*i*) two invigilated mid-term exams taking place in Week 6 and Week 10 of the semester, containing several essay questions, each worth 20% of the overall course grade; and (*ii*) one invigilated final exam, taking place at the end of the semester, containing only multiple-choice questions and counting as 40% of the overall course grade. The exam papers for all three invigilated exams are created by the course coordinator, who draws the corresponding questions from a pre-existing database of uniformly difficult questions. Each tutor marks an equal proportion of mid-term exam papers, not necessarily from their own tutorial students; marking is double-blind and follows a strict set of marking guidelines provided by the course coordinator with rigorous consistency checks in place. A machine automatically grades the multiple-choice questions of the final exam.

1.2. *Treatments*

As anticipated in the introduction, there are two basic measures of students' performance in the online assignment. The *completion rate*, at a given point in time, represents the proportion of exercises completed up to that moment. The *success rate*, at a given point in time, represents the percentage of correct decisions taken up to that moment.[17]

Let P denote the performance index, as generated by the product of a student's completion and success rate. As a feature of the online assignment, the software includes a leaderboard ranking all students in the course based on their P index. Every *five* minutes the server updates the leaderboard ranking and, if an individual's position has varied, notifies the student by displaying on their screen (*i*) their relative position with respect to all other participants, and (*ii*) their latest variation in ranking. This information disappears after *one* minute and appears again only once a subsequent server update of the leaderboard picks up another variation. Figure 1 shows an

[17] Thus, if two students have the same completion rate, the one who has made fewer mistakes has a higher success rate. However, a student who has taken only one decision, provided it is correct, has a higher success rate than one who is ahead in the assignment (and possibly already completed it) but has made mistakes.

Fig. 1. *Example of Ranking Feedback as Shown in the Assignment.*

example of a student who dropped 11 positions and is currently ranked 770th out of 1,101. Moreover, by clicking the button '*i*' on the icon, an information box would appear and explain how the individual rank is constructed.

To test the notification system, the software developer conducted a (platform-wide) RCT by varying the manner in which rank feedback was conveyed online to students. These tests were conducted (*i*) to ensure a student's rank was correctly calculated, displayed and updated, and (*ii*) to identify the rank disclosure format that would elicit the most user activity. To this effect, while the Main treatment was the default disclosure format, the developer also deployed three additional small variation disclosures compared to the Main format. Students were randomly divided into five groups (one control and four treatments) at the start of Semester 1, 2016, based on the last digit of their student ID. As mentioned, they were not aware of the nature of these tests and did not know that they were being treated.[18] Furthermore, the instructors were not involved in any of these tests and none of them was aware of the treatment group to which each student was assigned. This provides us with a unique opportunity to examine the impact of providing relative performance feedback both on students' performance in the online assignment and, more importantly, in the course.

Each of the five groups consists, on average, of 220 students who were provided at all times with their (absolute) assignment performance information. The first group of students received feedback in the way we described above (i.e., for one minute only every time the server picked up a ranking variation; we will refer to this as the *Main* treatment). The control group received no information about ranking. The remaining three treatments consist of small variations of the Main treatment. We will call these treatments *Non-stop*, *Positive* and *Negative*. In the Non-stop treatment, ranking would be *constantly* displayed on the screen. In the Positive treatment, feedback would only appear if the student's ranking had *improved* and would be visible for one minute before disappearing. In the Negative treatment, feedback would only appear if the student's ranking had *worsen* and would be visible for one minute before disappearing. Thus, Non-stop conveys the same type of information as Main, but in a more constant manner. Positive and Negative provide feedback in the same way as Main (i.e., only when rank varies and only for one minute), but their informational content is not as rich. In what follows, we attempt to investigate how these feedback disclosure variations affect performance and identify the corresponding mechanism.

[18] During the entire semester, only a handful of students—in 1,101 participants—asked why their assignment did not display their ranking as it did for some of their peers. They were told that this feature was not available to all students as the developers were testing it. Interestingly, no one asked why they *did have* access to their ranking or why this was provided in a certain way, which suggests that students were not puzzled by the provision of feedback, nor by the way in which it was provided.

2. Theory and Predictions

In our set-up, the rank indicator is located in the top-right corner of the screen. In the Non-stop treatment this indicator is always visible. Hence, students are somewhat passively exposed to it, while their main focus is actively directed at solving the exercises from their online assignment. As students solve their exercises, the rank information is, by all intents and purposes, task-irrelevant and potentially relegated to the background of students' attention. Neuroscientists define this type of stimuli as 'bottom-up' signals that need to be salient in order to capture one's attention (Connor *et al.*, 2004): 'Volitional shifts of attention are thought to depend on "top-down" signals derived from knowledge about the current task (e.g., finding your lost keys), whereas the automatic "bottom-up" capture of attention is driven by properties inherent in stimuli, that is, by salience (e.g., a flashing fire alarm)' (Buschman and Miller, 2007). Bottom-up attentional mechanisms are more likely to be engaged by pop-out objects (i.e., objects that stand out from the visual search) and stimuli such as a suddenly appearing information box notifying a rank change (as in the Main, Positive, and Negative treatments), rather than by an information box constantly present in the background (as in the Non-stop treatment). Our Non-stop treatment might then fail to attract sufficient attention, whereas the remaining treatments (Main, Positive and Negative) are more likely to draw attention because of their pop-out nature.

Based on these insights, in the remainder of this section we assume that greater attention is paid to the rank change in the Main treatment relative to the Non-stop treatment. As for the remaining treatments, remember that Positive and Negative feature the same pop-out stimulus, *but* convey information less frequently (either only when rank goes up or when it goes down). As a result, feedback is by definition less frequent than in Main. We remain agnostic as to the comparison between Positive and Negative and between either of these treatments and Non-stop.

We are going to present two models in which students learn how they compare to their peers by receiving a noisy signal on their relative performance. Depending on the model, this signal is important either because it triggers competitive preferences or because it provides students with information about their own ability.

Although there might be several ways that students can acquire information regarding their relative standing in their cohort, we assume that our rank information facilitates this process by reducing the variance associated with the pre-existing noisy signal that students receive. We also assume that the more salient the rank indicator, the more accurate the signal (i.e., the signal will be at its most accurate in the Main treatment compared to Non-stop, Positive and Negative). As long as the model predicts that a more accurate signal results in less volatile outcomes, the Main students should also display less variance in their academic outcomes.

2.1. *The Models*

How does relative performance feedback affect students' education production function? The literature has identified two potential mechanisms (Ertac, 2005; Azmat and Iriberri, 2010), namely *competitive preferences* and *self-perception theory*. According to a model of competitive preferences, students know their own ability but have incomplete information about others'; they choose effort in order to maximise their utility function, which depends both on their absolute and their relative performance: *ceteris paribus*, they prefer to do better than others in the course, as opposed to worse. Self-perception theory assumes that students do not perfectly observe their own ability; by receiving feedback they update their belief and use this information to choose

their optimal level of effort, maximising an objective function which depends solely on their absolute performance in the course.

We follow and expand the theoretical framework proposed by Azmat and Iriberri (2010). Assume $N \geq 2$ students, whose individual abilities are independently and randomly drawn from a distribution F on the interval $[\underline{a}, \bar{a}]$, with $\underline{a} > 0$. Student i's performance in the course depends on their ability and on their chosen level of effort e_i, according to $p(a_i, e_i)$. Performance is increasing and strictly concave in ability and effort, that is, $\frac{\partial p(a_i,e_i)}{\partial a_i} > 0$, $\frac{\partial^2 p(a_i,e_i)}{\partial a_i^2} < 0$, $\frac{\partial p(a_i,e_i)}{\partial e_i} > 0$ and $\frac{\partial^2 p(a_i,e_i)}{\partial e_i^2} < 0$. Effort is costly and its cost, $c(e_i)$, is increasing and strictly convex; hence, $\frac{\partial c(e_i)}{\partial e_i} > 0$ and $\frac{\partial^2 c(e_i)}{\partial e_i^2} > 0$.

To rationalize our results and in the same spirit as Azmat and Iriberri (2010), we next present these two (parsimonious) models and compare their predictions to see which is more compatible with our data. First, we will consider a model in which students, who are aware of their own ability, exhibit competitive preferences and use relative performance feedback as a signal to infer the average performance of other students. We will then compare this set-up with a self-perception model in which students do not know their own ability and use feedback to learn it.

2.2. *Competitive Preferences Theory*

Consider the following utility function

$$u_i = p_i(a_i, e_i) - c(e_i) + \alpha \left(\frac{p_i(a_i, e_i) - E\left[\frac{1}{N}\sum_{k=1}^{N} p_k(a_k, e_k)\right]}{\sigma_{\bar{p}_k}} \right) \text{ for } i = 1, 2, \ldots, N. \quad (1)$$

The first difference compares the benefit and cost of effort, while the remaining term represents the competitive part of the utility function, with $\alpha > 0$ being the weight given to competitiveness. $E[\frac{1}{N}\sum_{k=1}^{N} p_k(a_k, e_k)]$ is the expectation of the average course performance among all students, while $\sigma_{\bar{p}_k}$ is the variance of such expectation. All else equal, a student prefers to be above (as opposed to below) the average performance and, when above (below) average performance, their utility is higher the greater (lower) the distance from the average. This implies that a grade of 8 (out of 10) generates more utility when the class average is 6, rather than 9. The higher the variance (i.e., the noisier the expectation), the less importance is attributed to competitiveness.

This approach to modelling competitiveness is close to Kandel and Lazear (1992). Charness and Rabin (2002) use instead a linear utility that includes others' payoffs, potentially negatively. Dubey and Genakoplos (2010) and Moldovanu *et al.* (2007) assume full knowledge of the complete ranking, with positive utility derived from the number of individuals below one's rank and negative utility from the number of individuals above it. Although some of the details would differ, these models would qualitatively yield similar results to those presented here.

Given that students know their own performance, we can write

$$E\left[\frac{1}{N}\sum_{k=1}^{N} p_k(a_k, e_k)\right] = \frac{1}{N} p_i(a_i, e_i) + \bar{p}_k,$$

where $\bar{p}_k = E[\frac{1}{N}\sum_{k \neq i} p_k(a_k, e_k)]$ is an unknown random variable assumed to be distributed according to $N(\mu_{\bar{p}_k}, \sigma^2_{\bar{p}_k})$. It follows that, because students in the control group do not receive

any feedback, their utility function can be expressed as

$$u_i^C = p_i(a_i, e_i) - c(e_i) + \alpha \left(\frac{p_i(a_i, e_i) - \left(\frac{1}{N} p_i(a_i, e_i) + \mu_{\bar{p}_k} \right)}{\sigma_{\bar{p}_k}} \right) \text{ for } i = 1, 2, \ldots, N.$$

Treated students receive additional feedback in the form of an informative signal. As we have four distinct treatments, let us call T_j the generic treatment j and $\bar{s}^{T_j} = \bar{p}_k + \epsilon^{Tj}$ the signal received by students in T_j. The error term ϵ^{T_j} is assumed to be normally distributed according to $N(0, (\sigma_\epsilon^{T_j})^2)$ and the two random variables, $\bar{p}_k$ and ϵ^{Tj}, are independently distributed. Treated students choose the optimal effort level conditioning on the signal $\bar{s}^{T_j}$. Hence, the objective function of student i in treatment T_j can be written as

$$u_i^{T_j} = p_i(a_i, e_i) - c(e_i) + \alpha \left(\frac{p_i(a_i, e_i) - \left(\frac{1}{N} p_i(a_i, e_i) + \mu_{\bar{p}_k|\bar{s}^{T_j}} \right)}{\sigma_{\bar{p}_k|\bar{s}^{T_j}}} \right) \text{ for } i = 1, 2, \ldots, N. \quad (2)$$

Building on the analysis by Azmat and Iriberri (2010), we establish the following proposition:

PROPOSITION 1. *Assume competitive preferences as modelled in (1):*

(*i*) *The optimal effort level of students in T_j is greater than the optimal effort level of students in the control group, for any ability.*
(*ii*) *Given two treatments T_a and T_b, with $(\sigma_\epsilon^{T_a})^2 < (\sigma_\epsilon^{T_b})^2$, the optimal effort level is higher in T_a than T_b for any ability a_i.*

PROOF. The proposition is composed of two parts. The first part coincides with Result 1 in Azmat and Iriberri (2010). Let us prove (*ii*). From (2) we can derive the first-order condition by taking the derivative with respect to effort and obtaining

$$\left[\frac{\sigma_{\bar{p}_k|\bar{s}^{T_j}} + \alpha \frac{N-1}{N}}{\sigma_{\bar{p}_k|\bar{s}^{T_j}}} \right] \frac{\partial p_i(a_i, e_i)}{\partial e_i} - c'(e_i) = 0.$$

As shown in Azmat and Iriberri (2010) (see 'Proof of Result 1'), $\sigma_{\bar{p}_k|\bar{s}^{T_j}}$ is distributed according to

$$N\left[\mu_{\bar{p}_k} + \frac{\sigma_{\bar{p}_k}^2}{\sigma_{\bar{p}_k}^2 + \sigma_\epsilon^2} \left(\bar{s}^{T_j} - \mu_{\bar{p}_k} \right), \frac{\sigma_{\bar{p}_k}^2 \sigma_\epsilon^2}{\sigma_{\bar{p}_k}^2 + \sigma_\epsilon^2} \right].$$

We can now take the first derivative of $\frac{\sigma_{\bar{p}_k}^2 \sigma_\epsilon^2}{\sigma_{\bar{p}_k}^2 + \sigma_\epsilon^2}$ with respect to σ_ϵ^2, which is equal to $(\frac{\sigma_{\bar{p}_k}^2}{\sigma_{\bar{p}_k}^2 + \sigma_\epsilon^2})^2$. As the latter is positive, it follows that $\sigma_{\bar{p}_k|\bar{s}^{T_j}}$ increases as σ_ϵ^2 increases, in turn lowering the optimal level of effort. As a direct consequence, given two treatments, the one that features a lower signal variance is the treatment with higher optimal effort level for any ability. □

The intuition behind this result is that a more precise signal of the average performance in the course makes the competitive component of the utility function more salient, triggering a bigger behavioural response that induces an increase in effort. This occurs because the competitive part of the utility function always increases the effort level—if effort were costless, that component would induce students to select maximum effort no matter what. Our relative performance feedback makes the expected average performance more precisely estimated, and that in turn entails more weight being given to the competitive part of the utility function. Because that

component increases the effort choice, the optimal choice of effort is higher. For this reason, the result is clear cut: (*i*) the additional relative performance feedback always increases the chosen level of effort; and (*ii*) given two treatments, the one associated with a more precise signal leads to a higher optimal effort level for any ability.

In order to flesh out the model's predictions in terms of heterogeneous treatment effects, let us consider the special case where effort and ability are perfect substitutes: $c(e_i) = e_i^2/2$ and $p(a_i, e_i) = a_i + e_i$. A quick inspection reveals that the optimal levels of effort and performance are $e_i^* = 1 + (\alpha/\sigma_{\bar{p}_k|\bar{s}^{T_j}})(1 - 1/N)$ and $p(a_i, e_i^*) = a_i + e_i^*$, respectively. It is easy to see that a change in the variance of the signal has the same effect on all students, irrespective or their level of ability.

Furthermore, given that academic performance $p(a_i, e_i)$ is naturally bounded, the model is potentially consistent with a ceiling effect for high-performing students, which in turn would imply that the academic outcomes would tend to be less dispersed the lower the variance of the signal. To see this point, take, for example, a student with very high ability. Even with little effort, they might achieve top course grades as captured, for example, by a high distinction (in the Australian system this corresponds to 85/100 course marks or above). Our treatment would have a relatively small effect on this type of high ability student, because this student is already achieving high distinction to begin with.

One final note regarding the dynamic effect of our relative performance feedback. Despite not being captured in our static model, it is worth noting that the signal $\bar{s}^{T_j} = \bar{p}_k + \epsilon^{Tj}$ is received by students in T_j during the entire semester. Therefore its informativeness reaches its peak at the time when the students finish the current semester and are about to start the next. Given that students are likely to meet a subset of their peers again in future semesters (or meet new peers drawn from a similar pool of candidates), the signal received during the current semester has the potential to affect the variance of the expected average performance of future courses as well. In that case, our model would predict that the above results could extend into the future as well, with treatment effects potentially reaching their peak in the period between the intervention semester and the next one.

2.3. *Self-Perception Theory*

Assume students do not perfectly observe their own ability and use feedback to learn about it. All students, whether treated or controlled, receive a noisy signal of their own ability

$$s_i = a_i + \eta \text{ for } i = 1, 2, \ldots, N.$$

Ability a_i is distributed according to $N(\bar{a}, (\sigma)^2)$, while the common shock η, which can be interpreted as the easiness of the exam, is distributed according to $N(0, (\psi)^2)$. Ability and the common shock are independently distributed. Treated students, through feedback, also observe the average signal

$$\bar{s} = \frac{\sum_{k=1}^{N} s_k}{N} = \frac{\sum_{k=1}^{N} (a_k + \eta)}{N} = \frac{\sum_{k=1}^{N} a_k}{N} + \eta.$$

Both s_i and (when observed) $\bar{s}$ are informative about a student's own ability. Self-perceived ability, in turn, determines the optimal effort level. To simplify the analysis, following Ertac (2005) and Azmat and Iriberri (2010), we will assume $p(a_i, e_i) = a_i e_i$, which implies complementarity between ability and effort (i.e., $\frac{\partial^2 p(a_i, e_i)}{\partial a_i \partial e_i} > 0$).

Students' utility function is the same as in (1) without the competitive part (i.e., with $\alpha = 0$). Those in the control group, who only receive a signal of their own ability, maximise

$$u_i^C = E\left[a_i e_i - c(e_i) \mid s_i\right] = E\left[a_i \mid s_i\right] e_i - c(e_i), \tag{3}$$

while treated students, who also receive a signal of the average ability, maximise

$$u_i^T = E\left[a_i e_i - c(e_i) \mid s_i, \bar{s}\right] = E\left[a_i \mid s_i, \bar{s}\right] e_i - c(e_i). \tag{4}$$

As proven by Ertac (2005) and by Azmat and Iriberri (2010) (see 'Proof of Result 2'), expected ability conditional on the individually observed signal and conditional on both the individually observed signal and the average signal are, respectively, equal to:

$$E[a_i \mid s_i] = \bar{a} + (s_i - \bar{a})\frac{\sigma^2}{\sigma^2 + \psi^2},$$

and

$$E[a_i \mid s_i, \bar{s}] = \bar{a} + (s_i - \bar{a}) - (\bar{s} - \bar{a})\frac{N\psi^2}{\sigma^2 + N\psi^2}.$$

Notice that in both expectations a student's private signal enters positively. Also, in the second expectation the average signal enters negatively. Observing a good average signal lowers beliefs about own ability because it increases the likelihood that the task was easy, which in turn decreases the probability that one's own ability was high. This result is in line with evidence from psychology suggesting that self-confidence is negatively affected by unfavourable social comparisons (see, for instance, Alicke, 2000).

Differentiating with respect to effort (3) and (4), we obtain the first-order conditions for the utilities of control and treated students, respectively:

$$\frac{\partial u_i^C}{\partial e_i} = 0 \qquad E[a_i \mid s_i] = c'(e_i^C)$$

$$\frac{\partial u_i^T}{\partial e_i} = 0 \qquad E[a_i \mid s_i, \bar{s}] = c'(e_i^T).$$

The comparison of the two first-order conditions reduces to the comparison of the conditional expected abilities. We find the signal such that the two conditional expectations are equal:

$$E[a_i \mid s_i] = E[a_i \mid s_i, \bar{s}] \text{ when } s^* = (\bar{s} - \bar{a})\frac{N\left(\sigma^2 + \psi^2\right)}{\sigma^2 + N\psi^2} + \bar{a}.$$

Because the cost function is convex, the right-hand side of either first-order condition is increasing in effort, from which the next proposition follows (see Azmat and Iriberri, 2010, 'Result 2'):

PROPOSITION 2. *There exists a threshold* $s^* = (\bar{s} - \bar{a})\frac{N(\sigma^2+\psi^2)}{\sigma^2+N\psi^2} + \bar{a}$ *such that* $e^{C*}(s_i) > e^{T*}(s_i, \bar{s})$ *for* $s_i < s^*$ *and* $e^{C*}(s_i) < e^{T*}(s_i, \bar{s})$ *for* $s_i > s^*$.

Proposition 2 tells us that, according to the self-perception theory, treated students should do better than control students at the higher end of the ability distribution, while the opposite should occur at the lower end. The threshold s^* is equal to the unconditional expected ability $\bar{s}$ if $\bar{s} = \bar{a}$ (i.e., when the average signal does not provide any information about the easiness of the exam). However, if $s^* \neq \bar{a}$, then s^* is higher than the average signal when $\bar{s} > \bar{a}$, while it is lower than

the average signal when $\bar{s} < \bar{a}$. The intuition behind this proposition is that, because effort and ability are assumed to be complements, by acquiring information about their own ability, better students are encouraged to exert more effort, while lower ability students are discouraged.[19]

Armed with these predictions we are ready to list a number of predictions based on these two models.

2.4. *Summary of the Theoretical Predictions*

First and foremost, by looking at the performance distribution among treated and control students, we should be able to see whether our data are compatible with either competitive preferences or with the self-perception theory.

Based on Proposition 1:

Prediction 1 (competitive preferences theory): If students exhibit competitive preferences of the type assumed in equation (1), then treated students should perform better in the course than control students, for any ability.

If instead students are not competitive, but use feedback to learn about their own ability and choose effort accordingly, then, according to Proposition 2:

Prediction 2 (self-perception theory): If students use feedback to update their beliefs about their own ability and optimise effort, then treated students perform better in the course than control students at the higher end of the ability distribution, while the opposite occurs at the lower end.

Based on our assumption that the signal will be at its most accurate in the Main treatment compared to Non-stop, Positive and Negative, and if students exhibit competitive preferences as in equation (1):

Prediction 3 (bottom-up selective attention): Main students should perform better than students in other treatments (as well as better than control students).

Furthermore, if effort and ability are perfect substitutes:

Prediction 4 (perfect substitutes): Treatment effects should be largely constant and positive across most of the ability distribution, except for a ceiling effect at the very top, and Main students should display less variance in their academic outcomes compared to the other treatments.

Finally, assuming that the informativeness of our relative performance signal reaches its peak at the time when the students finish the current semester and enter the next, and that students are likely to meet a subset of their peers again in future semesters (or meet new peers drawn from a similar pool of candidates):

Conjecture (persistent treatment effects): Relative performance feedback continues to be informative past the current (intervention) semester, triggering competitive preference in future semesters as well.

[19] Note that if we assumed a utility function such as: $p(a_i, e_i) = a_i + \frac{e_i}{a_i}$, which implies substitutability between ability and effort (i.e., $\frac{\partial^2 p(a_i, e_i)}{\partial a_i \partial e_i} < 0$), we would obtain the opposite result. Indeed the first-order conditions would look like: $\frac{1}{E[a_i|s_i]} = c'(e_i^C)$ and $\frac{1}{E[a_i|s_i, \bar{s}]} = c'(e_i^C)$. Hence, high ability students would rest on their laurels, while lower ability ones would be encouraged to put more effort compared to those in the control group. If, instead, we assumed that ability and effort are perfect substitutes, i.e., $p(a_i, e_i) = a_i + e_i$, then there would be no difference between the first-order condition of control and treated students and, consequently, releasing information would produce no effect.

3. Data and Empirical Analysis

This section provides an overview of the data and then discusses our empirical approach. As we examine the effects of our intervention directly on the course, the mechanism that drives them and whether they extend in any way to the longer-term, we will present these three cases separately.

3.1. *Data and Descriptive Statistics*

The data we use in our analysis come from university administrative records, as well as from the software developer logs and two course discussion boards. Specifically, our sample consists of 1,101 students, coming from 32 countries and taking the Principles of Microeconomics course in Semester 1, 2016. During this period, there were 46 tutorial classes available, taught by 16 different tutors. As mentioned, students are randomly assigned to tutorials at the beginning of the semester and cannot change their class at any point during the term. Below we provide evidence that this randomisation worked well.

Table 1 presents the main descriptive statistics for our sample at the student (Panel A.1) and tutorial level (Panel A.2). A quick glance reveals an almost 50/50 split between males and females. Nearly 78% of the students are Australian, while a significant proportion come from Asia (around 20%). Only 18% are economics students, while most of the others study degrees in commerce, business, and science, technology, engineering and mathematics (STEM henceforth). As for tutors, roughly 48% are males and almost 35% are international.

As discussed in Subsection 1.2, in Semester 1, 2016, the software developer of the online assignment adopted a prototype leaderboard system. This offers a unique opportunity to examine the effects on academic performance of small variations in the way students' relative performance feedback was displayed. Our identification strategy relies on developer's random assignment of students into five groups, based on the last digit of their ID number.

Table 2 reports differences in students' pre-determined characteristics across these groups, both overall and at tutorial level. Such characteristics refer to students' age, gender, degree undertaken, international student status and country of birth, as well as prior academic performance when available. Specifically, while we do have an ex-ante unified measure of prior ability for domestic students, we do not, unfortunately, possess a similar measure for international students. This is because the international students in our sample come from (32) different countries, all with different academic standards. As their high-school graduation and international university admission exams are of different scales and difficulty, the university where our intervention took place uses this information for admission purposes but does not maintain it in its records. As a result, for domestic students (roughly 78% of our sample), we proxy prior academic performance by their comparable high-school score, called the Australian Tertiary Admission Rank (ATAR hereafter).[20] For international students enrolled before Semester 1, 2016 (35% of the international subsample), we will use their previous semester GPA ('GPA previous semester: international' in Table 2).

We compare students' characteristics in each of the feedback groups to those in the control group (see Main, Panel A; Non-stop, Panel B; Positive, Panel C; Negative, Panel D). The first two columns in Table 2 report means and SDs for each treatment group, while the following two

[20] ATAR is the primary criterion for entry into Australian undergraduate programmes and denotes a student's high-school ranking relative to their peers when completing secondary education.

Table 1. *Descriptive Statistics.*

Variable	Mean	SD	Min.	Max.	*N*
Panel A.1: Student level characteristics					
Age	19.459	2.621	16	47	1,101
Male	0.565	0.496	0	1	1,101
Undertaking economics degree	0.183	0.386	0	1	1,101
International status	0.219	0.414	1	1	1,101
COB: Australia	0.777	0.416	0	1	1,101
COB: other Oceania	0.003	0.052	0	1	1,101
COB: Europe	0.011	0.104	0	1	1,101
COB: Asia	0.203	0.403	1	1	1,101
COB: America	0.002	0.043	0	1	1,101
COB: Africa and Middle East	0.004	0.060	0	1	1,101
ATAR score	90.146	7.555	49	99	861
GPA previous semester: international	4.509	1.141	1.667	6.75	86
Panel A.2: Tutorial level characteristics					
Male tutor	0.478	0.505	0	1	46
Tutor international status	0.348	0.482	0	1	46
Australian tutor	0.652	0.482	0	1	46
European tutor	0.065	0.250	0	1	46
Asian tutor	0.283	0.455	0	1	46
Panel B: Performance and effort indicators					
Week 6 exam	7.238	1.898	0	10	1,101
Week 10 exam	6.476	2.142	0	10	1,099
Final exam	6.203	1.681	0	10	1,101
Adjusted GPA Semester 1, 2016	4.867	1.119	1	7	1,080
GPA Semester 2, 2016	4.678	1.333	0	7	1,028
Assignment completion rate (%)	94.507	17.938	0	100	1,095
Time spent on assignment (hours)	10.161	6.322	0	61.351	1,093
Total number of posts	1.652	1.241	1	8	184
Number of relevant posts	1.467	1.267	0	8	184
Number of irrelevant posts	0.185	0.477	0	2	184
Posting (Y/N)	0.167	0.373	0	1	1,101

Notes: The classification of the country of birth (*COB*) follows the Standard Australian Classification of Countries, 2011. The Oceania group includes Oceania countries other than Australia. *ATAR* (Australian Tertiary Admission Rank) score denotes a student's ranking relative to their peers when completing secondary education. *GPA previous semester: international* is Semester 2, 2015, GPA for international students enrolled at the university before the intervention semester (Semester 1, 2016). *Assignment completion rate (%)* captures how much progress (as a percentage of all the assignment) a student has done in terms of completing all corresponding exercises. *Time spent on assignment (hours)* denotes the total number of hours a student spends attempting the assignment during the entire semester. *Total number of posts* refers to the posts a student contributes on the two course discussion boards, while *Number of relevant (irrelevant) posts* refers to those posts that are related (unrelated) to the course content, i.e., posts discussing (not discussing) economics topics. *Posting* is an indicator variable denoting whether a student has ever posted on any of the two discussion boards during the intervention semester.

display the same descriptive statistics for the control group. The last two columns present the differences in means between the two groups and the related SEs, respectively. We can quickly confirm that there are no statistically significant differences between treated students and those in the control group in *any* of the pre-determined characteristics at our disposal.[21]

Next, we also check if indeed students are randomly assigned to tutorials. To do so, we compare tutors' characteristics (at tutorial level) for each treated group with the control group. The figures reported in the bottom section of each panel in Table 2 prove this is the case. And an

[21] Accounting for prior ability and tutorial fixed effects leaves our balance checks unchanged, except for some country of birth group differences which become significant. All our specifications control for these indicators.

Table 2. *Balance Tests for Treatment and Control Groups in Semester 1, 2016.*

	Treatment group		Control group		Difference	
Variable	Mean	SD	Mean	SD	Diff.	SE
Panel A: Main feedback and control groups						
Age	19.384	(2.220)	19.602	(2.787)	0.218	(0.239)
Male	0.558	(0.498)	0.557	(0.498)	−0.002	(0.047)
Undertaking economics degree	0.170	(0.376)	0.195	(0.397)	0.025	(0.037)
International status	0.268	(0.444)	0.217	(0.413)	−0.051	(0.041)
COB: Australia	0.723	(0.448)	0.783	(0.413)	0.060	(0.041)
COB: other Oceania	–	–	–	–	–	–
COB: Europe	0.018	(0.133)	0.018	(0.134)	0.0002	(0.013)
COB: Asia	0.246	(0.431)	0.195	(0.397)	−0.051	(0.039)
COB: America	0.005	(0.067)	0.000	(0.000)	−0.005	(0.005)
COB: Africa and Middle East	0.009	(0.094)	0.005	(0.067)	−0.004	(0.008)
ATAR score	90.798	(7.365)	89.460	(8.304)	−1.338	(0.857)
GPA previous semester: international	4.258	(1.418)	4.258	(1.175)	0.000	(0.410)
Male tutor	0.496	(0.501)	0.502	(0.501)	0.007	(0.048)
Tutor international status	0.321	(0.468)	0.358	(0.480)	0.036	(0.045)
Australian tutor	0.679	(0.468)	0.643	(0.480)	−0.036	(0.045)
European tutor	0.067	(0.251)	0.086	(0.281)	0.019	(0.025)
Asian tutor	0.255	(0.437)	0.272	(0.446)	0.017	(0.042)
Panel B: Non-stop feedback and control groups						
Age	19.356	(2.375)	19.602	(2.787)	0.246	(0.251)
Male	0.572	(0.496)	0.557	(0.498)	−0.016	(0.048)
Undertaking economics degree	0.183	(0.387)	0.195	(0.397)	0.012	(0.038)
International status	0.178	(0.383)	0.217	(0.413)	0.039	(0.039)
COB: Australia	0.817	(0.387)	0.783	(0.413)	−0.035	(0.039)
COB: other Oceania	0.005	(0.069)	0.000	(0.000)	−0.005	(0.005)
COB: Europe	0.005	(0.069)	0.018	(0.134)	0.013	(0.010)
COB: Asia	0.168	(0.375)	0.195	(0.397)	0.026	(0.037)
COB: America	–	–	–	–	–	–
COB: Africa and Middle East	0.005	(0.069)	0.005	(0.067)	−0.0003	(0.007)
ATAR score	90.567	(7.680)	89.460	(8.304)	−1.108	(0.862)
GPA previous semester: international	4.633	(0.989)	4.258	(1.175)	−0.376	(0.428)
Male tutor	0.452	(0.499)	0.502	(0.501)	0.050	(0.048)
Tutor international status	0.385	(0.488)	0.358	(0.480)	−0.027	(0.047)
Australian tutor	0.615	(0.488)	0.643	(0.480)	0.027	(0.047)
European tutor	0.067	(0.251)	0.086	(0.281)	0.019	(0.026)
Asian tutor	0.317	(0.467)	0.272	(0.446)	−0.046	(0.044)

F-test reassures that it is indeed not possible to statistically reject the hypothesis that students' assignment is random, both at the course and at the tutorial level. Finally, we arrive at the same conclusion (i.e., no significant differences) also when comparing student characteristics and tutorial assignment across our different treatment arms.

In what follows, we will control for all pre-determined characteristics, except for prior academic performance. Our reason is three-fold. First, doing so would see our sample significantly reduced. For instance, controlling for ATAR would imply using only 78% of the students enrolled in the course, making our estimates much noisier. Second, doing so would also imply dropping a selective, non-random subsample of the population—all the international students. International students are, however, likely to differ from domestic students across various unobservable dimensions and may have a different valuation of their university degree.[22]

[22] Note, for instance, that international students pay much higher tuition fees, more than three times higher than domestic students.

Table 2. *Continued.*

	Treatment group		Control group		Difference	
Variable	Mean	SD	Mean	SD	Diff.	SE
Panel C: Positive feedback and control groups						
Age	19.367	(2.244)	19.602	(2.787)	0.234	(0.243)
Male	0.512	(0.501)	0.557	(0.498)	0.045	(0.048)
Undertaking economics degree	0.172	(0.378)	0.195	(0.397)	0.023	(0.037)
International status	0.219	(0.414)	0.217	(0.413)	−0.001	(0.040)
COB: Australia	0.777	(0.417)	0.783	(0.413)	0.006	(0.040)
COB: other Oceania	0.005	(0.068)	0.000	(0.00)	−0.005	(0.005)
COB: Europe	0.005	(0.068)	0.018	(0.134)	0.013	(0.010)
COB: Asia	0.214	(0.411)	0.195	(0.397)	−0.019	(0.039)
COB: America	–	–	–	–	–	–
COB: Africa and Middle East	0.000	(0.000)	0.005	(0.067)	0.005	(0.005)
ATAR score	89.661	(7.147)	89.460	(8.304)	−0.201	(0.839)
GPA previous semester: international	4.778	(0.931)	4.258	(1.175)	−0.520	(0.341)
Male tutor	0.447	(0.498)	0.502	(0.501)	0.056	(0.048)
Tutor international status	0.354	(0.479)	0.358	(0.480)	0.004	(0.046)
Australian tutor	0.647	(0.479)	0.643	(0.480)	−0.004	(0.046)
European tutor	0.056	(0.230)	0.086	(0.281)	0.030	(0.025)
Asian tutor	0.298	(0.458)	0.272	(0.446)	−0.026	(0.043)
Panel D: Negative feedback and control groups						
Age	19.571	(3.278)	19.602	(2.787)	0.031	(0.286)
Male	0.622	(0.486)	0.557	(0.498)	−0.066	(0.046)
Undertaking economics degree	0.193	(0.396)	0.195	(0.397)	0.001	(0.037)
International status	0.210	(0.408)	0.217	(0.413)	0.007	(0.039)
COB: Australia	0.790	(0.408)	0.783	(0.413)	−0.007	(0.039)
COB: other Oceania	0.004	(0.066)	0.000	(0.000)	−0.004	(0.004)
COB: Europe	0.009	(0.092)	0.018	(0.134)	0.010	(0.011)
COB: Asia	0.193	(0.396)	0.195	(0.397)	0.001	(0.037)
COB: America	0.004	(0.066)	0.000	(0.000)	−0.004	(0.004)
COB: Africa and Middle East	0.000	(0.000)	0.005	(0.067)	0.005	(0.004)
ATAR score	90.270	(7.220)	89.460	(8.304)	−0.811	(0.820)
GPA previous semester: international	4.731	(1.060)	4.258	(1.175)	−0.473	(0.358)
Male tutor	0.476	(0.501)	0.502	(0.501)	0.026	(0.047)
Tutor international status	0.348	(0.477)	0.358	(0.480)	0.010	(0.045)
Australian tutor	0.652	(0.477)	0.643	(0.480)	−0.010	(0.045)
European tutor	0.064	(0.246)	0.086	(0.281)	0.022	(0.025)
Asian tutor	0.283	(0.452)	0.272	(0.446)	−0.012	(0.042)

Notes: Each panel reports differences in pre-determined characteristics for students: in the Main (Panel A) and Non-stop (Panel B) treatment group versus the control group, respectively; and in the Positive (Panel C) and Negative (Panel D) treatment group versus the control group, respectively. The last two columns report the difference in means and the corresponding SE of the difference, respectively.

Moreover, while 60% of domestic students are male, this percentage drops to 43% among internationals. Of course, on top of controlling for ATAR we could also control for previous semester GPA for that minority of international students who enrolled at the university before the intervention semester. While this would slightly alleviate the small sample issue (i.e., we would only drop 14% of our total sample in this case), we would still be discarding a non-random subsample of the population—all of the first year international students. Furthermore, previous semester GPA is a noisier proxy for prior ability than ATAR. Finally, the software developer randomised students *solely* according to their student ID number and, thus, participants are unlikely to differ in their observed and unobserved characteristics across groups. Indeed,

Table 2 shows that the randomisation worked properly for *all* observable characteristics for which we have information (including ATAR for domestic students and previous semester GPA for internationals). Hence, it is highly improbable that it failed in a single dimension, i.e. (a uniform measure of), prior ability. For all these reasons, we control in our main analysis only for those characteristics which are available to us for the entire sample. (A robustness check that includes an imputed measure of prior ability shows that our results remain unchanged—see Section 6.)

In addition to the administrative data, we also have access to the software developer's logs. For each student, this database provides the following assignment-related data: (*i*) the final completion rate, (*ii*) the final success rate, (*iii*) the final rank, and (*iv*) the total amount of time spent on the assignment. Finally, two course-related discussion boards provide further information on peer interactions within the course, via the posts written on these forums by each student in our sample.

3.2. *Empirical Methodology*

3.2.1. *Course effects*

Our identification strategy relies on comparing the outcomes of students with similar characteristics, similar classmates and the same tutors, but who are exposed to different feedback treatments. We will analyse two different types of course effects. First, we will start by examining whether the various feedback treatments—to which students were randomly assigned—had any impact on students' performance in the online assignment, as indicated by their final assignment ranking. Next, we will examine the effect of feedback on academic performance as captured by students' grades in the invigilated course exams administered during the semester, namely the two mid-terms and the final exam. To capture the overall effect, we also use the weighted average exam grade computed as the mean of the three aforementioned tests, each scaled to take values between zero and ten—average exam grade henceforth (see Figure 2). Note that this information provides a good measure of students' learning and academic performance. First, we can rule out the possibility that the instructors may, in any conceivable way, artificially drive the effects, as none of them was aware of the treatment group to which each student was assigned. Second, all exams are closed-book, administered in class and invigilated, which makes them an objective measure of individual attainment. Third, the two mid-terms are marked—according to strict guidelines—by tutors, who are unaware of the experiment. As for the final exam, marking is entirely computerised. Finally, exam grades are not adjusted or re-weighted and so, they reflect each student's absolute performance in the course.

Our basic estimating equation takes the following form:

$$Y_{i,d,c} = \alpha + \beta Treatment_{i,d,c} + \gamma X_{i,d,c} + TutorialFE_c + u_{i,d,c}, \tag{5}$$

where $Y_{i,d,c}$ is either (*i*) the final assignment rank of student i, enrolled in degree d, attending tutorial class c or (*ii*) the grade achieved by student i, enrolled in degree d, attending tutorial class c in any of the three course exams, either separately or on average. The dummy variable $Treatment_{i,d,c}$ takes the value of one if a student is in one of the treatment groups and zero for control students. $X_{i,d,c}$ refers to students' characteristics (i.e., age, gender, dummies for countries of birth groups, as well as a dummy taking the value of one if a student is enrolled in an economics degree). To account for any systematic differences between students' learning

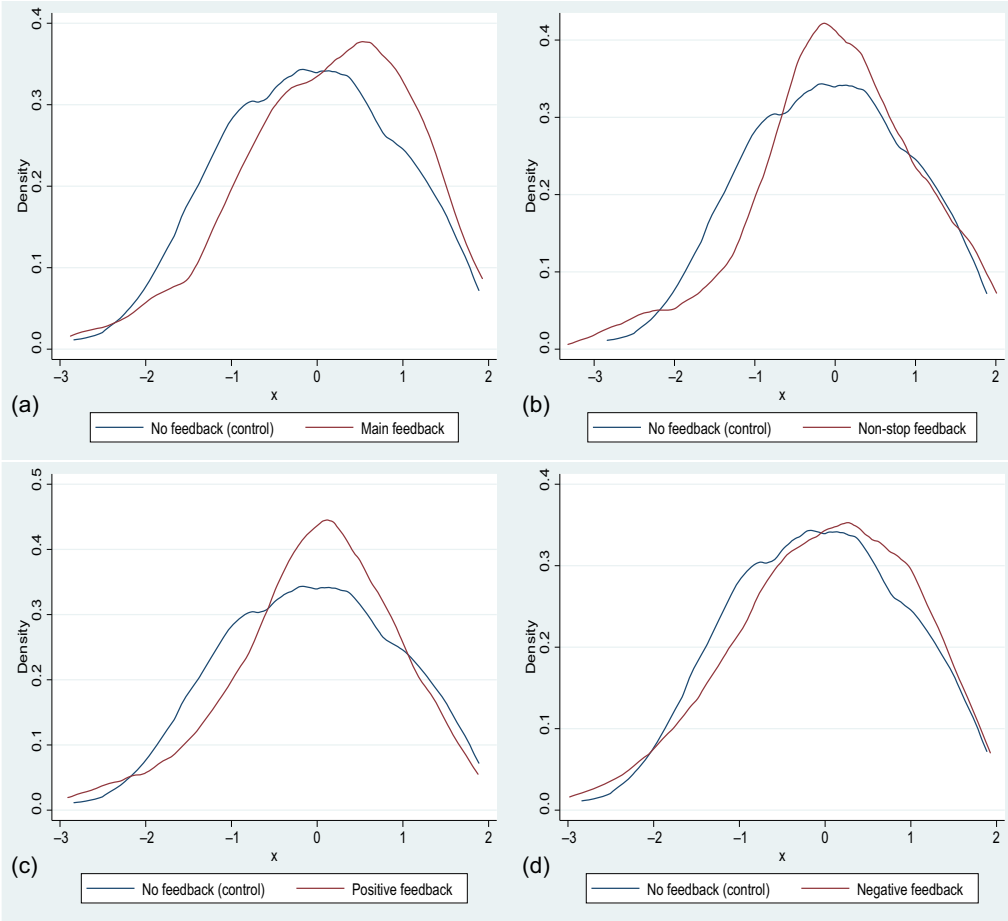

Fig. 2. *Densities of Average Exam Grades by Feedback Treatments.*

Notes: This figure shows the density functions for the standardised average exam grades that treated students achieve compared to their control counterparts. Plot (a) shows the densities for the Main treatment group compared to the control group; plot (b) shows the densities for Non-stop versus control; plot (c) shows the densities for Positive versus control; and plot (d) shows the densities for the Negative versus control.

experiences in tutorials, we include tutorial fixed effects. Standard errors are clustered at tutorial level to further account for the possibility of common shocks at the class level.

3.2.2. *Beyond course effects*

While our main focus is to identify the effects of feedback provision on the related course, it is also interesting to see whether our intervention had any long-lasting performance effects beyond the intervention course. Specifically, we explore whether learning one's ranking in the online assignment affects academic performance (*i*) in the other courses taken during the intervention semester as well as (*ii*) in the courses taken the following semester. To do so, we use two additional

outcome variables, namely Semester 1, 2016, GPA, adjusted to exclude the intervention course grade, and Semester 2, 2016, GPA. Both measures are continuously measured on a 0–7 scale.

On the one hand, students might end up putting more effort in the intervention course—in order to achieve and/or maintain a higher rank—and this, in turn, may come at the expense of other contemporaneous courses. Hence, feedback may indirectly affect students' performance in other courses taken during the intervention semester. On the other hand, as suggested by our model, the signal received during the intervention semester has the potential to affect the variance of the expected average performance in future courses as well. Being treated might also change students' behaviour more generally, with long-lasting effects on academic outcomes. This might be due to a higher propensity towards social learning, or to healthier study habits developed perhaps because of the extra incentives provided via leaderboards by the tighter effort–reward signal conveyed. Although our model seems to suggest this is not the likely channel, it is also possible that feedback provides information about one's self-perceived ability, while also resolving some of the uncertainty about returns to effort (Goulas and Megalokonomou, 2016). With this in mind, it will be interesting to explore our impacts' heterogeneity by gender, given the recent evidence that feedback provision might affect males and females differently (Mayo *et al.*, 2012; Goulas and Megalokonomou, 2016; Kugler *et al.*, 2017).

3.2.3. *The mechanism*

In an attempt to shed light on what drives students' response to feedback provision, we use additional data from the software developer's logs and the two course discussion boards. Both types of data provide us with a proxy for students' effort, either directly via a higher engagement with the online assignment or indirectly via a broader engagement with the course through social learning. For instance, the developer's logs provide data on the individual completion rate and time spent in the assignment, both overall and by set of exercises. To see whether feedback provision produces a greater direct engagement with the assignment, we will run a modified version of (5) with either the completion rate or the total amount of time spent working on the assignment as outcome variables. Furthermore, we use the two discussion boards associated with the course to obtain a measure of each student's total number of posts, while also categorising them as *relevant* and *irrelevant*.[23] These data allow us to examine whether students' reaction to feedback may be generated by a more active engagement with their peers in the course.

As a result, we employ the following regression model:

$$Posts_{i,d,c} = \alpha + \beta Treatment_{i,d,c} + \gamma X_{i,d,c} + TutorFE_c + u_{i,d,s}, \tag{6}$$

where $Posts_{i,d,c}$ refers to the total number of posts that student i, enrolled in degree d, assigned to tutorial class c writes on the discussion boards. We also run separate modified versions of (6) for specifications in which the outcome variable is either the number of relevant or the number of irrelevant posts, as well as whether one has ever posted on any of the two course discussion boards. In contrast to (5), note that (6) includes tutor fixed effects rather than tutorial fixed effects. This is because the number of students per treatment who post on the discussion boards is considerably smaller than the number of tutorial classes available, and so, in quite a

[23] Relevant posts relate to economic concepts, course content and materials; irrelevant posts refer, mostly, to course logistics or any other unrelated topic. The assignment of posts to these categories was done by two research assistants (RAs) who independently evaluated each post. When they disagreed on the classification of a post, which happened only a couple of times, a third RA classified that post. None of the RAs was aware of the research question.

few tutorial cases, there are an extremely low number of students posting or none whatsoever. In such situations, using tutorial fixed effects is not advisable (Greene, 2004). Including tutor fixed effects, on the contrary, is more appropriate because the total number of tutors is roughly one-third of the number of tutorial classes. Thus, including tutor fixed effects ensures that each tutor has contact with at least several students in the smaller subsample on which we have the course engagement data.

4. Results

4.1. *Baseline Estimates: Course Effects*

We start our results section by discussing estimates of the impact of feedback on students' performance in the online assignment. Results are reported in the first column in Table 3. The first row documents the effect of the Main feedback on students' final assignment ranking, as estimated via (5), while the remaining rows report impact estimates for the other three feedback treatments. The estimates show the Main feedback effect to be positive and statistically significant. The corresponding coefficient indicates that, by the end of the semester, *ceteris paribus*, students in the Main treatment group outrank their peers in the control by 62 (out of 1,093) positions. A different picture emerges, however, when looking at the other feedback groups. Estimates are still positive, but show no significant differences between the control students and those presented with the alternative feedback versions. That said, the impact of Non-stop is descriptively rather similar to the effect of the Main treatment (52 positions), while the effects of Positive and Negative are negligible.

We now turn to the effect of feedback on students' exams performance. Results are presented in Table 3, in (*i*) columns (2)–(4) and (5)–(7) for the two mid-terms, respectively, (*ii*) columns (8)–(10) for the final exam, and (*iii*) columns (11)–(13) for the average exam grade. For each exam, we show: (*i*) the actual performance—i.e., the grade out of 10 and rounded to two decimal places, (*ii*) the standardised performance—i.e., the grade transformed into z-scores to facilitate interpretation, and (*iii*) the ordinal grade rank to allow comparison with the assignment rank. The first row of Table 3 shows that students in the Main treatment group perform significantly better than control students across all course exams throughout the semester. Specifically, their grades are 0.31, 0.44 and 0.30 points (out of 10) higher than for control students in the first, second and final exam respectively, with an average performance rise of 0.34 points (out of 10). This translates into an improvement of 0.16, 0.21 and 0.18 SDs in Week 6, Week 10 and final exam respectively, for an average increase of 0.21 SDs. Finally, if we rank the grades of all students within each test, we note that the performance rise due to feedback provision translates into a rank increase almost identical to the one observed in the online assignment. Notably, this 62–63 exam positions boost appears constant across all three exams.

Again, a different picture emerges when we look at the alternative ways to provide rank feedback (vs. control), as reported in the remaining rows. For all these groups, the estimated coefficients are positive but not statistically different from zero. Similarly, when comparing them to the Main group, the estimates appear negative but not always significant. For instance, we find the Negative treatment to perform worse than Main, but no significant difference between Main and Non-stop (see Table A2). Thus, while we cannot definitely say that continuously presenting a student with their rank information is worse than showing it only when it varies, giving only bad news is clearly detrimental compared to the latter. Finally, we also run specifications that stack

Table 3. *Treatment Effects of Feedback on Academic Performance.*

	Assignment	Week 6 exam			Week 10 exam			Final exam			Average exam		
	Rank (1)	Perf. (2)	Std. (3)	Rank (4)	Perf. (5)	Std. (6)	Rank (7)	Perf. (8)	Std. (9)	Rank (10)	Perf. (11)	Std. (12)	Rank (13)
Main feedback	61.715 (31.858)*	0.311 (0.185)*	0.164 (0.098)*	61.948 (31.272)*	0.445 (0.214)**	0.208 (0.100)**	62.902 (33.063)*	0.299 (0.154)*	0.178 (0.092)*	63.582 (28.168)**	0.344 (0.159)**	0.210 (0.097)**	63.233 (25.825)**
Observations	442	445	445	445	444	444	444	445	445	445	444	444	445
Non-stop feedback	51.716 (33.790)	0.118 (0.232)	0.062 (0.122)	25.965 (38.492)	0.271 (0.195)	0.127 (0.091)	40.244 (28.699)	0.234 (0.178)	0.139 (0.106)	50.494 (34.501)	0.188 (0.171)	0.115 (0.104)	40.283 (29.659)
Observations	426	429	429	429	428	428	428	429	429	429	428	428	429
Positive feedback	6.762 (30.725)	0.006 (0.190)	0.003 (0.100)	16.851 (35.851)	0.135 (0.209)	0.063 (0.097)	21.328 (32.480)	0.087 (0.178)	0.052 (0.106)	18.986 (36.480)	0.084 (0.175)	0.051 (0.107)	18.581 (32.079)
Observations	434	436	436	436	434	434	434	436	436	436	434	434	436
Negative feedback	2.676 (36.429)	0.111 (0.199)	0.058 (0.105)	23.872 (34.900)	0.222 (0.201)	0.104 (0.094)	35.882 (29.321)	−0.035 (0.171)	−0.021 (0.102)	10.193 (32.453)	0.090 (0.166)	0.055 (0.101)	23.887 (28.824)
Observations	451	454	454	454	453	453	453	454	454	454	453	453	454
Pooled feedback	28.290 (23.062)	0.136 (0.149)	0.072 (0.078)	30.530 (26.497)	0.280 (0.156)*	0.131 (0.073)*	41.173 (23.965)*	0.124 (0.127)	0.074 (0.076)	31.353 (24.890)	0.172 (0.129)	0.105 (0.078)	34.927 (22.391)
Observations	1,093	1,101	1,101	1,101	1,099	1,099	1,099	1,101	1,101	1,101	1,099	1,099	1,101
Tutorial FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Student characteristics	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Notes: Each row presents estimates from separate OLS regressions. The dependent variable in column (1) is the ordinal rank a student achieves in the online assessment (out of 1,093). The dependent variable in columns (2)–(4) is the exam grade (out of 10), standardised exam grade and the grade rank in the first mid-term, while columns (5)–(7) and (8)–(10) refer to the same outcome formulations for the second mid-term and the final exam. Columns (11)–(13) average these series over all three course exams. The grade rank in each assessment is computed such that ties get assigned the average rank between the related two (equal) observations. *Pooled feedback* shows estimates from specifications that pool the four treatments (Main, Non-stop, Positive and Negative feedback) and compare them to the control group. In all specifications we include controls for age, gender, dummies for countries of birth groups, a dummy denoting whether a student is enrolled in an economics degree and tutorial fixed effects. Standard errors are clustered at tutorial level and reported in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% level, respectively.

together all four feedback treatments and compare them with control. The estimates labelled 'Pooled feedback' in Table 3 show that while we find significant (and positive) effects only in relation to Week 10 grades, all the other treatment estimates (e.g., on Week 6 exam, final exam, average exam) continue to point to the right direction, hence being indeed suggestive of an overall beneficial feedback effect.

Taken together, our Main feedback findings lend support to Prediction 3 and Prediction 4 and are indicative of Prediction 1 being more likely to be at play (rather than Prediction 2). Below we will indeed show this to be the case by specifically looking at the performance distribution. Additionally, we also note that the effect of the Main treatment (0.21 SDs) is remarkable. While in line with the effects documented in other feedback provision studies,[24] when comparing our findings with results from the education literature, we see, for instance, that our estimates are of comparable magnitude to being taught by a teacher between 1.5 and 2 SDs above the average (Hanushek *et al.*, 2005; Chetty *et al.*, 2014) or to reducing the class size by 20% (Angrist and Lavy, 1999; Krueger, 1999). While these are extremely costly interventions to implement, manipulating the way feedback is disclosed online is virtually costless to do in practice.[25] Moreover, from a practical perspective, feedback disclosure appears increasingly feasible by the day, due to the fast technological advancements that are quickly becoming an integral part of the education and training sector.

4.1.1. *Heterogeneity*

The results presented above are quite substantial, but they might vary greatly across different categories of students. To investigate the presence of heterogeneous treatment effects, we split the whole sample along various observable dimensions and re-run our benchmark specification in (5) separately for different subsamples. Results are reported in the Online Appendix, Tables A3–A6. For simplicity, we report only standardised performance, in all course exams and on average.

First, let us analyse the effect of feedback by gender as reported in Table A3. This dimension is particularly interesting as there exists a growing body of evidence showing that females are more sensitive to feedback interventions than males (Mayo *et al.*, 2012; Goulas and Megalokonomou, 2016; Kugler *et al.*, 2017). In our case, females in the Main treatment group indeed react to feedback provision more than males across all exams. Taken in isolation, female students in the Main treatment group perform, on average, better than females in the control group in all exams except for the first mid-term. This effect is quite substantial, amounting, on average, to a performance rise of 0.36 SDs. While still positive, the same effect drops to 0.10 SDs for males and is no longer statistically significant. Finally, the coefficients of interest for the alternative feedback groups (Non-stop, Positive, Negative) do not reveal any sizeable or significant pattern.

Second, note that another heterogeneous element of our results emerges when we focus on the effect of feedback by age. Table A4 reports results separately for students below 19 and above (or equal to) 19 years of age.[26] A quick glance reveals a more pronounced feedback effect for older students, statistically significant at the 5% level in all exams except the first

[24] See Azmat and Iriberri (2010), Bandiera *et al.* (2015) and Goulas and Megalokonomou (2016).

[25] We note that our intervention is not costless in absolute terms. It may be costless for the institution adopting it (both in the implementation phase and in terms of ongoing support), but to the extent that our effects are the result of higher effort, it is the student who bears the additional effort cost.

[26] The average student in our sample is 19.5 years old, hence our age split provides us with two rather comparable groups size-wise.

mid-term. In terms of magnitude, the coefficient appears almost three times larger for 19+ students than it is for those under 19, suggesting that the Main feedback may be more effective with a more mature sample. This is in line with Barankay (2012) who also reports a more sizeable effect of feedback among older salesmen. It contrasts, however, with Blanes i Vidal and Nossol (2011), who find that feedback provision is equally important for workers with different levels of experience. No further feedback alternatives provide any other significant treatment effects.

Third, a considerable proportion of our sample comes from abroad (see Subsection 3.1). In Table A5 we report treatment effects by international status, obtained by running regressions for each exam, separately for domestic and international students. The Main feedback impact is more prominent and more precisely estimated within the international subsample, albeit not statistically significant. Compared to domestic students, this effect is generally three to four times larger for internationals. No particular effects are present for any of the other feedback treatments too.

Finally, in Table A6 we look separately at how students majoring in economics respond to feedback compared to all other students. The effect of the Main feedback seems to be more pronounced, on average, among economics students, although not statistically different from zero. Indeed, we find no statistically significant effects of feedback in relation to any treatment group.

4.1.2. *Non-linearities*

Our baseline specification (5) assumes a linear impact of information provision on performance. The effect, however, may very well vary across the grade distribution. It is plausible, for instance, that feedback affects low- and high-achieving students differently. To address this issue, we allow for non-linear effects by running quantile regression models. We estimate the effect of feedback at each decile $\theta \in [0, 1]$ of the conditional distribution of grades as follows:

$$Y_{Quant_\theta} = \alpha_\theta + \beta_\theta Treatment_i + \gamma_\theta X_i + TutorFE + \epsilon_{i,\theta}. \tag{7}$$

Figure 3 plots the $\hat{\beta}_\theta$ coefficients from these quantile regressions (marginal effects) at each decile θ, as well as the associated 95% confidence interval. The outcome variable is the standardised average exam grade.[27] We use bootstrapping (with 500 replications) to compute the standard errors and estimate (7) for all four treatments.

Plot (a) in Figure 3 shows the effect of the Main feedback to be (largely) constant and positive across most of the distribution, lending support to our Prediction 1 (competitive preference) rather than Prediction 2 (self-perception theory). Although we observe the coefficients for only 3–4 deciles to be significant, the effect of rank incentives appears to increase for the first six deciles and then slightly decline at the highest decile. This type of pattern is also reported in Bandiera *et al.* (2015) and Goulas and Megalokonomou (2016), suggesting the presence of a ceiling effect or perhaps even a demotivating effect for the very best students. Taken as a whole, these results are in line with Prediction 4. As expected, none of the other treatments produces significant effects (see the other plots in Figure 3), except for Non-stop at the third decile. At a descriptive level, we note, however, that the effect of Non-stop and Negative is positive across the whole distribution. Vice versa, the bottom left plot in Figure 3 suggests a differential effect of the Positive treatment: the effect is positive up to the 60th percentile, becoming negative beyond

[27] Results are very similar when we use any of the three course exam grades in isolation.

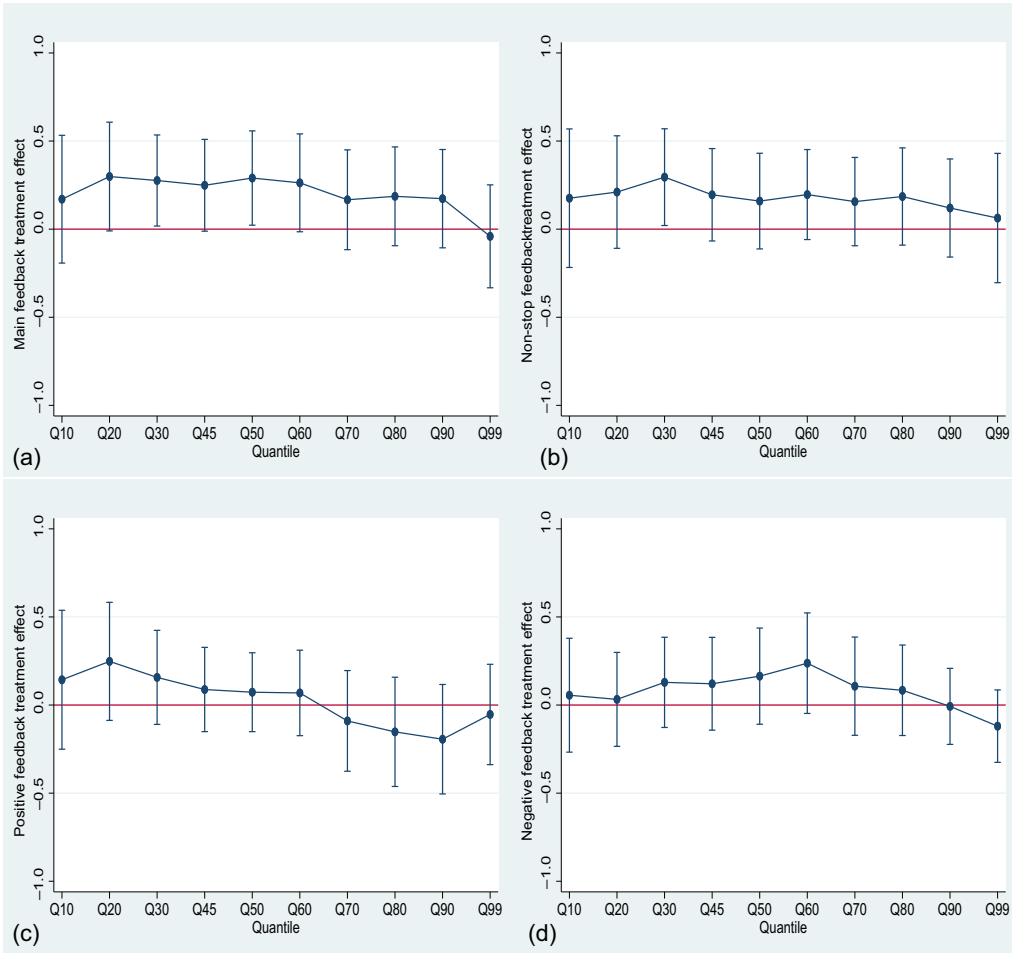

Fig. 3. *Quantile Regression Estimates of the Effect of Different Feedback Treatments.*
Notes: This figure presents the estimated quantile effects (marginal effects) of feedback provision on the standardised grades averaged across all exams at each decile, and the associated 95% confidence interval. The quantile regressions are conditional on students' age, gender, dummies for a student's country of birth and tutor fixed effects. We use bootstrapped standard errors with 500 repetitions. Plot (a) presents the quantile effects for Main treatment; plots (b), (c) and (d) present the quantile effects for the Non-stop, Positive and Negative treatment, respectively.

that point. Although these results are not significant, such pattern seems to suggest that reporting only positive news may hurt the best students.

4.2. *Beyond Course Effects*

The administrative records allow us to track students' performance in other courses taken in Semester 1, 2016, as well as in the subsequent semester. Table 4 reports evidence of the effect of feedback on these academic outcomes that go beyond the intervention course: column (1) refers to students' GPA in Semester 1, 2016, adjusted to exclude the intervention course grade,

Table 4. *Treatment Effects of Feedback Beyond the Intervention Course.*

	Adjusted GPA (std.) Semester 1, 2016	GPA (std.) Semester 2, 2016		
	(1)	(2)	(3)	(4)
Main feedback	0.093	0.285	0.184	0.167
	(0.095)	(0.114)**	(0.084)**	(0.083)*
Observations	437	416	414	414
Non-stop feedback	0.035	0.115	0.060	0.053
	(0.113)	(0.129)	(0.085)	(0.088)
Observations	426	399	398	398
Positive feedback	−0.035	0.089	0.108	0.100
	(0.122)	(0.113)	(0.079)	(0.078)
Observations	432	412	410	410
Negative feedback	0.029	−0.099	−0.090	−0.093
	(0.102)	(0.120)	(0.095)	(0.097)
Observations	445	422	418	418
Tutorial FE	✓	✓	✓	✓
Student characteristics	✓	✓	✓	✓
Adjusted GPA Semester 1, 2016	x	x	✓	✓
Performance in intervention course	x	x	x	✓

Notes: Each row presents estimates from separate OLS regressions. The dependent variables in column (1) and (2)–(4) are the standardised GPA of the intervention semester (Semester 1, 2016) adjusted to exclude the intervention course and the standardised GPA next semester (Semester 2, 2016), respectively. Adjusted current GPA and next semester GPA are measured continuously on a 0–7 scale. In all specifications, we include controls for age, gender, dummies for countries of birth groups, a dummy denoting whether a student is enrolled in an economics degree and tutorial fixed effects. Standard errors are clustered at tutorial level and reported in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% level, respectively.

while columns (2)–(4) focus on whether feedback has any long-lasting effects, by examining next semester GPA. The raw data for these two outcome variables takes continuous values between 0 and 7.

The estimated effects for the Main treatment are reported in the first row, while the remaining rows show the effects for our alternative feedback treatments. The fact that the Main feedback increases treated students' grades in the intervention course does not appear to come, however, at the expense of performance in other courses. Indeed, estimates in column (1) indicate that a student's average performance in all other courses taken in the same semester is not affected by feedback provision in the intervention course. We also find no difference in the number of other courses taken or passed, currently or in the next period (results available upon request). So, being treated did not crowd out effort in other courses.

The picture is very different when we look at the effect of the Main treatment on students' performance next semester. As reported in column (2), students in the Main feedback group experience a GPA increase in Semester 2, 2016, equal to 0.28 SDs. This is a remarkable effect that indeed supports our Conjecture (persistent treatment effects). That said, by performing this exercise, we are conflating two different channels: the direct long-term effect of feedback provision and its indirect impact through the intervention course. We are interested in the former. In order to isolate this direct effect, we also control for students' performance in all other courses taken in Semester 1, 2016, excluding the intervention course. Although this may introduce a slight endogeneity problem, it is the only way to net out the indirect effect. Moreover, because feedback does not affect performance in other contemporaneous courses, the potential

endogeneity does not appear overly problematic. As reported in column (3), the estimate drops to 0.18 SDs. Additionally, controlling for intervention course performance—see column (4)—leaves our estimates roughly unchanged (0.17 SDs). No further significant results are present for Non-stop, Positive or Negative.

These findings indicate that disclosing relative performance information can have a long-lasting positive impact on university students' academic performance. But is this behaviour driven by a particular subsample? Also, are these spillovers general or do they come from other economics courses? Table A7 presents estimates from specifications similar to the one in column (2) of Table 4, ran on separate subsamples split by gender and course type (economics vs. non-economics). We find that our long-term treatment effect is driven by male students, for whom we report a striking 0.38 SDs performance rise. We also note that this increase is not coming from subsequent economics courses. That said, this is most likely because 82% of our sample are students enrolled in degrees other than economics, for whom the option to take economics courses is quite limited. All in all, these results lend support to competitive preferences extending into next semester as relative performance information also extends into the future, as opposed to the alternative explanation that current semester economics learning is only useful in future related courses. This was already foreshadowed by the aggregate long-term effects remaining relevant even after controlling for students undertaking an economics degree and their (economics) learning in the intervention course, but it is a key point worth confirming.

5. The Mechanism

Our findings clearly show that students exposed to the Main feedback outperform control students both in the online assignment and, more importantly, in all course exams across the semester. Furthermore, this effect appears to be independent of a student's position across the grade distribution and is, thus, consistent with a model in which competitive preferences induce everyone to exert more effort (see Azmat and Iriberri, 2010). While we are not the first to document a positive impact of feedback provision on performance (e.g., Azmat and Iriberri, 2010; Tran and Zeckhauser, 2012; Katreniakova, 2014), up to our best knowledge, no other study has managed, so far, to provide direct evidence of the mechanism driving these results—an increase in the effort exerted. We are able to do so because of the RCT set-up, which presents two advantages. First, the feedback provided relates to a continuous drill, and its continuity allows us to observe students' activity over the entire semester. Second, the technology adopted to perform this drill helps us keep track of such activity.

With this in mind, note that effort can manifest itself in various ways. In our context, the most obvious ones are perhaps related to how one engages with the assignment. But it can also take the form of greater engagement at a higher level, both with the course and with other fellow students. To proxy for the first type of effort we will use two different measures of assignment engagement, namely the proportion of assignment completed and the amount of time spent doing it. To proxy for the second kind of effort, we will analyse several social learning indicators captured via the number (and type) of posts written by students on the two course discussion boards.

Table 1, Panel B, reports the relevant summary statistics on the assignment-related outcomes. Recall that 20% of the overall grade depends on completing 100% of the assignment; an equivalent proportion is awarded for partial completion. As we can see, 86% of students finish the assignment, with an average student completing 95% of it and spending roughly 10 hours doing so. This is a substantial amount of time, totalling about one-third of the overall face-to-face instruction

Table 5. *Treatment Effects of Feedback on Assignment Outcomes.*

	Assignment completion rate (1)	Time spent on assignment (2)
Main feedback	0.919	−0.061
	(1.604)	(0.728)
Observations	443	442
Non-stop feedback	0.079	−0.108
	(2.226)	(0.585)
Observations	427	426
Positive feedback	0.570	−1.419
	(1.670)	(0.594)**
Observations	435	434
Negative feedback	−1.861	−0.266
	(1.802)	(0.707)
Observations	453	451
Tutorial FE	✓	✓
Student characteristics	✓	✓

Notes: Each row presents estimates from separate OLS regressions. The dependent variable in column (1) is the assignment's completion rate. The dependent variable in column (2) is the total number of hours that a student spends on completing assignment exercises over the course of the semester. In both specifications we include controls for age, gender, dummies for countries of birth groups, a dummy denoting whether a student is enrolled in an economics degree and tutorial fixed effects. Standard errors are clustered at tutorial level and reported in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% level, respectively.

provided during the entire semester. Table 5 reports the ordinary least squares (OLS) estimates of the corresponding treatment effects.[28] We find that neither the completion rate nor the time spent working on the assignment are in any way robustly related with our treatments. Also, Positive treatment students do spend on average about 1.42 hours less on the assignment, but this does not seem to make a difference for their assessments. Thus overall, our treatments did not affect the way students engage with the assignment per se.

We now turn to the impact of our Main treatment (and other variations) on peer interactions, as captured by the posts that students upload on the two course discussion boards. We consider not only the total number of posts, but also their split into relevant and irrelevant. (The former group is related to the course material and assessments, while the latter includes logistics or fully unrelated issues.) Table 1 Panel B shows that students post on average 1.65 posts, with the relevant ones clearly representing the vast majority (mean = 1.46) and the irrelevant ones appearing only sparsely (mean = 0.19). Overall, 17% of students contribute to these forums (i.e., write at least one post in any of the course discussion boards).[29] We are interested in investigating whether any particular treatment group is more likely to post in the first place. The preliminary results in Table A8 show that students who post are very equally split among treatment groups and also compare well to the initial randomised group proportions. Importantly, note that treated students who post—even in this reduced sample—outperform control students. Indeed, we find

[28] One concern might be that our assignment time variable may not be very precise, as the server collects information about the total time students were logged in the software platform. There is no reason, however, to assume that this potential measurement error is different across treatments.

[29] There are no extreme outliers, with only three students posting more than six posts. Results continue to hold after dropping these observations.

Table 6. *Treatment Effects of Feedback on Course Engagement.*

	Total posts		Relevant posts		Irrelevant posts		Posting
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Compared to control group							
Main feedback	0.777	0.788	0.699	0.713	0.077	0.076	−0.009
	(0.325)**	(0.330)**	(0.348)*	(0.351)**	(0.108)	(0.110)	(0.033)
Observations	71	70	71	70	71	70	437
Non-stop feedback	0.184	0.286	0.081	0.214	0.102	0.072	0.044
	(0.342)	(0.344)	(0.337)	(0.326)	(0.090)	(0.083)	(0.035)
Observations	77	76	77	76	77	76	426
Positive feedback	−0.068	−0.061	−0.212	−0.223	0.143	0.162	−0.007
	(0.337)	(0.347)	(0.410)	(0.420)	(0.201)	(0.201)	(0.041)
Observations	69	69	69	69	69	69	432
Negative feedback	−0.024	0.014	−0.082	−0.048	0.058	0.062	0.006
	(0.306)	(0.288)	(0.308)	(0.288)	(0.105)	(0.130)	(0.035)
Observations	72	71	72	71	72	71	445
Panel B: Compared to main feedback							
Non-stop feedback	−0.396	−0.369	−0.453	−0.403	0.057	0.034	0.036
	(0.515)	(0.528)	(0.489)	(0.495)	(0.095)	(0.091)	(0.029)
Observations	78	78	78	78	78	76	423
Positive feedback	−0.508	−0.532	−0.460	−0.494	−0.048	−0.039	−0.009
	(0.318)	(0.298)*	(0.329)	(0.295)	(0.162)	(0.160)	(0.038)
Observations	70	69	70	69	70	69	429
Negative feedback	−0.673	−0.693	−0.757	−0.800	0.085	0.107	0.002
	(0.340)*	(0.340)**	(0.372)**	(0.375)**	(0.129)	(0.129)	(0.034)
Observations	73	71	73	71	73	71	442
Tutor FE	✓	✓	✓	✓	✓	✓	✓
Student characteristics	✓	✓	✓	✓	✓	✓	✓
Adjusted GPA Semester 1, 2016	x	✓	x	✓	x	✓	✓

Notes: Each row presents estimates from separate OLS regressions. The dependent variable in columns (1)–(2) is the total number of posts that students contribute to the two course discussion boards. Columns (3)–(4) and (5)–(6) show specifications with the number of relevant and irrelevant posts, respectively as dependent variable. The dependent variable in column (7) is a dummy taking the value one if a student has ever written on any of the discussion boards, and zero otherwise. In all specifications, we include controls for age, gender, dummies for countries of birth groups, a dummy denoting whether a student is enrolled in an economics degree and tutor fixed effects. Standard errors are clustered at tutorial level and reported in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% level, respectively.

a mean standardised average grade of treated students of 0.09 ($N = 36$), while for control this mean is 0.04 ($N = 35$).

Table 6 presents our regression results for (*i*) the total number of posts, in columns (1)–(2); (*ii*) the number of relevant posts, in columns (3)–(4); (*iii*) the number of irrelevant posts, in columns (5)–(6); and (*iv*) whether a student ever posted on any board during the semester. Specifications (1), (3) and (5) employ the baseline model (6), while (2), (4), (6) and (7) include extra controls for students' performance in the other Semester 1, 2016, courses. Panel A indicates that while there seems to be no treatment effect on whether one posts or not, the estimated impact of Main feedback (compared to no feedback) on number of posts is positive, sizeable, significantly different from zero, and robust across specifications. In other words, our intervention made active students even more active. In particular, conditional on posting, a student from the Main treatment

group writes, on average, 0.79 posts more than a control student. As the mean number of posts that control students write is close to 1.50, this is equivalent to a treated student writing 53% more often than a control one. Importantly, this result is entirely driven by the relevant posts, with a student in the Main treatment group writing about 0.71 more relevant posts than a control student. In contrast, the effect of feedback on the number of irrelevant posts that a treated student writes appears small and insignificant. This provides additional evidence supporting the social learning channel: students in the Main treatment react to feedback by getting more involved in the course and interacting more frequently with their peers. This seems to reflect their decision to exert more effort by engaging more actively in social learning.

One potential concern in this context is related to high-performing students perhaps being, on average, more 'vocal' on this type of forums. We are, however, not particularly worried about such selection as our treatment groups are balanced as far as prior ability is concerned. To allay any further concerns, we also control for a student's average performance in all other Semester 1, 2016, courses excluding the intervention course. (Note that Subsection 4.2 shows that feedback does not affect students' performance in any other contemporaneous courses.) Doing so hardly affects our results, with the corresponding estimates changing only by 2–6% (see columns (2), (4) and (6) in Table 6).

One additional remark. The increased participation in public forums could be problematic for our identification strategy if the additional posts generate positive externalities for the general student population (i.e., by being more engaged, the treated students benefit those in the control group as well). In this case, our impacts would be underestimated—they would capture the effect of being more active on public forums net of the positive externality.

Finally, we also explore if there are significant differences between our Main treatment effects and Non-stop, Positive and Negative in terms of posting behaviour, which is our proxy for social learning. Panel B in Table 6 shows that no such significant differences between Main and Non-stop are present, which is also the case for course performance when comparing the standardised average grades (see Table A2). In contrast, there are significant differences in posting behaviour between those receiving the Main feedback and those in the Positive or Negative group. In particular, students in the Positive (Negative) feedback group write 0.51 (0.67) posts less than students in the Main treatment group. Turning again to their grades, Table A2 shows that students in the Main treatment group outperform those in the Negative group in the final and average course exam grade by 0.18 and 0.15 SDs, respectively.

Overall, compared to control students, only students in the Main treatment group perform significantly better and are more socially engaged. One possible reason for this pattern is related to the effectiveness of feedback potentially relying on its ability to trigger attention, unwaveringly reminding a student of their current relative position with respect to their peers. Marketing experts, for instance, consider attention to be the key driver of advertising success: for consumers to be affected by an advert message, they first have to be paying attention. A similar thinking could apply in our context, but, while intriguing, this idea remains purely speculative. Alternatively, it could just be that we lack power, or perhaps our continuous-type feedback is very sensitive not only to the type of information disclosed, but also to the way in which it is disclosed. Finally, the treatment effects would also be underestimated if, by being more active on public forums, the treated students benefit those in the control group and, in turn, encourage them to post more. Further research is required to investigate these additional issues.

Table 7. *Placebo Treatment Effects of Feedback on Academic Performance.*

	Week 6 exam (1)	Week 10 exam (2)	Final exam (3)	Average exam (4)
Main feedback	0.134	0.055	0.126	0.119
	(0.113)	(0.102)	(0.086)	(0.100)
Observations	440	440	440	440
Non-stop feedback	0.139	0.030	−0.013	0.071
	(0.121)	(0.090)	(0.116)	(0.112)
Observations	441	440	441	440
Positive feedback	0.112	0.065	0.001	0.072
	(0.100)	(0.096)	(0.117)	(0.104)
Observations	440	440	440	440
Negative feedback	0.139	0.049	−0.033	0.070
	(0.109)	(0.099)	(0.100)	(0.104)
Observations	440	439	440	439
Tutorial FE	✓	✓	✓	✓
Students' characteristics	✓	✓	✓	✓

Notes: Each row presents estimates from separate OLS regressions. Data is generated as follows: first, we create a random variable, next we use it to sort the dataset and then we assign observations to placebo treatment groups. The dependent variables in columns (1)–(4) are the standardised exam grades as achieved in the first and second mid-term, the final exam and on average, respectively. In all specifications, we include controls for age, gender, dummies for countries of birth groups, a dummy denoting whether a student is enrolled in an economics degree and tutorial fixed effects. Standard errors are clustered at tutorial level and reported in parentheses. *, ** and *** denote significance at the 10%, 5% and 1% level, respectively.

6. Robustness

To alleviate potential concerns about any confounding factors which might affect our findings, we first perform a *randomised treatments*-type of robustness test. Specifically, we generate placebo treatments that do not reflect students' real treatment assignment and examine whether these treatments can produce a similar pattern as found in our main results. To do so, we randomly re-assign[30] all students into five groups and re-run our main model using the placebo treatments as variables of interest. If the placebo treatments are found to be significant determinants of performance, this would indicate that students might react to confounding factors (not perfectly coinciding with the real treatments) and get a performance boost.[31] If there is a correlation between these confounding factors and the assignment to actual treatment groups, then the placebo treatments would be picking up some of these effects. The results in Table 7 show no effect of these placebo treatments on performance. Hence, we conclude that our findings are unlikely to be driven by simultaneous effects other than our original treatments (as generated by the software developer).

We also wish to investigate whether our results may be, at least partially, driven by differential dropout rates across various treatments. Indeed, one may be concerned that feedback, instead of stimulating students to do better, may be discouraging the low-achieving ones and causing them to drop out at a higher rate than control students. We propose that students are unlikely to drop out from the course in response to the treatment group to which they are assigned. To support

[30] We create a random variable for all observations in our data and sort observations based on it. Using it, we then assign students to random treatment groups.

[31] For instance, students assigned to different treatments may create effective study groups and benefit from out-of-class interactions.

this claim, first note that students are eligible to drop out only during the first two weeks of the semester and, therefore, their limited exposure to the assignment in this time makes it unlikely that they quit because of the feedback treatment to which they are assigned. Furthermore, we look at the actual number of students who drop out from the course by treatment group. Column (2) in Table A9 shows a similar proportion of dropout students in all treatments (2.1–5.5% for all groups).

Another valid concern is related to whether there is a significant difference in the proportion of students who complete 100% of the online assignment. Indeed, students who complete the whole assignment are more exposed to the treatment than those who only partially complete it. Column (4) in Table A9 reports the number of students with 100% completion rate. We note that percentages are very similar across treatments (around 82–9%).

To account for prior ability, we also run a specification that includes an imputed measure of previous academic performance. As mentioned, previous ability (ATAR) scores are available only for domestic students, hence controlling directly for this variable would considerably reduce our sample. The missing prior ability values were thus imputed using the coefficients of (*i*) Semester 1, 2016, average course grade (adjusted to exclude the intervention course grade)[32] and (*ii*) our full set of controls (i.e., age, gender, countries of birth groups, and whether enrolled in an economics degree) from an OLS regression on the existing ATAR data (Dobrescu, 2015).[33] Table A10 estimates show that accounting for prior achievements leaves our treatment effects unchanged (i.e., robustly significant only for the Main group).

As mentioned, students were not aware of the nature of the leaderboard tests and did not know they were being treated. Class attendance was both voluntary and non-incentivised and during classes (*i*) neither the assignment exercises nor its structure and timeline were discussed, and (*ii*) no group work was undertaken. While all this goes some way to minimise the potential spillovers between our treatment arms, we cannot exclude this possibility. One avenue to investigate the magnitude of potential spillovers is to account in our baseline analysis (see Table 3) for the share of control students in one's tutorial class. Our estimates (available from the authors upon request) remain unchanged.

An additional concern worth investigating relates to the possibility of joint heterogeneous dimensions. For instance, the significant effect recorded for the 19+ year old subsample might be larger because females are more likely to be older than 19 and the female effect is substantial. To address this issue, we simultaneously include in our baseline specification all the interactions between our heterogeneity dimensions (e.g., age, gender, international status, field of study). We find that our main results remain unchanged (i.e., only the Main treatment remains statistically significant across the different assessments), with no further consistent significant estimates for the interaction terms. Combining the Main with the Non-stop group, and the Positive and Negative groups in an attempt to gain more power does not alter our findings (again, available upon request).

Our analysis so far has compared each treatment individually to the control group in order to capture the effect of different rank information disclosures in the most straightforward manner. As an extra robustness check, we are also pooling (*i*) the treatments, (*ii*) the standardised grades, and (*iii*) both the treatments and standardised grades, and re-run our baseline specifications as well as new ones that include interaction terms with gender. Specifically, we first pool all the standardised exam grades and run our baseline specification with treatment dummies for Main

[32] Note that our treatments are orthogonal to Adjusted GPA Semester 1, 2016 (see Table 4).

[33] Imputations affected 97.12% of the international sample, with the specification attaining an R^2 of 22.72%.

feedback against control, Non-stop feedback against control, and pooled Main and Non-stop feedback against control, each with and without interaction terms for gender in even- and odd-numbered specifications, respectively. Results are presented in the Online Appendix Table A11 and show significant treatment estimates for both sets of specifications involving the Main treatment group. Doing the same in the case of Positive and Negative groups yields no results, while gender interaction terms appear insignificant across the board. Next, we pool treatments and proceed in a similar fashion as for Table A11—also see Table A12. Panel A shows the effect of the Main feedback compared to all other treatments and control for each course exam grade, as well as for the average exam. Panel B does the same but compares the pooled Main and Non-stop feedback group to all other treatments and control. In both cases, our specifications (*i*) maintain significance for Week 10 exam, final exam and average exam, and (*ii*) yield no consistently significant gender-related estimates. Finally, we pool both the standardised grades and the treatments and proceed like before. Table A13 shows results (*i*) separately for Main feedback group against all the other treatment groups and control, as well as for the pooled Main and Non-stop feedback groups against all the other treatment groups and control. Similarly to our first two attempts in Tables A11–A12, we find significant treatment effects when the Main feedback treatment is involved and no additional results related to gender.[34]

Finally, we also re-run our assignment outcomes analysis with tutor fixed effects to ensure that the lack of results related to this type of effort is not an artefact of the fixed effects nature. Doing so does not change our assignment results, confirming that only our social learning mechanism is at play.

7. Concluding Remarks

Building a unified body of knowledge around the effectiveness of rank feedback is a difficult task because the behavioural response to rank feedback per se (rank incentives) is potentially compounded with many confounding factors. This paper is among the first to cleanly identify the impact of rank incentives and to shed light on its underlying mechanism.

In a higher education setting, we find that providing rank feedback had a sizeable positive impact on student performance (0.21 SDs higher exam grades) did not crowd out effort in the other contemporaneous courses and improved academic performance next semester.

The reason why our intervention was successful, compared to similarly well identified rank incentives studies (Barankay, 2011; 2012) or the other prominent randomised control trials in higher education (e.g., Azmat *et al.*, 2019), might lie in how salient, visible, and immediately actionable our feedback was. In our RCT, students were (randomly) assigned to four treatments that privately presented them with information on their real-time rank as achieved in a semester-long online assignment. Treatments varied in how often the rank feedback was displayed and in what type of information was presented: some students were shown their rank every time it changed either upwards or downwards (Main treatment), others were uninterruptedly exposed to their relative performance position (Non-stop), or they could only see it when their rank position improved (Positive) or worsened (Negative).

A number of instructional design issues emerge as potential avenues for further research. The timing of rank feedback provision seems to be crucial. Indeed, our results suggest that providing it

[34] We note that the loss of significance in specifications (2) and (4) that include gender effects is restored when the dummies for countries of birth groups are replaced by an international student indicator.

non-stop or not often enough might render the rank information not salient. This raises questions on the optimal frequency of information provision, with an eye on the balance of interest versus information overload.

How people receive the information also appears to matter greatly. In our case, feedback was provided in real time and continuously during the semester. Because real-time feedback directly resolves some of the uncertainty about returns to effort (and it does so in a fairly granular, decision-by-decision, manner), restricting it to a specific period would have likely limited its impact. With current technologies making the provision of real-time feedback over long periods virtually costless, it will be interesting to see if this type of approach can reduce the demoralisation effect (due to a lower-than-expected rank) also in other contexts.

The granularity of the information might also have played a key role. In our RCT, students learnt their exact rank rather than whether they belonged in a specific band. Providing the information partitioned differently would have made the rank changes more difficult to spot and would have likely triggered a different effort response. This direction of research warrants further study and may unveil some potentially interesting lessons.

Finally, an important avenue for future research has to do with the mechanism behind the impact of rank feedback. In our case, results seem to be driven by social learning, i.e., the extent to which a student engages with their peers by posting on the two course discussion boards. Main treatment students (and only they) do so 50% more often than control students—a considerable, robust and statistically significant effect. More research is required, however, to shed further light on how the education production function is affected by rank feedback.

Our findings have considerable policy implications. Improving students' attainments is a priority for all policymakers and practitioners who tend to focus on a variety of inputs, such as (*i*) reducing class size (Krueger, 1999; Bedard and Kuhn, 2008), (*ii*) improving quality of teachers (Glewwe *et al.*, 2010; Chetty *et al.*, 2014; Duflo *et al.*, 2015) and schools (Lavy, 2002),[35] (*iii*) extending term length (Pischke, 2007; McMullen and Rouse, 2012), (*iv*) improving peer group quality (Zimmerman, 2003; Duflo *et al.*, 2011), (*v*) providing financial and non-financial incentives (Benhassine *et al.*, 2015; Levitt *et al.*, 2016), and (*vi*) employing more student-level differentiation (Banerjee *et al.*, 2016), using frequent data to tailor classroom instruction and instilling a culture of high expectations (Abdulkadiroglu *et al.*, 2011; Fryer, 2014). All such interventions are, however, very costly to administer and their effectiveness is uncertain. Technology, on the other hand, is increasingly seen as the leading cost-effective avenue to boost instruction productivity (Bill and Melinda Gates Foundation, 2016; Mead, 2016). There are several channels through which this might occur, from shortening feedback time to creating environments that trigger people's engagement. We shed light on these issues and show that providing feedback in such contexts is technically feasible, beneficial for human capital accumulation and virtually costless to implement.

Beyond higher education, our findings bring important implications for the design of effective information—and potentially compensation—systems more generally. For instance, our intervention suggests that firms could benefit from providing real-time rank feedback to their employees, even in the absence of an associated reward (or punishment) structure. This is a significant insight in line with the notion that feedback is a key element of employee satisfaction with the job itself (Lam *et al.*, 2002). Better feedback could then create greater job satisfaction, which in turn contributes to employee well-being which has been causally linked to higher productivity

[35] See also Rockoff (2004), Rivkin *et al.* (2005), Aaronson *et al.* (2007) and Kane and Staiger (2008).

(Oswald *et al.*, 2015). And it may also boost one's confidence in having the skills needed to tackle a particular task, which then increases their motivation to complete that task (Benabou and Tirole, 2002), especially because rank feedback seems to trigger competitive preferences in a way that affects the whole distribution, i.e., demotivation following less than ideal feedback is less likely to be a concern. Finally, the long-term impacts of rank incentives that we report could form the basis of behavioural systems of incentives that foster the creation of good work habits.

University of New South Wales, Australia
University of Queensland, Australia
University of Queensland, Australia
University of New South Wales, Australia

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

References

Aaronson, D., Barrow, L. and Sander, W. (2007). 'Teachers and student achievement in the Chicago public high schools', *Journal of Labor Economics*, vol. 25(1), pp. 95–135.
Abdulkadiroglu, A., Angrist, J., Dynarski, S., Kane, T. and Pathak, P. (2011). 'Accountability in public schools: Evidence from Boston's charters and pilots', *Quarterly Journal of Economics*, vol. 126(2), pp. 699–748.
Alicke, M. (2000). 'Evaluating social comparison targets', in (J. Suls and L. Wheeler, eds.), *Handbook of Social Comparison: Theory and Research*, The Plenum Series in Social/Clinical Psychology, pp. 271–94, New York: Springer.
Angrist, J. and Lavy, V. (1999). 'Using Maimonides' rule to estimate the effect of class size on scholastic achievement', *Quarterly Journal of Economics*, vol. 114(2), pp. 533–75.
Azmat, G., Bagues, M., Cabrales, A. and Iriberri, N. (2019). 'What you don't know ... can't hurt you: A field experiment on relative performance feedback in higher education', *Management Science*, vol. 65(8), pp. 3714–36.
Azmat, G. and Iriberri, N. (2010). 'The importance of relative performance feedback information: Evidence from a natural experiment using high school students', *Journal of Public Economics*, vol. 94(7–8), pp. 435–52.
Azmat, G. and Iriberri, N. (2016). 'The provision of relative performance feedback: An analysis of performance and satisfaction', *Journal of Economics and Management Strategy*, vol. 25(1), pp. 77–110.
Bandiera, O., Barankay, I. and Rasul, I. (2013). 'Team incentives: Evidence from a firm level experiment', *Journal of the European Economic Association*, vol. 11(5), pp. 1079–114.
Bandiera, O., Larcinese, V. and Rasul, I. (2015). 'Blissful ignorance? Evidence from a natural experiment on the effect of individual feedback on performance', *Labor Economics*, vol. 34, pp. 13–25.
Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M. and Walton, M. (2016). 'Mainstreaming an effective intervention: Evidence from randomized evaluations of "teaching at the right level" in India', Working paper no. 22746, NBER.
Barankay, I. (2011). 'Rankings and social tournaments: Evidence from a crowd-sourcing experiment', Working paper, University of Pennsylvania.
Barankay, I. (2012). 'Rank incentives: Evidence from a randomized workplace experiment', Working paper, University of Pennsylvania.
Bedard, K. and Kuhn, P. (2008). 'Where class size really matters: Class size and student ratings of instructor effectiveness', *Economics of Education Review*, vol. 27(3), pp. 253–65.
Benabou, R. and Tirole, J. (2002). 'Self-confidence and personal motivation', *Quarterly Journal of Economics*, vol. 117(3), pp. 871–915.
Benhassine, N., Devoto, F., Duflo, E., Dupas, P. and Pouliquen, V. (2015). 'Turning a shove into a nudge: A "labeled cash transfer" for education', *American Economic Journal: Economic Policy*, vol. 3(7), pp. 86–125.
Bill and Melinda Gates Foundation. (2016). 'Finding what works: Results from the leap innovations pilot network', Technical report, Bill and Melinda Gates Foundation.
Blader, S., Gartenberg, C.M. and Prat, A. (2016). 'The contingent effect of management practices', Research paper no. 15-48, Columbia Business School.
Blanes i Vidal, J. and Nossol, M. (2011). 'Tournaments without prizes: Evidence from personnel records', *Management Science*, vol. 57(10), pp. 1721–36.

Brade, R., Himmler, O. and Jäckle, R. (2020). 'Relative performance feedback and the effects of being above average—field experiment and replication', Working paper no. 88830, MPRA.
Bursztyn, L. and Jensen, R. (2015). 'How does peer pressure affect educational investments?', *Quarterly Journal of Economics*, vol. 130(3), pp. 1329–67.
Buschman, T. and Miller, E. (2007). 'Top-down versus bottom-up control of attention in prefrontal and posterior parietal cortices', *Science*, vol. 315, pp. 1860–2.
Charness, G., Masclet, D. and Villeval, M.C. (2013). 'The dark side of competition for status', *Management Science*, vol. 60(1), pp. 38–55.
Charness, G. and Rabin, M. (2002). 'Understanding social preferences with simple tests', *Quarterly Journal of Economics*, vol. 117(3), pp. 817–69.
Chen, Y., Harper, F., Konstan, J. and Li, S.X. (2010). 'Social comparisons and contributions to online communities: A field experiment on MovieLens', *American Economic Review*, vol. 100(4), pp. 1358–98.
Chetty, R., Friedman, J. and Rockoff, J. (2014). 'Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood', *American Economic Review*, vol. 104(9), pp. 2633–79.
Connor, C., Egeth, H. and Yantis, S. (2004). 'Visual attention: Bottom-up versus top-down', *Current Biology*, vol. 14, pp. R850–2.
Dobrescu, L. (2015). 'To love or to pay: Savings and health care in older age', *Journal of Human Resources*, vol. 50(1), pp. 254–99.
Dubey, P. and Genakoplos, J. (2010). 'Grading exams: 100, 99, 98, ... or A, B, C?', *Games and Economic Behavior*, vol. 69(1), pp. 72–94.
Duflo, E., Dupas, P. and Kremer, M. (2011). 'Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya', *American Economic Review*, vol. 101(5), pp. 1739–74.
Duflo, E., Dupas, P. and Kremer, M. (2015). 'School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools', *Journal of Public Economics*, vol. 123, pp. 92–110.
Ely, J.C. (2017). 'Beeps', *American Economic Review*, vol. 107(1), pp. 31–53.
Eriksson, T., Poulsen, A. and Villeval, M.C. (2009). 'Feedback and incentives: Experimental evidence', *Labour Economics*, vol. 16, pp. 679–88.
Ertac, S. (2005). 'Social comparisons and optimal information revelation: Theory and experiments', Working Paper, University of California.
Fryer, R. (2014). 'Injecting charter school best practices into traditional public schools: Evidence from field experiments', *Quarterly Journal of Economics*, vol. 129(3), pp. 1355–407.
Gerhards, L. and Siemer, N. (2014). 'Private versus public feedback—the incentive effects of symbolic awards', Working paper, University of Aarhus.
Gerhards, L. and Siemer, N. (2016). 'The impact of private and public feedback on worker performance: Evidence from the lab', *Economic Inquiry*, vol. 54(2), pp. 1188–201.
Gill, D., Kissová, Z., Lee, J. and Prowse, V. (2019). 'First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision', *Management Science*, vol. 65(2), pp. 459–507.
Glewwe, P., Ilias, N. and Kremer, M. (2010). 'Teacher incentives', *American Economic Journal: Applied Economics*, vol. 2(3), pp. 205–27.
Goulas, S. and Megalokonomou, R. (2016). 'Knowing who you actually are: The effect of feedback on short and long term outcomes', Research paper no. 1075, University of Warwick.
Greene, W. (2004). 'Fixed effects and bias due to the incidental parameters problem in the Tobit model', *Econometric Reviews*, vol. 23(2), pp. 125–47.
Hannan, R.L., Krishnan, R. and Newman, A.H. (2008). 'The effects of disseminating relative performance feedback in tournament versus individual performance compensation plans', *The Accounting Review*, vol. 83(4), pp. 893–913.
Hanushek, E., Kain, F. and Rivkin, G. (2005). 'Teachers, schools and academic achievement', *Econometrica*, vol. 2(73), pp. 417–58.
Jessoe, K. and Rapson, D. (2014). 'Knowledge is (less) power: Experimental evidence from residential energy use', *American Economic Review*, vol. 104(4), pp. 1417–38.
Kandel, E. and Lazear, E. (1992). 'Peer pressure and partnerships', *Journal of Political Economy*, vol. 100(4), pp. 801–17.
Kane, T. and Staiger, D. (2008). 'Estimating teacher impacts on student achievement: An experimental validation', Working paper 14607, NBER.
Katreniakova, D. (2014). 'Information, aspirations and incentive to learn: A randomized field experiment in Uganda', Working paper, CERGE-EI.
Krueger, A. (1999). 'Experimental estimates of education production functions', *Quarterly Journal of Economics*, vol. 114(2), pp. 497–532.
Kugler, A., Tinsley, C. and Ukhaneva, O. (2017). 'Choice of majors: Are women really different from men?', Working paper no. 23735, NBER.
Kuhnen, C.M. and Tymula, A. (2012). 'Feedback, self-esteem, and performance in organizations', *Management Science*, vol. 58(1), pp. 94–113.
Lam, S., Yik, M. and Schaubroeck, J. (2002). 'Responses to formal performance appraisal feedback: The role of negative affectivity', *Journal of Applied Psychology*, vol. 87(1), pp. 192–201.

Lavy, V. (2002). 'Evaluating the effect of teachers' group performance incentives on pupil achievement', *Journal of Political Economy*, vol. 110(6), pp. 1286–317.
Levitt, S., List, J., Neckermann, S. and Sadoff, S. (2016). 'The behavioralist goes to school: Leveraging behavioral economics to improve educational performance', *American Economic Journal: Economic Policy*, vol. 8(4), pp. 183–219.
Mas, A. and Moretti, E. (2009). 'Peers at work', *American Economic Review*, vol. 99(1), pp. 112–45.
Mayo, M., Kakarika, M., Pastor, J. and Brutus, S. (2012). 'Aligning or inflating your leadership self-image? A longitudinal study of responses to peer feedback in MBA teams', *Academy of Management Learning & Education*, vol. 11(4), pp. 631–52.
McMullen, S. and Rouse, K. (2012). 'The impact of year-round schooling on academic achievement: Evidence from mandatory school calendar conversions', *American Economic Journal: Economic Policy*, vol. 4(4), pp. 230–52.
Mead, R. (2016). 'Learn different: Silicon Valley disrupts education', *New Yorker*, 5 March.
Moldovanu, B., Sela, A. and Shi, X. (2007). 'Contests for status', *Journal of Political Economy*, vol. 115, pp. 338–63.
Oswald, A., Proto, E. and Sgroi, D. (2015). 'Happiness and productivity', *Journal of Labor Economics*, vol. 33(4), pp. 789–822.
Pischke, J. (2007). 'The impact of length of the school year on student performance and earnings: Evidence from the German short school years', Economic Journal, vol. 117(523), pp. 1216–42.
Rivkin, S., Hanushek, E. and Kain, J. (2005). 'Teachers, schools and academic achievement', *Econometrica*, vol. 73(2), pp. 417–58.
Rockoff, J. (2004). 'The impact of individual teachers on student achievement: Evidence from panel data', *American Economic Review*, vol. 94(2), pp. 247–52.
Tiefenbeck, V., Goette, L., Degen, K., Tasic, V., Fleisch, E., Lalive, R. and Staake, T. (2016). 'Overcoming salience bias: How real-time feedback fosters resource conservation', *Management Science*, vol. 64(3), pp. 1458–76.
Tran, A. and Zeckhauser, R. (2012). 'Rank as an inherent incentive: Evidence from a field experiment', *Journal of Public Economics*, vol. 96(9), pp. 645–50.
Wedel, M. and Pieters, R., eds. (2012). *Visual Marketing: From Attention to Action*, London: Psychology Press.
Wolitzky, A. (2018). 'Learning from others' outcomes', *American Economic Review*, vol. 108(10), pp. 2763–801.
Zimmerman, D. (2003). 'Peer effects in academic outcomes: Evidence from a natural experiment', *Review of Economics and Statistics*, vol. 85(1), pp. 9–23.