Normatively Framed Relative Performance Feedback – Field Experiment and Replication

Raphael Brade and Oliver Himmler and Robert Jäckle

5 September 2018

Normatively Framed Relative Performance Feedback – Field Experiment and Replication

RAPHAEL BRADE[a] OLIVER HIMMLER[b] ROBERT JÄCKLE[c]

September 5, 2018

Abstract

Feedback can help individuals put their performance into perspective, especially when transitioning into a new environment such as university or a different job. In a randomized field experiment we give first-year university students normatively framed relative performance feedback about their accumulated course credits. We find an increase in subsequent performance, but only when the feedback is positive. Using a regression discontinuity design, we show that the improved performance is not driven by unobserved characteristics of those receiving positive feedback, but that it is indeed due to the positive rather than negative nature of the feedback. We administer a replication experiment with the next wave of first-year students one year later and reproduce the results. Survey data provides suggestive evidence that positive feedback has an effect on behavior when students underestimate their relative performance, and that consistent with a mechanism of selective information processing, individuals focus on positive feedback to adjust their beliefs.

Keywords: Relative Performance Feedback, Higher Education, Randomized Field Experiment, Replication, Selective Information Processing

JEL Classification: I23, C93

---

[a]University of Göttingen, Chair of Public Economics, Platz der Göttinger Sieben 3, 37073 Göttingen, Germany. email: raphael.brade@uni-goettingen.de.

[b]Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Strasse 10, 53113 Bonn, Germany, University of Erfurt, and University of Göttingen. email: ohimmle@gwdg.de.

[c]TH Nürnberg Georg Simon Ohm, Faculty of Business Administration, Bahnhofstrasse 87, 90402 Nürnberg, Germany. email: robert.jaeckle@th-nuernberg.de.

We would like to thank Robert Schwager, Philipp Weinschenk, and seminar participants at the ESA World Meeting, ifo Center for the Economics of Education, Imebess at the European University Institute Florence, Spring Meeting of Young Economists, the University of Augsburg, the University of Göttingen, and the University of Nürnberg for valuable comments and discussions. Fabian Escher, Peter Frenzel, Burak Akkaymak, Milad Zargartalebi, and Nadine Gerlich provided excellent research assistance. We gratefully acknowledge financial support from the German Federal Ministry of Education and Research under grant 01PX16003A, 01PX16003B, and administrative and financial support from the TH Nürnberg and the Max Planck Institute for Research on Collective Goods. Jäckle also gratefully acknowledges financial support from the Staedtler Stiftung.

1 Introduction

Whenever individuals enter new environments, such as university or a different job, feedback can help them appraise their performance. Absolute feedback, however, is often insufficient to put performance into perspective, because objective benchmarks have not yet been established or are unknown to the individuals, leaving them with no appropriate frame of reference. The resulting uncertainty may compromise motivation, or lead to misguided and non-optimal decision-making, for example when it comes to effort provision. Under such conditions, social comparison theory suggests that information about the behavior of similar others can provide an important benchmark or reference point against which individuals can compare their performance and gauge their abilities (Bandura 1991, Corcoran, Crusius and Mussweiler 2011, Festinger 1954, and Taylor, Wayment and Carrillo 1996). Offering performance feedback relative to a suitable peer group may therefore enhance decision-making and motivation, and therefore facilitate an efficient transition into the new environment.

The performance feedback literature largely relies on providing relative feedback in terms of objective information like rank or percentile in the performance distribution.[1] A complementary approach is frequently employed in the social norms literature, but has not been explored when it comes to performance feedback:[2] the provision of normative feedback.[3] Normative frames or injunctive messages can provide cues which make individuals aware of reference points and especially of what type of behavior is approved in a new environment (see e.g., Allcott 2011, Cialdini, Reno and Kallgren 1990, Deutsch and Gerard 1955, and Schultz, Nolan, Cialdini, Goldstein and Griskevicius 2007). We combine objective relative performance information with normative frames to clearly convey what constitutes positive and negative feedback, respectively. There is some theoretical reason to believe that positive and negative feedback have diverging effects on performance, by differentially affecting motivation or because negative information may be discounted (see Bénabou and Tirole 2016 or Golman, Hagmann and Loewenstein 2017). Augmenting objective relative performance feedback with normative messages allows us to separately investigate the effects of these

[1]Examples for the provision of rankings or distributional information range from workplace settings (e.g., Azmat and Iriberri 2016, Blanes i Vidal and Nossol 2011, Card, Mas, Moretti and Saez 2012, Gibbons and Murphy 1990, or Gill, Kissová, Lee and Prowse 2018) to educational environments (e.g., Azmat and Iriberri 2010, Azmat, Bagues, Cabrales and Iriberri 2018, and Goulas and Megalokonomou 2015) and have shown mixed effects on performance and other outcomes.

[2]Where we take performance to broadly encompass accomplishments at the workplace, in school or at other productive tasks.

[3]Normative frames are popular in social comparison contexts that are not concerned with performance. They have been applied to foster e.g. tax compliance (Hallsworth, List, Metcalfe and Vlaev 2017, Slemrod 2016) or environmentally-friendly behavior (see e.g., Allcott and Rogers 2014, Costa and Kahn 2013, and Goldstein, Cialdini and Griskevicius 2008).

different types of feedback.

Feedback is presumably especially important when the new tasks are complex and challenging, and the stakes are high. An economically and socially important area to which this applies is higher education. Individuals at the start of their academic career may lack information on how difficult it will be to succeed in complex tasks like passing university exams and how much effort they should expend.[4] Additionally, the stakes are relatively large. Failing exams always comes with a substantial cost in terms of re-taking classes and psychological pressures, but at the beginning of study programs this is often exacerbated by certain core exams being prerequisites to continue on with the program. This makes higher education an interesting and relevant setting to test the effectiveness of normatively framed performance feedback.

We conduct a randomized field experiment with a cohort of first-year students at a German university of applied sciences, and combine relative performance feedback on obtained course credits with normative cues. After the first semester exams, control group students receive letters in the mail informing them about how many credits they obtained in the previous semester. Students in the treatment group receive the same information but the letters also inform them about how well they performed relative to the average student and the student on the 80th percentile.[5] This information is augmented with normative messages of approval for those who obtained at least the average amount of credits. For those below the average the approving normative messages are greyed out. These individuals therefore receive no explicitly disapproving normative framing, but should be aware that their performance failed to qualify for an approving message. This design allows us to provide first evidence on the effects of performance feedback by type of feedback. We label the different types of feedback as follows: *positive* (= above average performance + approving normative message), *ambiguous* (= average performance + approving normative message) and *negative* (= below average performance + no approving normative message).

We provide feedback on obtained course credits. The choice of credits over grades as the feedback dimension is driven by two important considerations. First, for students at the start of their university careers it is likely harder to assess their abilities with respect to how many exams they can take than to assess their ability to obtain good grades (Figure 16 provides

[4]Among other causes, this may contribute to a large share of students in higher education never obtaining a degree or taking much longer than scheduled to graduate. In the U.S., less than 40% of a cohort entering four-year institutions obtain a bachelor's degree within four years (See the National Center for Education Statistics (NCES), Jan 19 2017, retrieved from http://nces.ed.gov/programs/digest/ d13/tables/dt13_326.10.asp). The U.S. is no exception. In many countries delayed graduation is prevalent, and about 30% of students entering a four-year tertiary education in the OECD do not complete their studies (OECD 2013).

[5]In the first experiment "average student" refers to the median student, in the second experiment to the mean performance of the other students.

some evidence on this). This is because in contrast to university, high school typically leaves little freedom to vary the amount of exams taken and so students have no experience in this domain. Such situations where individuals lack experience and therefore information are exactly where feedback can be most helpful. Second, obtained credits are an interesting outcome from a policy perspective, as they are not only a measure for academic performance, but ultimately determine time to graduation. Lower time to graduation has substantial payoffs: on an individual level people forgo income with each semester taken longer to graduate, and from a social perspective faster graduation means higher contributions in terms of taxes and payments into the social security systems.[6] It is thus not surprising that time to graduation has started to receive more attention in recent years, possibly triggered by the observation that in many countries students take much longer than scheduled to obtain a degree (e.g., Bound, Lovenheim and Turner 2012, Garibaldi, Giavazzi, Ichino and Rettore 2012, Himmler, Jäckle and Weinschenk (forthcoming), and Leuven, Oosterbeek and Klaauw 2010).

Another important feature of our experiment is that we observe the entire universe of performance, i.e. the results of all exams taken. We thus measure the treatment effect of feedback on overall exam performance and not in a single course (in which case one may find effects simply due to students re-allocating effort between courses). Any positive effects can thus be interpreted as net gains. We give feedback on performance in terms of passing exams (obtained course credits), and expect to find effects in that domain. Yet an important but often neglected question is whether such gains can be had without sacrificing on another dimension. We thus also monitor student grades and dropout behavior, and we consider survey responses on measures of well-being such as life satisfaction and satisfaction with the study program. This allows us to paint a more encompassing picture of the effects of performance feedback on participants' welfare.

We find that in the second semester (i.e. the first treatment semester), students who receive positive feedback earn about 2.1 credit points (.2 standard deviations) more than the controls. With a regression discontinuity design we make use of the sharp cut-off for positive feedback, and show that the treatment effect is indeed due to the positive feedback, and not due to unobserved characteristics of those receiving positive feedback. Negative feedback has no statistically significant effect on behavior. Neither does ambiguous feedback. The latter suggests that sending a normative message of approval has no effect without the fundamentals to back it up, i.e. without evidence of above average performance. Finally, we show that students who receive positive feedback maintain the same grade point aver-

[6] In many ways, the benefits of obtaining credits are therefore quite straightforward and possibly easier to quantify than the benefits of obtaining better grades.

age as the controls, indicating that they do not buy gains on one performance dimension with losses on another. Similarly, we find no evidence of students being worse off on the well-being domains we observe.

We administer a replication experiment one year later with the new cohort of first-year students. The design is identical with one small tweak. Because we found effects of positive feedback, we now report the mean instead of the median for the average performance in the feedback letters. Since the performance of many students in the first experiment was exactly equal to the median and because the mean is smaller than the median in our data, this enables us to give positive feedback to a larger share of the treatment group (56% vs 37.5% in the original experiment). We reproduce all findings of the original experiment. Most importantly, we validate the result that positive feedback significantly increases performance (although the effects are somewhat smaller: 1.7 credit points or .16 standard deviations). This is particularly interesting, because it implies that through the simple design tweak, we can extend the positive effects to also benefit weaker students (who obtain median or slightly below median credits).

In both the original experiment and the replication we find no further performance gain of the treatment groups versus the controls in the third semester. In order to better understand these dynamics and the differential effects of feedback types, in the replication experiment we gather survey data on the pre- and post-treatment expectations about relative performance. In particular, we investigate two questions: first, do effects depend on whether the feedback provides new or unexpected information, and, second, are the results driven by the different feedback types being asymmetrically processed? We find suggestive evidence that positive feedback is effective when students underestimate their relative performance pre-treatment. Further, a post-treatment survey in the second semester shows that students update their expectations in response to positive relative feedback, and that the expectations of these students are significantly more accurate than the expectations of the controls qualifying for positive feedback. At the same time, we see no evidence of updated expectations in response to negative feedback, suggesting that students possibly ignore or disregard such information (although the sample size here is very small and the confidence intervals wide). Finally, in the third semester, the expectations of the controls who would have qualified for positive feedback are almost as accurate as the expectations in the treatment group. This indicates that after a while there is no longer an informational difference between treatment and control group, in line with no further effects of positive feedback in the third semester. In contrast, with negative feedback, neither control nor treatment group improve the accuracy of their expectations in the third semester – which again suggests that students do not process negative information.

The results are broadly consistent with the idea that a positive view of one's ability is a significant motivational and potentially performance enhancing factor (see Bénabou and Tirole 2002). Individuals may therefore maintain a positive self-assessment by selectively processing positive feedback while discarding negative feedback.[7] This idea is also expressed in Eil and Rao (2011): individuals who receive negative feedback have little willingness do update their self-concept, whereas people who receive positive information are willing to incorporate the positive news in their future behavior. The finding that positive feedback increases performance is also in line with the models of Ertac (2005) as well as Azmat and Iriberri (2010), who argue that individuals use feedback to update self-perceived ability. Similarly, in the presence of selective information processing, competitive preferences and social norms theory can be aligned with our findings.

Overall, our results show that identifying information deficiencies and providing normatively framed feedback can help individuals who underestimate their abilities. This strategy is not only effective but also inexpensive. The total cost of our measure per semester and student is less than €2.5 (see Table 3). In education contexts, this compares very favorably to more traditional measures of improving outcomes, such as hiring new faculty in order to reduce class size or grant schemes.[8]

Relation to the Literature Our findings contribute to the literature on several dimensions. First, we add to the relative performance feedback literature in general. Most of the research in this area is concerned with effects on the performance of employees, and is implemented in the lab or workplace contexts. In lab experiments the subjects are typically asked to complete real-effort tasks such as decoding, adding, or multiplying numbers and are given feedback on their performance (Azmat and Iriberri 2016, Charness, Masclet and Villeval 2013, Eriksson, Poulsen and Villeval 2009, and Kuhnen and Tymula 2012). When it comes to field evidence, the tasks include dispatching goods, picking fruits, and selling clothing or furniture (Bandiera, Barankay and Rasul 2013, Barankay 2012, Blanes i Vidal and Nossol 2011, and Delfgaauw, Dur, Sol and Verbeke 2013). The tasks in those studies usually are repeatedly performed over a longer period of time, and the relative performance feedback is introduced

[7] A strand of the psychological literature has argued that individuals increase their efforts only after receiving positive feedback, and underweight adverse information about themselves (Ilgen, Fisher and Taylor 1979, Ilgen and Davis 2000, and Pearce and Porter 1986). In economics the reasoning is similar: Bénabou and Tirole (2002) argue that individuals choose to selectively remember positive information in order to maintain a positive self-image, whereas in Compte and Postlewaite (2004) individuals put little weight on negative information because they attribute negative outcomes to e.g. bad luck, as opposed to positive outcomes, which are attributed to own abilities or efforts.

[8] For example, the US Pell Grant initiative pays individual students up to $5,815 per annum (2016-17 award year). It aims at keeping students on track by rewarding "accelerated completion" and paying an "On-Track Pell Bonus"; see US Department of Education, `http://www.ed.gov/news/press-releases/fact-sheet-helping-more-americans-complete-college-new-proposals-success`, retrieved on Jan 19 2017.

at a time when workers are familiar with the task (Barankay 2012, Blanes i Vidal and Nossol 2011, or Delfgaauw, Dur, Sol and Verbeke 2013). We complement this literature by investigating the effects of relative performance feedback in a new environment, where the stakes are relatively high, the task is rather complex, and individuals have little intuition about the underlying performance distribution. Furthermore, the performance feedback literature has focused on providing feedback in the form of rankings or distributional information. Our approach of combining this type of information with normative messages is novel in the performance feedback literature. It clearly communicates to the subjects what constitutes (un-)desirable performance and what therefore should be perceived as positive and negative feedback – allowing us to separately analyze the effects of these different feedback types.

More specifically, our results contribute to the scarce literature about relative performance feedback in higher education, which so far provides mixed results.[9] There are two studies which find that relative feedback on the performance in a practice test or midterm exam improves the outcome of the final exam (Kajitani, Morimoto and Suzuki 2017 and Tran and Zeckhauser 2012). However, two other studies that provide students with information about their rank in the GPA distribution find zero or negative effects (Azmat, Bagues, Cabrales and Iriberri 2018 and Cabrera and Cid 2017). Our study adds a number of aspects to this literature. First, we provide evidence that feedback type (positive or negative) matters and that positive relative performance feedback increases subsequent performance. Replicating the original experiment with a later cohort of students shows that this is a robust finding. Second, we are the first to provide relative performance feedback on obtained credits instead of GPA; a domain where it is likely hard for students to assess their own ability (as they just entered university), and where it is also likely that they have only little prior knowledge about the performance distribution of others students. Third, by also considering effects on grades we can make sure that the positive effects are not accompanied by students trading off credits and grades. Finally, by observing the entire universe of performance we can check that any effects are indeed net gains, and are not generated by sacrificing performance in other courses (which could be a concern when only observing outcomes in one course).

Furthermore, the results of our study are related to the literature that studies the relation between confidence and performance. Bénabou and Tirole (2002) develop a model in which higher confidence in their abilities motivates individuals to work harder and take beneficial risks. Consequently, individuals may engage in confidence enhancing and/or preserving behavior, e.g. in asymmetric updating of beliefs in the face of good or bad news. Both aspects have been studied empirically. Individuals with higher levels of confidence have

[9]There is also a literature on the effects of relative performance feedback in primary and secondary education that mainly finds positive effects (Azmat and Iriberri 2010, Fischer and Wagner 2017, Goulas and Megalokonomou 2015, and Katreniakova 2014).

been found to work harder (Puri and Robinson 2007), even in tasks that are unrelated to the source of confidence (Pikulina, Luc and Philippe 2018). Also, individuals do indeed engage in asymmetric updating, depending on whether news are good or bad (Eil and Rao 2011, and Möbius, Niederle, Niehaus and Rosenblat 2014). Our findings are consistent with both aspects. First, we see an increase in performance in response to positive feedback. Second, we observe that individuals correctly revise their beliefs about their expected relative performance in the face of positive feedback, while there is no evidence they do so when presented with negative feedback. Our field experiments therefore add to the existing literature on the asymmetric updating of beliefs, especially the very scarce evidence from the field. Taken together, the results are in line with the notion that enhanced confidence constitutes the link between positive feedback and increases in performance.

Finally, we contribute to the literature that applies informational nudges in public policy, where the provision of information has been shown to e.g. influence medical choices, advance environmentally friendly behavior, tax compliance, and social benefit take-up (for an overview see e.g., Madrian 2014 or Chetty 2015). In particular, our intervention adds to the strand of research which aims to improve outcomes in higher education by providing information.[10] For example, Hoxby and Turner (2013) find that high-achieving, low-income students attend more selective colleges, when they are given information about the application process and the net cost of colleges. Castleman and Page (2015, 2016) use personalized text messages to increase the number of individuals that enrol in college among low-income high school graduates, and to increase college persistence, respectively. Wiswall and Zafar (2015a, 2015b) find that students revise their beliefs about expected earnings in response to information about the true population distribution of earnings. The revised beliefs then influence the college major choice of students. Our study shows that the provision of (normatively framed) information can be an effective means of targeting performance directly. This complements studies that e.g. try to directly influence student performance via financial incentives (e.g., De Paola, Scoppa and Nisticò 2012 and Leuven, Oosterbeek and Klaauw 2010), by providing a low cost and easily scalable alternative.

The remainder of the paper is structured as follows. Section 2 describes the institutional background, data, and design of our two experiments. We discuss different theoretical frameworks in Section 3. The empirical analysis is presented in Section 4. Section 5 concludes.

[10] Damgaard and Nielsen (2018) review the current research on (informational) nudging in education.

2 Institutional Background and Research Design

Our field experiments take place at one of the largest universities of applied sciences in Germany, with twelve faculties and more than 13,000 students. All students in our experiments are enrolled in bachelor's degree programs at the largest and third-largest faculties of the university: Business Administration and Mechanical Engineering.

2.1 Institutional Background

In order to obtain their degree, students need to acquire a total of 210 credits (ECTS), and the scheduled study duration is seven semesters, i.e. on average they are supposed to obtain 30 credits per semester.[11] The study programs follow a modular structure, where credits are awarded for passed exams. For many, these exams are challenging and complex tasks, especially at the beginning of an academic career. Arguably, they also have a high stakes nature for several reasons. First, passing obviously determines whether or not the student obtains the credits needed to graduate. Second, passing more exams per semester reduces the time to graduation, which typically translates into earlier entry into the workforce and higher lifetime income.[12] Third, students have to pass certain exams early in their studies in order to be allowed to continue with the program.[13] Fourth, the cost of not passing can be substantial, both in terms of time as well as psychologically: Students will have to study for the exams again, and potentially have to attend the same classes again in the next semester. On top, failing exams may incur psychological distress.[14]

Students can at all times access information on their progress via a web portal maintained by the university. The portal provides data on absolute performance – credits and GPA. It is important to note, however, that in absence of our treatment, the university does not provide any information or feedback on a student's relative performance.

2.2 Field Experiment I

The original field experiment is conducted with a cohort of first-year students who just started in their bachelor's programs. Treatment commences as soon as information on previous performance is available, i.e. at the start of the second semester. At this point 812 students study

[11]The European Credit Transfer and Accumulation System (ECTS) is the Europe-wide standardized point system which allows students to transfer their credit points from one European university to another.

[12]However, since these benefits materialize far in the future, they may not be very salient to students.

[13]For example, the programs typically include an internship semester and students are not allowed to start the practical part of their education before they have earned at least 60 credit points.

[14]See, e.g. Brunstein and Gollwitzer (1996) for a discussion of the (psychological) effects of failure.

towards degrees in five bachelor programs at the faculties for Business Administration and Mechanical Engineering. Roughly 80% of these students are enrolled in two large programs: Business Administration and Mechanical Engineering (Table 2 provides an overview of all degree programs and the number of students in our intervention).

Randomization. Randomization was carried out after the first semester, using stratification and balancing (see Morgan and Rubin 2012). We applied information on study programs and obtained accumulated first-semester credits (ACP) to divide students into separate strata and balanced on age, sex, high school grade, time since high school graduation, pre-treatment grade point average and type of high school degree (only Experiment II).[15] The rightmost columns of Table 2 display the fraction of students in treatment and control, by degree program. Tables 4 and 5 show the balancing properties across the control and treatment groups for the full sample and the subsample of students that were eligible to receive positive feedback.

Feedback Letter I. In the week before the second semester lectures start (see Figure 1), students in the control and treatment group receive an unannounced letter in the mail, providing them with information on their accumulated credits (ACP) and their cumulative GPA (AGPA). The envelope bears the official seal of the university, in order to ensure that students will open and read the letter. To further stress the official character, the letters are signed by the dean of the respective faculty. In the few cases where letters came back as undeliverable, we re-sent them via email (N= 32).

In both groups the letter states that the faculty "[...] would like to assist you in the further organization and planning of your studies. To this end we provide you with feedback information about your current academic performance" (see Figures 3 and 4). The control group letter then lists the student's credits and GPA obtained in the first semester – with no comparison to fellow students. In contrast, the treatment group letter continues with a graphical illustration that provides relative performance feedback on accumulated credit points. The visualization of this social comparison feedback information is shown in Figures 2 and 3, and we explain the design in detail below. Finally, as in the control group, the letter also quotes the student's cumulative first semester GPA (with no relative comparison).

Feedback Letter II. About four to five weeks before the exam period, students of the control and treatment group receive a second feedback letter (see Figure 1 and Figures 5 and 6). The letter design is identical to the first letter, and for most students the contained information

[15] In Experiment I rerandomization was only conducted in the largest Bachelor's program (BA). See the Data and Methods Appendix for more details on our randomization procedure.

will also be identical to the first letter. In some cases the university updated the information on grades and credits (e.g., because course results were not yet available), which can lead to different feedback compared to the first letter.[16] Apart from providing the most accurate information, the purpose of the second letter is to keep the feedback information salient as the exam period draws nearer. Consequently, we base our estimates of the effects of different feedback types on the content of the second letter.

Social Comparison. The visualization of the relative performance information in the treatment group can be seen in Figure 2.[17] A bar chart compares the individual student's obtained credits to the "Top 20%" and to the "Average" student enrolled in the same bachelor's program and the same semester as the student receiving the letter. "Average" performance is defined as the number of credits obtained by the median student(s), and "Top 20%" refers to the performance of the student(s) on the 80th percentile.

Social comparison theory suggests that perceived similarity increases the tendency to engage in social comparisons (Festinger 1954). In order to increase perceived similarity with and to minimize 'psychological distance' to the reference group (compare Trope and Liberman 2010) we further personalize the relative performance information. We define several comparison groups within each degree program: In smaller programs the comparison group consists of students "who in/before ⟨*year*⟩ earned their school leaving certificate.", where *year* equals the year in which the addressee of the letter received their school leaving certificate (this is in the International Business, Business Engineering, and the Energy and Building Services Engineering programs; see Table 2). In the large bachelor programs Business Administration and Mechanical Engineering, we further decompose the comparison groups by additionally referring to "students who in/before ⟨*year*⟩ earned *the same kind* of school leaving certificate *as you*", and we distinguish between the certificates "vocational track degree (or below)" and "general track degree".

Finally, an important feature of the feedback is that it provides a normative framing of the student's relative performance. The framing conveys that performing at least at the average level is approved of. It categorizes the students' performance as "good" (plus one "smiley" emoticon) for students at or above the average, and "great" (plus two "smiley" emoticons) for students in the top 20%. Students below average receive a neutral statement "currently below average" (and no emoticon), and the approving normative messages are greyed out.

We define as "positive feedback" a situation where the student both receives an approv-

[16] See the Data and Methods Appendix for details on the reasons and the number of observations that are affected.

[17] The design is inspired by the Social Comparison Module of OPOWER's Home Energy Reports (Allcott 2011, Allcott and Rogers 2014, and Schultz, Nolan, Cialdini, Goldstein and Griskevicius 2007).

ing normative framing ("great" or "good") and the information that they have obtained an above average number of credits, i.e. they are also given fundamentals that match the approving normative framing – this is true for all students above the average. Students on the average receive "ambiguous feedback": an approving normative statement that is not backed by matching fundamentals, because the student receives information that the obtained number of credits is only average.[18] Finally, we label the feedback that students below the average receive as "negative" since they do not see an approving normative framing, and also receive information that their performance was below average.

2.3 Field Experiment II: Replication

We replicate the experiment one year later (N= 797, Table 2 provides an overview of all degree programs and the number of students in the replication; Tables 4 and 5 show the balancing properties[19]). The aim is to establish with a new cohort of students whether the results are reproducible. Finding credible evidence that results can be reproduced is an important goal of our study, especially given the inconclusive results in the literature and the recent debate about replicability in economics and other fields (Camerer, Dreber, Forsell, Ho, Huber, Johannesson, Kirchler, Almenberg, Altmejd, Chan et al. 2016, Duvendack, Palmer-Jones and Reed 2017, and Open Science Collaboration 2015). The replication experiment uses the same design as the original experiment. The only small tweak is that as a reaction to finding performance enhancing effects of positive feedback in Experiment I (see Section 4.3), we now use the mean instead of the median as the cutoff above which students receive positive feedback. This retains all features of the original experiment while at the same time allowing us to provide 56% of the students instead of 37.5% with positive feedback and reducing the number of students that receive ambiguous feedback from 165 to 29 (because many students obtained exactly the median amount of credits in the first semester and because the mean is smaller than the median for observed credits in the first semester).

3 Theoretical Considerations

This section provides some theoretical intuition on how our intervention may affect behavior. As described above, we provide individuals with normatively framed information about

[18]In one of the smaller study programs in Experiment I, the median coincides with the 80th percentile. 15 students therefore receive feedback indicating that they are in the top 20%, but at the same time their performance is average. Because they do not meet all our criteria necessary for the label "positive feedback" they are in the "ambiguous feedback" category. Our results are robust to classifying them into the "positive feedback" category.

[19]We adjusted the randomization procedure in Experiment II. See the Data Appendix for details.

their performance relative to their peers. Social comparison theory suggests that such information will set social reference points (e.g., Festinger 1954, Suls and Wheeler 2000, or Wood 1989), and there are several theoretical reasons why individuals may change their behavior in response to reference points established by the feedback.

First, feedback may affect self-confidence and have a positive effect on motivation, thus increasing willpower and perseverance – which can ultimately lead to better performance. In Bénabou and Tirole (2002), positive news serves individuals to maintain a positive self-image, which in turn motivates them. Negative news on the other hand does not adequately enter into beliefs: individuals selectively process good information. In Compte and Postlewaite (2004) the mechanism is similar (but not a deliberate choice of the individual): positive outcomes are attributed to own abilities or efforts. Negative outcomes are attributed to e.g. unfortunate circumstances and therefore do not appropriately depress self-confidence. In both models, the induced optimism and confidence in own abilities can then lead to better performance. Applied to our setting, a lower perceived probability of failure in an exam may lead to higher effort levels and subsequent performance. Beyond this mechanism, confidence may also have a direct effect on utility, i.e. individuals may simply enjoy feeling good about themselves (Bénabou and Tirole 2002, Compte and Postlewaite 2004, and Köszegi 2006). This may also be a motivational factor for effort allocation.

The normative framing of feedback with approving messages (or their absence) gives the recipient of the feedback clear indication of when performance should be considered good and therefore is suited to bolster confidence.[20] Applied to our setting, for students who receive good news (positive feedback) the treatment may cause an increase in self-confidence, which helps subsequent performance. On the other hand, the postulated selective processing of information suggests that we may not see adverse effects of bad news, i.e. of negative feedback (this potential dichotomy in information processing is closely related to the literature on the disregarding and discounting of negative information, which we also discuss later in this section).

The idea in Bénabou and Tirole (2002), where confidence is manifested as beliefs over ability, is closely related to the models in Azmat and Iriberri (2010) and Ertac (2005). Here, individuals have only incomplete knowledge of their own ability.[21] Formally, they only know that their ability is drawn from a normal distribution with known parameters. All individuals

[20]The approving normative frame should e.g. make it more likely that a very ambitious individual will take performance between the average and the 80th percentile as good news rather than negative information.

[21]Azmat, Bagues, Cabrales and Iriberri (2018) develop a further theoretical framework to explain how information can influence performance. They focus on a situation where individuals already have good knowledge of their own ability in terms of GPA. Our setting is different, as we provide feedback with regard to credit points and show that individuals have little prior knowledge about their own ability in terms of credit points (see Section 4.6).

then receive a private signal about their performance, which can be used to update beliefs about ability. In our setting one could think of the individually passed exams and obtained credits as this private signal. When relative feedback is given, this signal about the average performance of the other students will additionally affect self-perceived ability by providing information about how hard the task was. It can be shown that if the task was of average difficulty[22] and under the assumption that ability and effort are complements in generating performance, students will invest more effort if they receive information that they have performed above average, and less effort if they receive information that they have performed below average. In our setting this translates to higher (lower) performance after receiving positive (negative) feedback. Combined with the idea of self-confidence managing individuals in Bénabou and Tirole (2002), we may however only see an effect of positive feedback, as the negative feedback may be discarded.

Second, relative performance information can influence the behavior of individuals with competitive preferences. One way to specify these is to include an additive social comparison component in the utility function that penalizes individuals for performing below the expected average and rewards them for performing above the expected average (see, e.g. Azmat and Iriberri 2010 or Kandel and Lazear 1992). Following Azmat and Iriberri (2010), relative performance information will increase the precision of the expectation about average performance, and individuals will put more weight on the competitive part of the utility function. The weight should be especially high when the competitive situation is stressed and made salient, as may be the case when normative messages are included. As a consequence, we expect treated students to increase performance relative to the control group. Note that this is the case irrespective of a student's position relative to the reference point.

Third, the reference points in the feedback letter may convey a descriptive social norm, by describing how others behave. The focus theory of normative conduct suggests that individuals try to comply with descriptive norms, predicting positive treatment effects for those below the descriptive norm (Cialdini, Reno and Kallgren 1990 or Cialdini 2011). However, at the same time negative effects for individuals who perform above the norm are implied (labeled a "boomerang" effect in the social norms literature).

The solution suggested in the literature is to add an approving normative message, i.e. an injunctive norm for those who exceed the descriptive norm. This can prevent boomerang effects and sometimes generates additional positive treatment effects (Allcott 2011, Cialdini 2003, Hallsworth, List, Metcalfe and Vlaev 2017, and Schultz, Nolan, Cialdini, Goldstein and Griskevicius 2007). Our normatively framed relative performance feedback can be under-

[22]This seems reasonable, because we give relative feedback on the performance in the entire semester and students usually write multiple exams per semester. The assumption is then that while each particular exam may be more or less difficult, in sum these exams are of average difficulty.

stood as such a combination of descriptive and injunctive norms. An important caveat is that this may fail to produce beneficial effects if the two norms are not in alignment (Cialdini, Demaine, Sagarin, Barrett, Rhoads and Winter 2006). In our case this could mean that the normative message only works if it is aligned with the information of the relative performance feedback, e.g. if the approving normative message is supported by information about an above average performance. Ambiguous feedback could then possibly result in zero effects because the descriptive norm and the normative message are not aligned. Negative feedback should increase performance, and students who receive positive feedback should keep their behavior unchanged or exert more effort.[23]

All three mechanisms predict that students who receive positive feedback will increase their subsequent performance, or leave it unchanged. The prediction for negative feedback is less clear. While descriptive norms and competitive preferences suggest an increase in performance, a downward adjustment of perceived ability should lower performance when effort and ability are complements. This ambiguity is further complicated by the above described potentially selective processing of information when self-confidence matters, and information avoidance in general. Individuals may choose not to receive information at all, or discount negative information (see e.g., Eil and Rao 2011, Grossman and Owens 2012, and Möbius, Niederle, Niehaus and Rosenblat 2014). Potential reasons for this behavior in our context could be disappointment aversion, anxiety, and optimism maintenance (Golman, Hagmann and Loewenstein 2017).[24] Hence, it could be the case that students who receive negative feedback will discount or disregard it. As a consequence, the relative performance information would not set a social reference point for those students, and we would thus expect that they do not change their subsequent performance.

It is important to note that the experiments were not designed with the goal of testing the different mechanisms against each other. In the empirical analysis, we will therefore not be able to pin down what mechanism exactly is driving the results. However, this section has shown that the notion of positive feedback being (weakly) beneficial to performance is common to all of the discussed theories. Consistent with this overarching sentiment, our results will provide robust evidence that positive feedback does indeed increase performance.

[23]Note that in our case the average performance and the performance of the 80th percentile may both constitute descriptive norms. Students below average (negative feedback) are below both reference points, so no matter which norm they pick, the theory suggests positive effects of feedback. Similarly, students in the top 20% are above both reference points and the approving framing should prevent boomerang effects and possibly even generate performance gains. Students above average and below the 80th percentile may choose a reference point above or below. The predictions differ depending on this choice: positive effects if they choose the higher reference point, and a possible boomerang effect if they choose the lower reference point. However, since they also receive a normative message of approval, we expect zero or positive effects in either case.

[24]The idea of information avoidance is also supported by psychologists. For surveys of the literature see e.g., Hertwig and Engel (2016) or Sweeny, Melnyk, Miller and Shepperd (2010).

In our field experiments we find no evidence of negative feedback significantly affecting behavior in any direction. This may be because multiple mechanisms work in different directions, or because individuals discount negative information. We will provide some tentative evidence pointing to the latter.

4 Results

In this section we first present our data, specifications, and the main findings. We then investigate drivers of the effect and potential mechanisms, as well as the effects of repeated treatment.

4.1 Data

We use anonymized administrative student-level data provided by the university. As described in Section 2.1, students received feedback based on the accumulated credits (ACP) and their cumulative GPA (AGPA) and we used those variables together with demographic information for the randomization. The administrative data is augmented by four online surveys.

For the empirical analysis, we use exam-level data instead of the cumulative figures. This provides more accurate information on student performance in each semester and as explained in the Data Appendix it is also the more conservative approach.[25] Our main outcome of interest is the number of obtained credits.[26] In auxiliary analyses, we will also investigate potential effects on other domains such as the number of attempted and failed exams, GPA, dropout, and well-being. Failed exams are conditional on attempting at least one exam, the GPA is based on passing grades, and we elicited the measures of well-being with online surveys (see Figure 1 and Table 21 for the timing and response rates of the surveys and Table 18 for the questions and the variables used in the estimations). As shown in the next section, we use demographic information and baseline outcomes (first semester credits and first semester GPA) as covariates in our estimations (see Table 1).

[25] The Data and Methods Appendix provides an example, and a detailed discussion of the differences between the accumulated and the exam-level figures.

[26] Credit points are net of credits granted for internships, which are scheduled later in the study program (4th/5th semester). Some students are awarded these credits at the start of their studies because they completed an apprenticeship and have work experience. As we are interested in the effect of treatment on academic performance, we do not count these internship credits.

4.2 Estimation

Unless otherwise specified, we provide intention-to-treat (ITT) effects from OLS estimations that compare the average outcomes of the control and the treatment group.

The baseline specification is:

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \varepsilon_i, \tag{1}$$

where Y_i^k denotes the level of outcome measure k for individual i. $Treatment_i$ is an indicator for being randomized into the treatment group.

In the second specification we follow the recommendations of Bruhn and McKenzie (2009) and control for the method of randomization:

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \mathbf{s_i}\boldsymbol{\alpha_2} + \varepsilon_i, \tag{2}$$

The vector $\mathbf{s_i}$ includes strata fixed effects which control for the random assignment of treatment and control units within blocks. Strata are defined by study program dummies and accumulated credits. In the estimation with pooled data from both experiments we also include cohort fixed effects and their interactions with the other strata variables.

For the third specification, we add a vector $\mathbf{x_i}$ which includes covariates that capture student ability:

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \mathbf{s_i}\boldsymbol{\alpha_2} + \mathbf{x_i}\boldsymbol{\alpha_3} + \varepsilon_i. \tag{3}$$

The vector contains the high school GPA and the first semester credits (description of variables in Table 1).

In the fourth specification, we add a vector $\mathbf{z_i}$ of additional control variables:

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \mathbf{s_i}\boldsymbol{\alpha_2} + \mathbf{x_i}\boldsymbol{\alpha_3} + \mathbf{z_i}\boldsymbol{\alpha_4} + \varepsilon_i. \tag{4}$$

It includes the age at randomization, an indicator for being female, the time since high school graduation, and an indicator for the type of high school degree.

Lastly, we estimate a fifth specification in which we control for the baseline GPA (GPA first semester):

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \mathbf{s_i}\boldsymbol{\alpha_2} + \mathbf{x_i}\boldsymbol{\alpha_3} + \mathbf{z_i}\boldsymbol{\alpha_4} + \alpha_5 BaselineGPA_i + \varepsilon_i. \tag{5}$$

We do this in a separate specification because the first semester GPA is missing for students who attempted no exams or failed all exams they attempted. In order to keep all observations in the sample, we impute values of the first semester GPA for students with a missing GPA.[27]

4.3 Effects of feedback types on performance

Field Experiment I. The top panel of Table 6 shows that across all types of relative performance feedback, the treatment group obtains on average roughly .6 additional credits (Column 1). This effect is not statistically significant, and if anything, adding the control variables in Columns (2) to (5) reduces the estimated effect.

Based on the theoretical considerations in Section 3, we are interested in whether different feedback types lead to different behavioral responses. As explained in Section 2, students below the average received negative feedback, while students above the average received positive feedback. Column (7) shows that students who received positive feedback perform significantly better than the control group counterparts (who would have also received positive feedback had they been allocated to the treatment group): they obtain on average 2.4 credits more. Adding covariates in Columns (8) to (11) reduces the effect to about 2.1 credits, which corresponds to an effect size of roughly .2 control group standard deviations.

Columns (5) to (8) in Table 7 show that among the students receiving positive feedback, those who range at or above the 80th percentile do not react as strongly to the treatment as those students between the 50th and 80th percentile. How positive feedback affects the distribution of the outcome variable can be seen in the top panel of Figure 8, which shows the density of standardized credits for individuals receiving positive feedback in the treatment group and individuals eligible for positive feedback in the control group. Credits are standardized within study programs in order to make the distribution of the outcome variable comparable between the different study programs.[28] The distribution is left-skewed in the treatment and the control group, but more so in the treatment group. The reason is that positive feedback predominantly moves up students from the bottom part of the distribution. At the same time there appears to be no treatment effect for the top part of the distribution. We analyze this more formally in the top panel of Figure 9, which depicts the coefficients of simultaneous quantile regressions (QR) for the 10th to the 90th percentile of the credits (using the set of controls from Equation 5). The OLS estimate is also depicted for comparison, and the figure shows a clear pattern: While the QR estimates are roughly the same as the OLS estimate for the first five deciles, the QR estimates decrease for the upper deciles. This shows that there are indeed no effects of positive feedback for the individuals at the top of the credit point distribution.

[27] See the Data and Methods Appendix for details.

[28] In the regressions the credit points are not standardized. Instead, we account for the different levels of credit points across programs by including study program fixed effects.

The heterogeneous effects of positive feedback across the distribution of the outcome variable could arise from ceiling effects if students are only capable of passing a certain number of exams per semester. The curriculum may also generate “artificial” ceilings if students do not go beyond the 30 credits per semester that are on average required to graduate on time.

While positive feedback has large and statistically significant effects, the same is not true for other types of feedback. Table 7 shows that negative feedback (Columns 1 and 2) has no significant effect on subsequent performance. Students whose achieved credits exactly match the average receive ambiguous feedback, i.e. approving framing plus descriptive data which shows that their performance was average. Interestingly, the coefficient is negative, albeit not statistically significant (Columns 3 and 4). We further investigate the effects of ambiguous treatment in Section 4.5.

The finding that positive feedback leads to an increase in subsequent performance is broadly consistent with the theoretical mechanisms discussed in Section 3. We find no effect of negative feedback, which is in line with the literature if either multiple mechanisms are relevant and work in different directions (within or across individuals), or if individuals discount or discard negative information, e.g. in order to preserve their self-concept. We will provide tentative evidence for the latter in Section 4.6.

Field Experiment II: Replication. In order to validate the results we found in the original experiment, we replicate our study one year later with a new cohort of first-year students. Our prediction for the replication was that there would again be an effect of positive feedback on performance, and no significant effects of the other feedback types. The middle panel of Table 6 shows the results for the replication experiment. The estimated treatment effects for the entire sample in Columns (1 to 5) are very similar to the original experiment. More importantly, we replicate the result that receiving positive feedback significantly increases performance (Columns 7 to 11). This is not only interesting because the positive feedback again generated performance-enhancing effects, but in particular because the minimal change in design now extends the positive effects to a number of weaker students (who obtain median or slightly below median credits). Although the differences in treatment effects across original experiment and replication are not statistically significant, it appears that the effects are somewhat smaller in the replication study. Even so, positive feedback increases performance by 1.7 credits, or .16 control group standard deviations.

As in the original experiment, the results of positive feedback are mainly driven by those between the 50th and the 80th percentile (Columns 5 to 8 in Table 7). From the bottom panels in Figure 8 and Figure 9 the same picture as before emerges. The effects of positive feedback are larger in the bottom half of the credit distribution of positive feedback, and we again see only small effects for the top part of the credit distribution.

The patterns for the other subgroups in Table 7 are also comparable. Again, we find no significant change in performance with regard to negative feedback (though the coefficient is negative). For the few people who still obtained exactly the average number of credits (N= 29), we again find an insignificant negative coefficient.

Pooled results. Because the two experiments share the same design, we can pool the observations from both cohorts to increase the power of our statistical analysis. The bottom panel of Table 6 presents the results for the pooled sample. As expected, the estimated treatment effect lies between the original and replication experiment at roughly 1.8 credits (.17 control group standard deviations). The estimates are more precise due to the larger sample size, and the effects of positive feedback are statistically significant at the 1%-level.

Overall, the results of the two field experiments provide consistent and robust evidence that normatively framed relative performance feedback leads to an increase in subsequent performance, but only if it is positive. The fact that we are able to reproduce the results is an important feature of our study, especially against the backdrop of inconclusive findings on relative performance feedback in the literature.

Magnitude of the effects. The effect sizes for positive feedback are in the range of .16 to .2 control group standard deviations and correspond to a 6-8% increase in performance relative to the control group. In the literature, the effect sizes for relative performance feedback are very much context-dependent. Studying the effects of relative performance information for workers, Blanes i Vidal and Nossol (2011) find an average effect of 6.8%. In a lab setting Kuhnen and Tymula (2012) find an increase in output of 12.2%. In education, Azmat and Iriberri (2010) find that the provision of information leads to a 5% improvement in grades and Tran and Zeckhauser (2012) find increases in performance after rank provision of about 10%. Our effects are in the mid-range of these estimates.

The same is true when considering the effect sizes that are typically found for other types of interventions in higher education: Dobkin, Gil and Marion (2010) find that a 10 percentage point increase in university class attendance increases performance by .17 standard deviations. Bandiera, Larcinese and Rasul (2010) find a decrease in performance of .11 standard deviations in response to a one standard deviation increase in lecture size. Carrell and West (2010) show that an increase in professor quality raises academic performance by .05 standard deviations, and Carrell, Fullerton and West (2009) find that an increase in peer quality has an effect of .08 standard deviations. Finally, we can compare our results to other studies that use obtained credits as an outcome. Leuven, Oosterbeek and Klaauw (2010) find that monetary incentives increase the obtained credits of high ability students by about 17%. De Paola, Scoppa and Nisticò (2012) also look at financial incentives, and find that it leads to an increase in obtained credits by .18 standard deviations.

Overall, this tentatively suggests that providing a specific type of feedback – positive feedback – may be an attractive way of inducing changes in behavior, especially when considering the low cost compared to some of the other interventions just mentioned.[29]

4.4 Positive feedback or unobserved factors?

So far, we have shown robust evidence that positive feedback substantially improves student performance. One possible explanation could be the following: Students receive positive feedback (or qualify for positive feedback but were allocated to the control group), because they have a higher first semester performance, and are therefore also likely to have on average different underlying characteristics than students who do not receive positive feedback. For example, they may have higher ability and a better learning technology at their disposal, and therefore they may be more able to respond to the treatment. In that case, the positive nature of the treatment may not matter, it could simply be the case that high ability students react better to relative performance feedback. In the following, we show that higher ability or other unobserved characteristics of students that receive positive feedback cannot explain our findings, and that it is in fact the positive nature of the feedback that changes the subsequent performance.

Similar to Allcott (2011) the design of our feedback intervention allows us to employ a sharp regression discontinuity design (RDD) for the treated students, as receiving positive or negative feedback follows a clear rule. Because here we are interested in the effect of positive versus negative feedback, we exclude individuals that receive ambiguous feedback. When implementing the RDD we follow in large parts the suggestions of Lee and Lemieux (2010). If the usual assumption for RDD holds, i.e. if there are no other discontinuities around the cut-off, it will provide a causal local average treatment effect (LATE) of receiving positive instead of negative feedback. To gather some intuition if this assumption is likely to hold, we can look at the behavior of the control group in absence of relative feedback information.[30] As running variable we are using the accumulated credits (ACP) a student obtained in the

[29] Table 3 shows that our intervention has a per student cost of under €2.5 per semester.

[30] Another assumption is that individuals have no or only imprecise control over the assignment variable (Lee and Lemieux 2010). This is very likely to hold in our case as when studying for their first semester exams, individuals do not know that they are going to receive feedback (let alone what form the feedback will have). Even if they did know, it would be virtually impossible to infer the exact value of the average performance in their comparison group or to precisely determine their position in the distribution of the assignment variable.

first semester, divided by the average ACP of her respective comparison group[31] (the corresponding distributions for the first experiment are shown in the left panels of Figure 10). Next, the left panels in Figure 11 provide a graphical depiction of the behavior of the outcome variable around the cut-off. For the treatment group we see a notable jump in the second semester credits, while we only observe a small jump for the control group. One explanation for the small jump in the control group could be the exclusion of individuals that receive ambiguous feedback, which could lead to discontinuities in the distributions of unobserved variables that are correlated with the obtained credits in the second semester.[32] To estimate the size of the jumps we implement a parametric RDD, using the following equation:

$$Y_i^k = \alpha_0 + \alpha_1 P_i + f(x_i) + f(x_i) P_i + \mathbf{s_i} \boldsymbol{\alpha_2} + \varepsilon_i, \tag{6}$$

where P_i indicates if a person receives positive feedback or, in case of the control group, is eligible to receive it. $f(x_i)$ is any smooth function of the running variable x_i that we allow to vary between the left and the right side of the cut-off and $\mathbf{s_i}$ is a vector including study program fixed effects and, in pooled estimations, also a cohort fixed effect and its interaction with the study program fixed effects.[33] The two upper panels of Columns (1) and (2) in Table 8 provide the estimated coefficients of α_1 for the treatment and the control group using two different discontinuity samples and a first order polynomial. The estimations confirm the results from the graphical illustration. For the treatment group the RDD indicates that receiving positive instead of negative feedback increases subsequent performance by at least seven credits. However, because of their size, the insignificant estimates for the control group could still suggest unobserved discontinuities around the cut-off. To account for any jump in our outcome variable due to unobserved discontinuities that are the same in the treatment and the control group we estimate the following regression discontinuity difference-in-difference (RD-DID) specification:[34]

$$\begin{aligned} Y_i^k = {} & \alpha_0 + \alpha_1 Treatment_i + \alpha_2 P_i + \alpha_3 Treatment_i P_i \\ & + x_i + x_i Treatment_i + x_i P_i + x_i Treatment_i P_i + \mathbf{s_i} \boldsymbol{\alpha_4} + \varepsilon_i, \end{aligned} \tag{7}$$

[31]This provides a smoother distribution around the cut-off than using the raw distance to the cut-off, because of differences in the credit point distributions within the different comparison groups.

[32]Another reason for a jump in the control group outcome at the cut-off could be that students share the content of the letters with each others, which could create spillovers. Observing no jump for the control group would therefore indicate that this is either not the case, or that there are no effects from this spillover on control group behavior.

[33]We do not include the vector $\mathbf{x_i}$, $\mathbf{z_i}$, and the baseline GAP as covariates, as this can make it difficult to differentiate between an inappropriate functional form and discontinuities in the covariates (see Lee and Lemieux 2010).

[34]See e.g., Danzer and Lavy (2018) or Dustmann and Schönberg (2012) for papers that make use of similar RD-DID specifications.

where we are interested in the parameter α_3. The results for this specification are shown in the bottom panel of Table 8. The coefficients in Column (1) and (2) indicate that there is still an effect of receiving positive instead of negative feedback, though the effect loses much of its significance.

As before, we can study the robustness of those findings by looking at the results for the second experiment. Using the mean as a cut-off leads to some differences. The right panel of Figure 10 shows that larger parts of the distribution are now above the cut-off, in part because the median of this distribution lies above the mean. Moreover, there are now far fewer students that receive ambiguous feedback, and therefore, less observations are excluded from the analysis. The results of this can be seen in the RDD plots in the right panel of Figure 11. The jump for the treatment group is larger in the second experiment and we do not observe any jump for the control group, which suggests that it is indeed the higher number of excluded individuals that leads to the jump in the control group of the first experiment. The corresponding estimates of Equations 6 and 7 are shown in Columns (3) and (4) of Table 8 and they confirm the results from the first experiment.

Columns (5) and (6) of Table 8 show the results for the pooled sample. The higher number of observations also allows us to show that the estimated coefficients are robust to different polynomial specifications and discontinuity samples (Table 9).[35] As an additional robustness check, Figures 12 to 15 depict the behavior of pre-treatment covariates around the cut-off for the two experiments. Many of them behave smoothly. In the cases where we do observe discontinuities, they behave very similar in the treatment and the control group. Therefore, Equation 7 should provide credible estimates.

The RDD results provide robust evidence that receiving positive instead of negative feedback leads to an increase in performance by at least six credits, for a student around the cut-off. This shows that for students of the same underlying ability the type of feedback is crucial for subsequent performance. This is in stark contrast to Allcott (2011), Brent, Lott, Taylor, Cook, Rollins and Stoddard (2017), and Costa and Kahn (2013) who implement similar RDD designs in an environmental context and find zero effects around the cutoff – which they interpret as evidence that varying normative framing alone does not affect behavior. The fact that we find a different effect and reproduce it in another experiment indicates that the context in which relative feedback and normative frames are provided plays an important role, and that the combination of both should receive more attention in contexts with complex and high-stakes tasks.

Regarding the theoretical considerations in Section 3, the results of the RDD tentatively

[35]With smaller discontinuity samples, higher order polynomials lead to overfitting and the corresponding estimates for the jump become imprecise and unreliable.

indicate that more precise knowledge about one's ability alone cannot fully explain the increase in performance. The RDD minimizes the difference in the credit points to the left and right of the cut-off point, and information updating about one's ability should lead to a continuous increase in performance. The approving normative framing of above average performance (aligned with matching descriptive information) could generate a discontinuous jump in subsequent performance, and if positive and negative feedback have diverging effects on performance, e.g. because negative information is discounted, this could also trigger a jump in performance.

4.5 Approving normative framing with(out) matching fundamentals

With the RDD we have shown that receiving positive instead of negative feedback matters for students around the cut-off. The positive feedback consists of two elements that negative feedback does not have: an approving normative framing and matching fundamentals showing that the student is better than average. This leads to the question of whether the approving normative framing alone can raise performance. The fact that in the original experiment we find no effect (even negative point estimates) for students on the median suggests otherwise (these students receive a normative message of approval without the matching fundamentals; see the results in Table 7). However, the lack of positive effects on these individuals may be due to some unobserved underlying characteristic.

The replication experiment allows us to further investigate this. Again, we saw no significant effect and a negative point estimate for students with ambiguous feedback in Experiment II (Table 7). But in addition, switching the peer information from the median to the mean in Experiment II allows more insights. We can calculate the median performance, and identify students who under the median rule of the original experiment would have received ambiguous feedback. Because they are now being compared to the mean, and mean credits are lower than median credits in most comparison groups, 80 of those students now receive positive feedback or are eligible to receive it. Assuming that those obtaining median credits are similar across cohorts, we can compare the treatment effect for those on the median in the original experiment with the treatment effect for those on the median in the replication. The only difference between the two groups is that those in the replication are shown fundamentals that match the approving normative message, and hence receive positive feedback.

In Table 10 we show the results from this exercise and see that the treatment coefficient for those on the median in the replication is positive – although not statistically significant. We take this as tentative evidence that the absence of a treatment effect for ambiguous feedback is not due to underlying characteristics of those students that have a median first semester performance. It is more plausible that an approving normative framing without the matching fundamentals has no effect on performance. Our results then tentatively suggest that providing the same individuals with matching fundamentals in addition to the normative message of approval may actually turn the zero-effect into a positive effect on performance.

The above finding additionally suggests that competitive preferences alone do not explain our results. Students who receive ambiguous feedback do not improve their performance. However, unless ambiguous feedback is disregarded, individuals with competitive preferences should exert more effort when the information about the reference group's performance becomes more precise (see e.g., Azmat and Iriberri 2010).

4.6 Expectations and Treatment Effects

As discussed in Section 3 there are a number of theoretical reasons why students might react to feedback. One prerequisite in all considered mechanisms is that the feedback provides new or unexpected information. Therefore, a natural question that arises is how student expectations about relative performance line up with actual relative performance. In the replication experiment, we surveyed students on their expected performance in terms of credits and GPA (Table 11 shows the questions[36]).

One caveat applies when considering the data from the surveys: the response rates are between 15 and 20% (see Table 21 in the Data and Methods Appendix). Accordingly our sample is rather small, and we find evidence that the respondents are a positively selected subpopulation. It is reassuring that in the subsample we use for the estimations in Table 12 the covariates are overall well balanced between the treatment and the control group (see Table 22).

We first assess how accurately students predict their actual rank in terms of credits and GPA in the pre-treatment semester. The top panel in Figure 16 plots the rank students expected to have in the first semester against the actual rank, in terms of credits. The insignificant rank correlation coefficient of .084 indicates that students have very little intuition about their actual relative performance. The bottom panel shows the corresponding figure for the GPA. In this case, the significant rank correlation coefficient of .228 suggests

[36]The wording in the pre- and post-treatment survey is not the same. While this may have level effects on control and treatment groups simultaneously, it should not affect the argument about differential updating between treatment and control we make in this section. Also, in the surveys we asked students about their expectations in terms of the rank/percentile, while our feedback provides them with the information if they performed above/below the average. To make our argument consistent we assume that expecting to belong to the top 50% (bottom 50%) is equivalent to expecting an above (below) average performance for the students. If students actually make a distinction between median and mean, due to the mean being smaller than the median this means that e.g. not everyone who expected to be below the 50th percentile will be surprised by positive feedback – this measurement error should make it harder for us to find results.

that students have at least some idea about their actual relative performance in terms of GPA, even before they receive their first grades. This supports our notion that relative performance feedback on grades as in Azmat, Bagues, Cabrales and Iriberri (2018) may well have different effects compared to our study, because students already have a better idea about the underlying performance distribution in terms of grades.

The second important question is whether students incorporate the feedback information into their post-treatment expectations. The top panel of Figure 17 displays the share among above average students who correctly expected an above average performance, before and after receiving positive feedback. The left panel supports what we saw in Figure 16. Before our treatment, only around 50% of students eligible for positive feedback correctly expected to perform above average, both in the treatment and the control group. In the semester after the start of our intervention, among above average performers, 87% in the treatment group and 68% in the control group correctly expected to be above average performers, and the difference of 19 percentage points is significant at the 5% level. This suggests two things. First, those who performed above average learn about their relative performance, even in the absence of relative feedback. Second, and more importantly, providing students with feedback leads to a stronger update in the expected relative performance.

The bottom panel of Figure 17 shows the share among below average students who incorrectly expected an above average performance. In the first semester, nearly half of these students overestimated their performance. Interestingly, in this case neither the control nor the treatment group appears to update their expected relative performance in the second or third semester. However, the number of below average students who took part in the survey is very low (N in the three surveys is never larger than 30), and this is reflected in the large confidence intervals. We thus very cautiously take this as evidence that students who receive negative feedback do indeed discard or discount the information.

Finally, as discussed in Section 3, we expect that individuals who initially expected to perform below average respond to positive feedback. These students received new information, in contrast to students who already expected that they would perform above average. To test if this does indeed affect the treatment effect, we create a dummy U_i that is 1 if students underestimated their performance in the first semester, i.e. they expected a below average performance, although they then actually performed better than the average. We estimate the following equation for all students that were better than average and therefore received positive feedback, or were eligible to receive it:[37]

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \alpha_2 U_i + \alpha_3 Treatment_i U_i + \mathbf{s_i}\boldsymbol{\alpha_4} + \varepsilon_i, \quad (8)$$

where α_1 gives the treatment effect for those who correctly expected to be above average, and α_3 gives the difference in the treatment effect for those who underestimated their relative performance. Table 12 presents the results. Column (2) shows that the treatment effect for those who correctly estimated their position is roughly 1.6 credits, which is 1.1 points smaller than the treatment effect that we find in the survey sample (Column 1) . Control group students who underestimated their relative performance obtain on average 2.1 credits less in the second semester. However, the interaction suggests that informing those students that they actually are above the average can increase their performance to the level of those who correctly anticipated to be above average. In Column (3 to 5) we add our controls. Especially the pre-treatment performance measures are likely to be correlated with the first semester expectations. The negative effect of underestimating performance becomes stronger and significant. Given the controls, this may indicate that at equal ability, lower confidence is detrimental to performance. Again, the effect is completely offset by receiving positive feedback. The treatment generates a significant effect of 5.3 credits for students that underestimated their relative performance. This indicates that relative performance feedback at the beginning of university might be especially helpful for students who underestimate their relative performance in the absence of such information.

The last insight comes from Figure 17. The top right shows that in the third semester a large majority of students, both in the treatment and the control group, correctly expected an above average performance. If differences in beliefs about being an above average student between treatment and control group produce our results (e.g. by increasing motivation and effort), one could expect that there will be no differences in the third semester performance, because both groups now share the same beliefs.

4.7 Repeated Treatment

It is not straightforward how to evaluate the consequences of repeated treatment. Once the interest is centered on the effects of the different types of feedback, some caveats regarding the analysis have to be considered. First, it is endogenous what type of feedback students receive in the third semester, because their relative performance in the second semester is affected by the initial feedback letters. This is illustrated with the transition probabilities

[37] The number of survey respondents who perform below the average is too low to study their behavior more formally.

(and their differentials) of the pooled sample in Table 20. They show the movement between different types of feedback, from the beginning of the second to the beginning of the third semester. Both in the treatment and the control group, only about 5% of the students that were below the average at the end of the first semester are able to move above it by the end of the second semester. Also, while students of both groups tend to stay above average, treated students are more likely to do so, because of their improved performance. As a consequence of this differential movement within the credit distribution after the initial feedback letters, we only analyze the effects of receiving a certain type of feedback in the second semester on performance in the third semester.

Second, as is also indicated by the transition matrices, our intervention could lead to differential drop-out behavior. Columns (3 and 4) and (7 and 8) of Table 17 show the effects of treatment on the dropout decision for both experiments and the pooled sample. While most of the coefficients are insignificant, the signs indicate that positive feedback leads to lower dropout. For the analysis of repeated treatment we exclude all students that dropped out before the start of the third semester. This is conservative if we assume that students who drop out are those with lower performance, and we therefore might underestimate the true effect of repeated treatment.[38]

The results are displayed in Table 13 and in Table 14. We find no significant effects on the third semester performance across all types of feedback. The findings hold in both experiments and in the pooled sample. Table 15 shows Lee bounds (Lee 2009) for the effects of the different feedback types in the third semester. Overall, the lack of effects for repeated treatment is in line with our findings in Section 4.6 and the idea that treatment is effective if students receive new information from the feedback letters.

4.8 Channels and effects on other domains: exams taken, exams failed, GPA, and well-being

We have found robust effects of positive feedback on credit points, the target of the intervention. What are the channels by which students generate these gains in performance? Do they sacrifice grade achievement for gains in credits in the process? Are there perhaps other (psychological) costs associated with encouraging students to obtain more credits?

First, we investigate the two ways in which students can obtain more credits: attempt more exams and/or fail fewer exams. Columns (5) to (8) of Table 16 show that students who receive positive feedback do both. The effect on attempted exams is larger in the first exper-

[38]Tables 23 and 24 in the Data and Methods Appendix show that this leaves us with a sample in which the covariates are overall well balanced between treatment and control group.

iment, while the effect on the number of failed exams is larger in the second experiment. In the pooled sample, treated students attempt about .17 more exams and fail .11 less. Multiplying those coefficients by the average number of credits students are awarded for an exam (5.75 credit points) and adding them up comes very close to the treatment effect for the pooled sample in Table 6.

Second, a concern could be that encouraging students to obtain more credits may come at a cost in terms of lower grades, because students may shift attention away from grades. In Columns (1 and 2) and (5 and 6) of Table 17 we report the effects on the second semester GPA. We do not find any significant effects. It appears that positive feedback can raise performance in terms of obtained credits, without negatively affecting the other performance dimension.

Lastly, one might worry that the feedback affects other dimensions that are not directly related to performance. For both our experiments we conducted a post-treatment survey (Figure 1) in which we asked students how often they attend the lectures, how satisfied they are with their life, the degree to which they are satisfied with their study program, the degree to which they are satisfied with their performance, and how stressful they find their studies (see Table 18 for the survey questions and the variables used in the estimations).[39] Table 19 shows the effects of negative and positive feedback on these outcomes.[40]. We find no indication that the effects of positive feedback are driven by an increase in lecture visits. We also find no evidence that they come at the cost of being less satisfied with life, studying or performance, or that they are accompanied by a higher stress level. If anything, positive feedback might increase life and study satisfaction. For the small number of students who received negative feedback and answered the survey, the results tentatively suggest that it increases lecture attendance, life, study, and performance satisfaction.

5 Conclusion

We investigated the effects of normatively framed relative performance feedback in a setting where individuals enter a new environment, and are therefore unfamiliar with the underlying performance distribution. In the first experiment we find evidence for differential effects of positive and negative feedback. The replication shows that the results are robust: we reproduce all results from the original experiment.

[39]Table 25 in the Data and Methods Appendix shows that even in this subsample most of the covariates are still balanced between the treatment and the control group. Note that the table also includes students who received ambiguous feedback, which explains the higher number of observations compared to Table 18

[40]The number of observations can vary between the outcomes, as students were allowed to give no answers to the questions in the survey.

Positive relative performance feedback improves performance, and we show that the increase in obtained course credits is indeed the effect of the positive nature of the feedback, and not of some unobserved characteristics. Because we observe the entire universe of university performance (including grades as the second performance dimension), our results can be interpreted as net gains. We find no evidence for effects of negative feedback, and normative messages of approval alone (i.e. ambiguous feedback) do not increase subsequent performance, either. Furthermore, we provide suggestive evidence that positive feedback is especially effective for those who underestimate their relative performance, and that the treatment effects are linked to a more accurate assessment of performance. Absent treatment, above average students over time also learn about their relative ability, and the performance of treatment and control group converges. This is also in line with the notion that it is indeed the informational advantage that generates the performance gains.

Our experiments were not designed with the goal of testing different theories. Yet the lack of evidence for significant effects of negative feedback at first glance seems to suggest that neither competitive preferences nor social norms can explain the results, as both would predict performance gains with negative feedback. However, we find survey evidence that negative feedback is not processed in the same way as positive feedback. If students discard negative feedback, our results are, in fact, compatible with all the theoretically stipulated mechanisms. While the findings are thus broadly consistent with theory, they leave some interesting questions and open up avenues for future research that can disentangle the potential mechanisms. It would also be worth investigating whether the combination of normative frame with information leads to a larger effect than relative feedback without normative messages. Due to the fixed cohort size and the resulting power, our experiments did not test this.

The results also have important implications from a policy perspective. The intervention increases student performance at the beginning stages of university, arguably a crucial time for getting on track to graduation. It is also low cost, and easy to implement in other contexts where individuals enter a new environment, e.g. when starting a new job. Importantly, the fact that we find and replicate performance enhancing effects where Azmat, Bagues, Cabrales and Iriberri (2018) and Cabrera and Cid (2017) do not find performance gains, suggests that the chosen performance dimension and the initial level of information about the underlying performance distribution may matter. In addition, our results suggest that individuals who perform above average benefit from relative feedback while there is no evidence that negative feedback is helpful. Future feedback schemes can be designed to take this heterogeneity into account and attempt to extend the benefits to below average students. One approach could be to combine relative feedback for above average students with other forms, e.g. self-referential feedback, for struggling students.

References

Allcott, Hunt, "Social Norms and Energy Conservation," *Journal of Public Economics*, 2011, *95* (9), 1082–1095.

___ **and Todd Rogers**, "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation," *The American Economic Review*, 2014, *104* (10), 3003–3037.

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

___ **and** ___ , "The Provision of Relative Performance Feedback: An Analysis of Performance and Satisfaction," *Journal of Economics & Management Strategy*, 2016, *25* (1), 77–110.

___ **, Manuel Bagues, Antonio Cabrales, and Nagore Iriberri**, "What You Don't Know... Can't Hurt You? A Natural Field Experiment on Relative Performance Feedback in Higher Education," *mimeo*, 2018.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul, "Team Incentives: Evidence from a Firm Level Experiment," *Journal of the European Economic Association*, 2013, *11* (5), 1079–1114.

___ **, Valentino Larcinese, and Imran Rasul**, "Heterogeneous Class Size Effects: New Evidence from a Panel of University Students," *The Economic Journal*, 2010, *120* (549), 1365–1398.

Bandura, Albert, "Social Cognitive Theory of Self-Regulation," *Organizational Behavior and Human Decision Processes*, 1991, *50* (2), 248 – 287.

Barankay, Iwan, "Rank Incentives: Evidence from a Randomized Workplace Experiment," *mimeo*, 2012.

Bénabou, Roland and Jean Tirole, "Self-Confidence and Personal Motivation," *The Quarterly Journal of Economics*, 2002, *117* (3), 871–915.

___ **and** ___ , "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, September 2016, *30* (3), 141–64.

Blanes i Vidal, Jordi and Mareike Nossol, "Tournaments Without Prizes: Evidence from Personnel Records," *Management Science*, 2011, *57* (10), 1721–1736.

Bound, John, Michael F Lovenheim, and Sarah Turner, "Increasing Time to Baccalaureate Degree in the United States," *Education Finance and Policy*, 2012, *7* (4), 375–424.

Brent, Daniel A, Corey Lott, Michael Taylor, Joseph Cook, Kimberly Rollins, and Shawn Stoddard, "Are Normative Appeals Moral Taxes? Evidence from a Field Experiment on Water Conservation," *mimeo*, 2017.

Bruhn, Miriam and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 200–232.

Brunstein, Joachim C and Peter M Gollwitzer, "Effects of Failure on Subsequent Performance: The Importance of Self-Defining Goals," *Journal of Personality and Social Psychology*, 1996, *70* (2), 395.

Cabrera, José María and Alejandro Cid, "Gender Differences to Relative Performance Feedback: A Field Experiment in Education," *mimeo*, 2017.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan et al., "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 2016, *351* (6280), 1433–1436.

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez, "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction," *American Economic Review*, 2012, *102* (6), 2981–3003.

Carrell, Scott E and James E West, "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy*, 2010, *118* (3), 409–432.

___ , Richard L Fullerton, and James E West, "Does Your Cohort Matter? Measuring Peer Effects in College Achievement," *Journal of Labor Economics*, 2009, *27* (3), 439–464.

Castleman, Benjamin L and Lindsay C Page, "Summer Nudging: Can Personalized Text Messages and Peer Mentor Outreach Increase College Going Among Low-Income High School Graduates?," *Journal of Economic Behavior & Organization*, 2015, *115*, 144–160.

___ **and** ___ , "Freshman Year Financial Aid Nudges: An Experiment to Increase FAFSA Renewal and College Persistence," *Journal of Human Resources*, 2016, *51* (2), 389–415.

Charness, Gary, David Masclet, and Marie Claire Villeval, "The Dark Side of Competition for Status," *Management Science*, 2013, *60* (1), 38–55.

Chetty, Raj, "Behavioral Economics and Public Policy: A Pragmatic Perspective," *American Economic Review*, 2015, *105* (5), 1–33.

Cialdini, Robert B, "Crafting Normative Messages to Protect the Environment," *Current Directions in Psychological Science*, 2003, *12* (4), 105–109.

___ , "The Focus Theory of Normative Conduct," *Handbook of Theories of Social Psychology*, 2011, *2*, 295–312.

___ , Linda J Demaine, Brad J Sagarin, Daniel W Barrett, Kelton Rhoads, and Patricia L Winter, "Managing Social Norms for Persuasive Impact," *Social Influence*, 2006, *1* (1), 3–15.

___ , Raymond R Reno, and Carl A Kallgren, "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places," *Journal of Personality and Social Psychology*, 1990, *58* (6), 1015–1026.

Compte, Olivier and Andrew Postlewaite, "Confidence-Enhanced Performance," *American Economic Review*, 2004, *94* (5), 1536–1557.

Corcoran, K., J. Crusius, and T. Mussweiler, "Social Comparison: Motives, Standards, and Mechanisms," in Derek Chadee, ed., *Theories in Social Psychology*, Wiley-Blackwell, 2011.

Costa, Dora L and Matthew E Kahn, “Energy Conservation “Nudges” and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment,” *Journal of the European Economic Association*, 2013, *11* (3), 680–702.

Damgaard, Mette Trier and Helena Skyt Nielsen, “Nudging in Education,” *Economics of Education Review*, 2018, *forthcoming.*

Danzer, Natalia and Victor Lavy, “Paid Parental Leave and Children’s Schooling Outcomes,” *The Economic Journal*, 2018, *128* (608), 81–117.

Delfgaauw, Josse, Robert Dur, Joeri Sol, and Willem Verbeke, “Tournament Incentives in the Field: Gender Differences in the Workplace,” *Journal of Labor Economics*, 2013, *31* (2), 305–326.

Deutsch, Morton and Harold B Gerard, “A Study of Normative and Informational Social Influences upon Individual Judgment,” *The Journal of Abnormal and Social Psychology*, 1955, *51* (3), 629.

Dobkin, Carlos, Ricard Gil, and Justin Marion, “Skipping Class in College and Exam Performance: Evidence from a Regression Discontinuity Classroom Experiment,” *Economics of Education Review*, 2010, *29* (4), 566–575.

Dustmann, Christian and Uta Schönberg, “Expansions in Maternity Leave Coverage and Children’s Long-Term Outcomes,” *American Economic Journal: Applied Economics*, 2012, *4* (3), 190–224.

Duvendack, Maren, Richard Palmer-Jones, and W Robert Reed, “What Is Meant by "Replication" and Why Does It Encounter Resistance in Economics?,” *American Economic Review*, 2017, *107* (5), 46–51.

Eil, David and Justin M Rao, “The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself,” *American Economic Journal: Microeconomics*, 2011, *3* (2), 114–138.

Eriksson, Tor, Anders Poulsen, and Marie Claire Villeval, “Feedback and Incentives: Experimental Evidence,” *Labour Economics*, 2009, *16* (6), 679–688.

Ertac, Seda, “Social Comparisons and Optimal Information Revelation: Theory and Experiments,” *Job Market Paper, UCLA*, 2005.

Festinger, Leon, “A Theory of Social Comparison Processes,” *Human Relations*, 1954, *7* (2), 117–140.

Fischer, Mira and Valentin Wagner, “Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment in Secondary Schools,” *mimeo*, 2017.

Garibaldi, Pietro, Francesco Giavazzi, Andrea Ichino, and Enrico Rettore, “College Cost and Time to Complete a Degree: Evidence from Tuition Discontinuities,” *Review of Economics and Statistics*, 2012, *94* (3), 699–711.

Gibbons, Robert and Kevin J Murphy, “Relative Performance Evaluation for Chief Executive Officers,” *ILR Review*, 1990, *43* (3), 30–S.

Gill, David, Zdenka Kissová, Jaesun Lee, and Victoria Prowse, “First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision,” *Management Science*, 2018, *forthcoming.*

Goldstein, Noah J, Robert B Cialdini, and Vladas Griskevicius, "A Room With a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels," *Journal of Consumer Research*, 2008, *35* (3), 472–482.

Golman, Russell, David Hagmann, and George Loewenstein, "Information Avoidance," *Journal of Economic Literature*, 2017, *55* (1), 96–135.

Goulas, Sofoklis and Rigissa Megalokonomou, "Knowing Who You Are: The Effect of Feedback Information on Short and Long Term Outcomes," *Warwick Economics Research Paper Series 1075*, 2015.

Grossman, Zachary and David Owens, "An Unlucky Feeling: Overconfidence and Noisy Feedback," *Journal of Economic Behavior & Organization*, 2012, *84* (2), 510–524.

Hallsworth, Michael, John A List, Robert D Metcalfe, and Ivo Vlaev, "The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance," *Journal of Public Economics*, 2017, *148*, 14–31.

Hertwig, Ralph and Christoph Engel, "Homo Ignorans: Deliberately Choosing Not to Know," *Perspectives on Psychological Science*, 2016, *11* (3), 359–372.

Himmler, Oliver, Robert Jäckle, and Philipp Weinschenk, "Soft Commitments, Reminders and Academic Performance," *American Economic Journal: Applied Economics*, forthcoming.

Hoxby, Caroline and Sarah Turner, "Expanding College Opportunities for High-Achieving, Low Income Students," *SIEPR Discussion Paper 12-014*, 2013.

Ilgen, Daniel and Cori Davis, "Bearing Bad News: Reactions to Negative Performance Feedback," *Applied Psychology*, 2000, *49* (3), 550–565.

Ilgen, Daniel R, Cynthia D Fisher, and M Susan Taylor, "Consequences of Individual Feedback on Behavior in Organizations.," *Journal of Applied Psychology*, 1979, *64* (4), 349.

Kajitani, Shinya, Keiichi Morimoto, and Shiba Suzuki, "Relative Performance Information Feedback and Just-Pass Behavior: Evidence from a Field Experiment," *Meisei University Working Papers 36*, 2017.

Kandel, Eugene and Edward P Lazear, "Peer Pressure and Partnerships," *Journal of Political Economy*, 1992, *100* (4), 801–817.

Katreniakova, Dagmara, "Social Comparison, Rewards and Incentives to Learn: A Randomized Control Trial in Uganda," *mimeo*, 2014.

Köszegi, Botond, "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, 2006, *4* (4), 673–707.

Kuhnen, Camelia M and Agnieszka Tymula, "Feedback, Self-Esteem, and Performance in Organizations," *Management Science*, 2012, *58* (1), 94–113.

Lee, David S, "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 2009, *76* (3), 1071–1102.

—— **and Thomas Lemieux**, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 2010, *48* (2), 281–355.

Leuven, Edwin, Hessel Oosterbeek, and Bas Klaauw, "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment," *Journal of the European Economic Association*, 2010, *8* (6), 1243–1265.

Madrian, Brigitte C, "Applying Insights from Behavioral Economics to Policy Design," *Annual Review of Economics*, 2014, *6* (1), 663–688.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat, "Managing Self-Confidence," *mimeo*, 2014.

Morgan, Kari Lock and Donald B Rubin, "Rerandomization to Improve Covariate Balance in Experiments," *The Annals of Statistics*, 2012, *40* (2), 1263–1282.

OECD, "Education at a Glance 2013: OECD Indicators," Technical Report, OECD Publishing 2013.

Open Science Collaboration, "Estimating the Reproducibility of Psychological Science," *Science*, 2015, *349* (6251), aac4716.

Paola, Maria De, Vincenzo Scoppa, and Rosanna Nisticò, "Monetary Incentives and Student Achievement in a Depressed Labor Market: Results from a Randomized Experiment," *Journal of Human Capital*, 2012, *6* (1), 56–85.

Pearce, Jone L and Lyman W Porter, "Employee Responses to Formal Performance Appraisal Feedback.," *Journal of Applied Psychology*, 1986, *71* (2), 211.

Pikulina, Elena, Renneboog Luc, and Tobler N Philippe, "Do Confident Individuals Generally Work Harder?," *Journal of Multinational Financial Management*, 2018, *44*, 51–60.

Puri, Manju and David T Robinson, "Optimism and Economic Choice," *Journal of Financial Economics*, 2007, *86* (1), 71–99.

Schultz, P Wesley, Jessica M Nolan, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius, "The Constructive, Destructive, and Reconstructive Power of Social Norms," *Psychological Science*, 2007, *18* (5), 429–434.

Slemrod, Joel, "Tax Compliance and Enforcement: New Research and its Policy Implications," *Ross School of Business Working Paper 1302*, 2016.

Suls, Jerry and Ladd Wheeler, *Handbook of Social Comparison: Theory and Research* The Plenum Series in Social/Clinical Psychology, New York: Kluwer Academic/Plenum Publishers, 2000.

Sweeny, Kate, Darya Melnyk, Wendi Miller, and James A Shepperd, "Information Avoidance: Who, What, When, and Why," *Review of General Psychology*, 2010, *14* (4), 340.

Taylor, S.E., H.A. Wayment, and M. Carrillo, "Social Comparison, Self-Regulation, and Motivation," in R. M. Sorrentino and E.T. Higgins, eds., *Handbook of Motivation and Cognition, Vol. 3. The Interpersonal Context*, Guilford Press, 1996.

Tran, Anh and Richard Zeckhauser, "Rank as an Inherent Incentive: Evidence from a Field Experiment," *Journal of Public Economics*, 2012, *96* (9), 645–650.

Trope, Yaacov and Nira Liberman, "Construal-Level Theory of Psychological Distance," *Psychological Review*, 2010, *117* (2), 440.

Wiswall, Matthew and Basit Zafar, "Determinants of College Major Choice: Identification using an Information Experiment," *The Review of Economic Studies*, 2015, *82* (2), 791–824.

____ **and** ____ , "How Do College Students Respond to Public Information about Earnings?," *Journal of Human Capital*, 2015, *9* (2), 117–169.

Wood, Joanne V, "Theory and Research Concerning Social Comparisons of Personal Attributes.," *Psychological Bulletin*, 1989, *106* (2), 231.

Appendix

Table 1: DESCRIPTION OF VARIABLES

Variable	Description
Treatment Variables	
Treatment	Random assignment to the treatment group
Stratification Variables	
Study Program	Indicators for study programs; for more information see Table 2.
Credit Strata	Indicating strata based on accumulated first-semester credit points (ACP).[a)]
Control Variables	
Age	Age in years
Female	Indicator for being female
HS Degree Abitur	Indicator for a general track degree ("Abitur"); reference category includes vocational track degree ("Fachhochschulreife") and students who hold other degrees.
Time since HS Degree	Time in years since high school graduation
HS GPA	Final high school grade point average (1=best, 4=worst); missing values imputed.[a)]
GPA 1st Semester	First semester grade point average (course-level[b)]; 1=best, 4=worst); failed exams are not included in calculation. Missing values imputed.[a)]
Credits 1st Semester	Number of credit points (course-level)[b)] obtained in the first semester net of credits granted for an internship.[a)]
Feedback Variables	
ACP	Accumulated first semester credit points (module-level)[b)]; feedback type is based on this variable.
AGPA	Cumulative first semester grade point average (module-level[b)]; 1=best, 4=worst); failed exams are not included in calculation. Variable was used as balancing covariate in the randomization.
Outcome Variables[c)]	
Credits	Number of credit points obtained in the semester, net of credits granted for an internship
Attempted Exams	Number of attempted exams in the semester
Failed Exams	Number of failed exams in the semester; excludes students who attempted no exams.
GPA	Current grade point average (1=best, 4=worst); failed exams are not included in calculation.
Dropout	Indicator for having dropped out of the study program.

Note: a) For details see Data and Methods Appendix. *b)* Course-level: includes partly completed multiple-course-modules (= passed sub-modules). Module-level: considers only fully completed modules. For more details see Data and Methods Appendix. *c)* All outcome variables are measured on the course-level.

Table 2: STUDY PROGRAMS, NUMBER OF STUDENTS AND TREATMENT RATES

		Observations at start of intervention					
		Faculty		Experiment		Fraction in Treatment	
Study program	Degree	Business Administration	Mechanical Engineering	Experiment I	Replication Experiment II	Experiment I	Replication Experiment II
Business Administration (BA)	B.A.	735	-	402	333	50.25%	50.15%
International Business (IB)	B.A.	122	-	63	59	49.21%	50.85%
Business Engineering[a)] (BE)	B.Eng.	124	-	61	63	50.82%	50.79%
Mechanical Engineering (ME)	B.Eng.	-	533	235	298	50.21%	49.66%
Energy and Building Services Engineering (EBSE)	B.Eng.	-	95	51	44	49.02%	50.00%
N – Overall	1,609	981	628	812	797	50.12%	50.06%

Note: a) BE is a joint degree program of the business and the tech faculty. During the first semesters most courses are related to business administration and economics. We therefore assign BE to the business faculty.

Table 3: Summary of Cost Incurred by the Relative Performance Feedback (in Euros)

Cost calculation for relative performance feedback (cohort of 800)		
Student assistant	(60 hours per semester * € 11.70)	€ 702
Postage	(2 letters * € 0.48 * 800 students)	€ 768
Printing of letters	(2 letters * 2 pages * € 0.12 * 800 students)	€ 384
Printing of letters 2nd language	(2 letters * 2 pages * € 0.12 * 140 students)	€ 67.20
Envelopes	(2 letters * € 0.02 * 800 students)	€ 32
Total cost per semester		€ 1,953.20
Cost per student per semester		**€ 2.44**

Table 4: DESCRIPTIVE STATISTICS AND BALANCING PROPERTIES, ALL STUDENTS

	Experiment I			Experiment II: Replication		
	(1) Control Group Mean (Std. Deviation)	(2) Treatment Coefficient[a)] (Robust SE)	(3) p-Value	(4) Control Group Mean (Std. Deviation)	(5) Treatment Coefficient[a)] (Robust SE)	(6) p-Value
Age	22.514	-0.086	0.696	22.417	0.084	0.689
	(3.376)	(0.220)		(3.078)	(0.210)	
Female	0.395	0.001	0.976	0.344	-0.001	0.973
	(0.489)	(0.030)		(0.476)	(0.029)	
HS Degree Abitur	0.430	-0.010	0.766	0.399	0.017	0.604
	(0.496)	(0.033)		(0.490)	(0.034)	
Time since HS Degree	1.341	-0.094	0.561	1.171	-0.005	0.968
	(2.523)	(0.161)		(1.885)	(0.133)	
HS GPA	2.567	-0.011	0.758	2.555	-0.042	0.225
	(0.563)	(0.036)		(0.622)	(0.035)	
% HS GPA Imputed[b)]	0.012	0.002	0.754	0.020	-0.002	0.824
	(0.111)	(0.008)		(0.141)	(0.009)	
GPA 1st Semester	2.504	-0.057	0.168	2.602	-0.043	0.290
	(0.627)	(0.041)		(0.640)	(0.041)	
% GPA 1st Semester Imputed[b)]	0.067	-0.001	0.968	0.088	-0.009	0.585
	(0.250)	(0.015)		(0.284)	(0.017)	
Credits 1st Semester	20.236	0.348	0.263	18.660	0.207	0.557
	(10.187)	(0.311)		(11.170)	(0.353)	
Attempted Exams 1st Semester	4.699	0.041	0.609	4.814	-0.052	0.587
	(1.558)	(0.081)		(1.683)	(0.096)	
Failed Exams 1st Semester[c)]	0.768	0.084	0.140	1.072	-0.028	0.689
	(1.169)	(0.057)		(1.386)	(0.069)	
AGPA at Randomization	2.491	-0.039	0.479	2.584	-0.031	0.493
	(0.713)	(0.054)		(0.683)	(0.046)	
% AGPA at Randomization NA[b)]	0.264	-0.014	0.426	0.116	-0.007	0.707
	(0.441)	(0.017)		(0.320)	(0.018)	
N		812			797	

Note: Columns (1) and (4) present the unadjusted control group means and standard deviations of the covariates. For details on the variables see Table 1 and the Data and Methods Appendix. *a)* Columns (2) and (5) present the estimated coefficients of regressing the covariates on the treatment indicator using Equation 2. Columns (3) and (6) test the null hypothesis of no treatment effect. *b)* See the Data and Methods Appendix for details on the missing values and the imputation. *c)* Failed Exams excludes students who attempted no exams (27 in Experiment I and 32 in Experiment II). * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 5: DESCRIPTIVE STATISTICS AND BALANCING PROPERTIES, STUDENTS ELIGIBLE TO RECEIVE POSITIVE FEEDBACK

	Experiment I			Experiment II: Replication		
	(1) Control Group Mean (Std. Deviation)	(2) Treatment Coefficient (Robust SE)	(3) p-Value	(4) Control Group Mean (Std. Deviation)	(5) Treatment Coefficient (Robust SE)	(6) p-Value
Age	22.581	-0.412	0.175	22.085	-0.065	0.790
	(2.905)	(0.303)		(2.828)	(0.243)	
Female	0.394	-0.023	0.636	0.359	0.009	0.823
	(0.490)	(0.049)		(0.481)	(0.038)	
HS Degree Abitur	0.445	0.063	0.231	0.413	0.017	0.702
	(0.499)	(0.052)		(0.493)	(0.045)	
Time since HS Degree	1.426	-0.223	0.348	1.081	-0.035	0.819
	(2.433)	(0.237)		(1.699)	(0.152)	
HS GPA	2.445	-0.026	0.658	2.377	-0.037	0.435
	(0.515)	(0.060)		(0.581)	(0.048)	
% HS GPA Imputed$^{b)}$	0.019	-0.008	0.489	0.004	-0.004	0.321
	(0.138)	(0.012)		(0.067)	(0.004)	
GPA 1st Semester	2.252	-0.088	0.190	2.383	-0.019	0.723
	(0.597)	(0.067)		(0.601)	(0.053)	
% GPA 1st Semester Imputed$^{b)}$	0.006	-0.007	0.321	0.000	0.005	0.321
	(0.080)	(0.007)		(0.000)	(0.005)	
Credits 1st Semester	26.252	0.841	0.166	25.697	0.057	0.892
	(8.016)	(0.606)		(7.632)	(0.423)	
Attempted Exams 1st Semester	5.239	-0.073	0.475	5.404	-0.032	0.733
	(1.295)	(0.103)		(1.162)	(0.092)	
Failed Exams 1st Semester$^{c)}$	0.142	0.009	0.840	0.323	-0.063	0.225
	(0.418)	(0.046)		(0.719)	(0.052)	
AGPA at Randomization	2.286	-0.030	0.711	2.392	-0.017	0.753
	(0.711)	(0.081)		(0.624)	(0.054)	
% AGPA at Randomization NA$^{b)}$	0.071	-0.001	0.966	0.000	0.008	0.160
	(0.258)	(0.017)		(0.000)	(0.006)	
N		305			448	

Note: Columns (1) and (4) present the unadjusted control group means and standard deviations of the covariates. For details on the variables see Table 1 and the Data and Methods Appendix. *a)* Columns (2) and (5) present the estimated coefficients of regressing the covariates on the treatment indicator using Equation 2. Columns (3) and (6) test the null hypothesis of no treatment effect. *b)* See the Data and Methods Appendix for details on the missing values and the imputation. *c)* Failed Exams excludes students who attempted no exams (1 in Experiment II). * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 6: TREATMENT EFFECT ON CREDITS

	All Feedbacks					Control Mean (SD)	Positive Feedback					Control Mean (SD)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Experiment I	0.643	0.665	0.533	0.530	0.287	21.07	2.411**	2.539**	2.460**	2.501**	2.073**	26.32
	(0.884)	(0.730)	(0.713)	(0.709)	(0.695)	(12.34)	(1.151)	(1.075)	(1.042)	(1.060)	(1.000)	(10.50)
N	812	812	812	812	812		305	305	305	305	305	
Experiment II: Replication	0.677	0.617	0.357	0.415	0.312	19.75	1.837*	1.922**	1.735**	1.725**	1.706**	26.39
	(0.953)	(0.736)	(0.711)	(0.710)	(0.702)	(13.32)	(0.958)	(0.903)	(0.867)	(0.867)	(0.808)	(10.39)
N	797	797	797	797	797		448	448	448	448	448	
Pooled	0.661	0.641	0.457	0.480	0.327	20.42	2.067***	2.170***	2.044***	1.997***	1.810***	26.36
	(0.650)	(0.518)	(0.504)	(0.501)	(0.493)	(12.84)	(0.736)	(0.691)	(0.668)	(0.670)	(0.626)	(10.42)
N	1609	1609	1609	1609	1609		753	753	753	753	753	
Strata FE	NO	YES	YES	YES	YES		NO	YES	YES	YES	YES	
Ability Controls	NO	NO	YES	YES	YES		NO	NO	YES	YES	YES	
Background Variables	NO	NO	NO	YES	YES		NO	NO	NO	YES	YES	
GPA 1st Semester	NO	NO	NO	NO	YES		NO	NO	NO	NO	YES	

Note: Outcome variable: credits 2nd semester; *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 7: TREATMENT EFFECT ON CREDITS, BY FEEDBACK TYPE

	Negative Feedback		Control Mean (SD)	Ambiguous Feedback		Control Mean (SD)	Positive Feedback, (50,80)		Control Mean (SD)	Positive Feedback, Top 20%		Control Mean (SD)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Experiment I	0.039	0.046	14.14	-1.591	-1.580	26.26	4.283**	4.333**	24.35	1.930	1.284	26.87
	(1.134)	(1.138)	(12.04)	(1.336)	(1.190)	(7.99)	(2.101)	(2.112)	(11.14)	(1.182)	(1.115)	(10.30)
N	342	342		165	165		67	67		238	238	
Experiment II: Replication	-1.019	-1.129	10.35	-3.386	-3.671	24.17	3.404**	3.105**	24.22	0.802	0.992	28.06
	(0.985)	(0.985)	(11.34)	(5.787)	(5.992)	(10.18)	(1.322)	(1.276)	(10.43)	(1.199)	(1.085)	(10.09)
N	320	320		29	29		202	202		246	246	
Pooled	-0.477	-0.513	12.30	-1.319	-1.455	25.98	3.450***	3.191***	24.26	1.344	1.116	27.48
	(0.751)	(0.752)	(11.84)	(1.299)	(1.186)	(8.29)	(1.132)	(1.086)	(10.57)	(0.836)	(0.771)	(10.19)
N	662	662		194	194		269	269		484	484	
Strata FE	YES	YES		YES	YES		YES	YES		YES	YES	
Ability Controls	YES	YES		YES	YES		YES	YES		YES	YES	
Background Variables	YES	YES		YES	YES		YES	YES		YES	YES	
GPA 1st Semester	NO	YES		NO	YES		NO	YES		NO	YES	

Note: The control group means of 2nd semester performance for different feedbacks can be similar because eligibility for feedback types is based on 1st semester performance. *Outcome variable:* credits 2nd semester; *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 8: RDD ESTIMATES – EFFECT OF POSITIVE VERSUS NEGATIVE FEEDBACK, 1ST ORDER POLYNOMIAL

	Experiment I		Experiment II: Replication		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
	0.25 < Ratio < 1.75	0.5 < Ratio < 1.5	0.25 < Ratio < 1.75	0.5 < Ratio < 1.5	0.25 < Ratio < 1.75	0.5 < Ratio < 1.5
Treatment Group	8.889***	7.509**	7.191***	10.489***	7.016***	8.554***
	(2.454)	(3.304)	(2.229)	(2.899)	(1.596)	(2.145)
N	247	206	303	221	550	427
Control Group	1.928	1.882	0.144	0.413	0.484	0.613
	(2.644)	(3.283)	(2.702)	(3.511)	(1.815)	(2.293)
N	268	226	301	226	569	452
Diff-in-Diff	6.773*	5.571	6.680*	9.858**	6.246**	8.171***
	(3.608)	(4.565)	(3.507)	(4.562)	(2.429)	(3.120)
N	515	432	604	447	1119	879
Study Program FE	YES	YES	YES	YES	YES	YES

Note: Outcome variable: credits 2nd semester; *study program FE:* in the pooled estimation we also include a cohort dummy and its interaction with the study program FE. *Running variable:* ratio of accumulated 1st semester credits (ACP) to the comparison group median (mean) credits (ACP). Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 9: RDD ESTIMATES – EFFECT OF POSITIVE VERSUS NEGATIVE FEEDBACK, DIFFERENT POLYNOMIALS AND DISCONTINUITY SAMPLES

Pooled: Treatment Group

	(1)	(2)	(3)	(4)
	0 < Ratio < 2	0.25 < Ratio < 1.75	0.5 < Ratio < 1.5	0.75 < Ratio < 1.25
1st Order Polynomial	8.230***	7.016***	8.554***	10.096***
	(1.541)	(1.596)	(2.145)	(3.292)
2nd Order Polynomial	7.923***	10.564***	9.022**	16.225**
	(2.569)	(2.755)	(3.693)	(6.278)
3rd Order Polynomial	10.711***	7.963*	12.214**	-1.967
	(3.727)	(4.124)	(5.631)	(13.831)
4th Order Polynomial	10.973**	12.914**	15.443*	-8.142
	(4.956)	(5.672)	(9.305)	(29.994)
N	594	550	427	246
Study Program FE	YES	YES	YES	YES

Pooled: Control Group

	(1)	(2)	(3)	(4)
	0 < Ratio < 2	0.25 < Ratio < 1.75	0.5 < Ratio < 1.5	0.75 < Ratio < 1.25
1st Order Polynomial	1.992	0.484	0.613	-1.847
	(1.693)	(1.815)	(2.293)	(4.152)
2nd Order Polynomial	-1.257	-0.114	-2.675	5.775
	(2.744)	(3.152)	(4.326)	(7.392)
3rd Order Polynomial	0.739	-1.087	6.102	-7.244
	(4.198)	(4.951)	(6.569)	(14.912)
4th Order Polynomial	-0.650	2.182	-3.333	37.970
	(5.826)	(6.789)	(10.206)	(41.316)
N	611	569	452	254
Study Program FE	YES	YES	YES	YES

Note: Outcome variable: credits 2nd semester; *study program FE:* include a cohort dummy and its interaction with the study program FE. *Running variable:* ratio of accumulated 1st semester credits (ACP) to the comparison group median (mean) credits (ACP). Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 10: MEDIAN PERFORMANCE AND TREATMENT EFFECTS

	(1)	(2)	(3)	(4)	(5)
Experiment I					
Ambiguous Feedback	-1.241	-1.313	-1.448	-1.591	-1.580
	(1.394)	(1.404)	(1.295)	(1.336)	(1.190)
N	165	165	165	165	165
Experiment II: Replication					
Positive Feedback	3.430	2.615	2.593	3.126	2.515
	(2.117)	(2.214)	(2.219)	(2.177)	(2.100)
N	80	80	80	80	80
Strata FE	NO	YES	YES	YES	YES
Ability Controls	NO	NO	YES	YES	YES
Background Variables	NO	NO	NO	YES	YES
GPA 1st Semester	NO	NO	NO	NO	YES

Note: The samples include students who obtained the median number of credits. In Experiment I these students received ambiguous feedback. In Experiment II these students received positive feedback (but would have received ambiguous feedback under the design of Experiment I because they performed above the mean but not above the median). *Outcome variable:* credits 2nd semester; *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 11: SURVEY QUESTIONS ON STUDENT EXPECTATIONS IN EXPERIMENT II

First semester	
Credits	*Assume that there are 100 students who have started studying at the same time and are enrolled in the same degree. If you were to rank all 100 students by their credit points (ECTS), such that rank 1 is the student with the highest number of credit points and 100 is the student with the lowest ECTS. In which position do you think you would be?*
GPA	*And in which rank you would be with respect to your grade point average, if rank 1 is the student with the best grade point average and 100 is the student with the worst grade point average?*
Second and third semester	
Credits	*What do you think? How many per cent of your fellow students will have achieved more credit points (ECTS) than you at the end of the current semester?*
GPA	*And in your opinion, how many per cent of your fellow students will have achieved a better overall grade point average than you at the end of the current semester?*

Note: All questions provide the option to give no answer.

Table 12: EFFECT ON CREDITS, BY PRE-TREATMENT EXPECTATIONS – AMONG STUDENTS ELIGIBLE FOR POSITIVE FEEDBACK)

	(1)	(2)	(3)	(4)	(5)
Treatment	2.700*	1.563	2.110	1.877	1.220
	(1.590)	(1.686)	(1.452)	(1.514)	(1.183)
Underestimated Performance		-2.122	-2.495	-2.670	-3.036*
		(2.280)	(1.986)	(1.919)	(1.765)
Treatment*Underestimated		2.542	2.290	2.630	4.061
		(3.182)	(2.975)	(2.751)	(2.523)
Treatment+(Treatment*Underestimated)		4.105	4.400*	4.507*	5.281**
		(2.769)	(2.616)	(2.412)	(2.374)
N	110	110	110	110	110
Strata FE	YES	YES	YES	YES	YES
Ability Controls	NO	NO	YES	YES	YES
Background Variables	NO	NO	NO	YES	YES
GPA 1st Semester	NO	NO	NO	NO	YES

Note: See Table 11 for the survey questions on student expectations. *Outcome variable:* credits 2nd semester; *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 13: EFFECT OF REPEATED TREATMENT ON CREDITS – THIRD SEMESTER

	All Feedbacks					Control Mean (SD)	Positive Feedback					Control Mean (SD)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Experiment I	0.067	0.143	-0.025	-0.012	-0.119	23.35	0.240	0.364	0.202	0.131	-0.110	26.51
	(0.798)	(0.689)	(0.676)	(0.669)	(0.667)	(10.50)	(0.895)	(0.881)	(0.859)	(0.863)	(0.837)	(7.36)
N	713	713	713	713	713		294	294	294	294	294	
Experiment II: Replication	0.418	0.103	-0.077	-0.022	-0.046	21.43	-0.657	-0.587	-0.686	-0.746	-0.730	25.51
	(0.874)	(0.760)	(0.754)	(0.740)	(0.733)	(11.58)	(0.810)	(0.803)	(0.786)	(0.770)	(0.751)	(8.38)
N	661	661	661	661	661		431	431	431	431	431	
Pooled	0.251	0.124	-0.005	0.032	-0.031	22.42	-0.293	-0.202	-0.268	-0.414	-0.495	25.92
	(0.592)	(0.511)	(0.505)	(0.498)	(0.494)	(11.07)	(0.605)	(0.596)	(0.586)	(0.575)	(0.558)	(7.99)
N	1374	1374	1374	1374	1374		725	725	725	725	725	
Strata FE	NO	YES	YES	YES	YES		NO	YES	YES	YES	YES	
Ability Controls	NO	NO	YES	YES	YES		NO	NO	YES	YES	YES	
Background Variables	NO	NO	NO	YES	YES		NO	NO	NO	YES	YES	
GPA 1st Semester	NO	NO	NO	NO	YES		NO	NO	NO	NO	YES	

Note: Students who dropped out before the start of the third semester are excluded. *Outcome variable:* credits 3rd semester; *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 14: EFFECT OF REPEATED TREATMENT ON CREDITS – THIRD SEMESTER, BY FEEDBACK TYPE

	Negative Feedback		Control Mean (SD)	Ambiguous Feedback		Control Mean (SD)	Positive Feedback, (50,80)		Control Mean (SD)	Positive Feedback, Top 20%		Control Mean (SD)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Experiment I	-0.410	-0.407	18.46	0.863	0.701	25.68	2.347	2.466	26.13	-0.398	-0.774	26.62
	(1.343)	(1.345)	(12.74)	(1.244)	(1.202)	(8.00)	(2.064)	(2.004)	(8.29)	(0.977)	(0.964)	(7.13)
N	257	257		162	162		63	63		231	231	
Experiment II: Replication	1.391	1.327	13.39	-1.599	-2.042	21.00	0.302	0.253	25.06	-1.353	-1.308	25.85
	(1.580)	(1.573)	(12.52)	(5.739)	(6.253)	(14.93)	(1.236)	(1.229)	(8.79)	(1.001)	(0.954)	(8.09)
N	202	202		28	28		192	192		239	239	
Pooled	0.439	0.440	16.17	1.181	0.928	25.09	0.569	0.540	25.33	-0.942	-1.095	26.22
	(1.011)	(1.013)	(12.87)	(1.240)	(1.212)	(9.17)	(1.047)	(1.036)	(8.64)	(0.702)	(0.674)	(7.64)
N	459	459		190	190		255	255		470	470	
Strata FE	YES	YES		YES	YES		YES	YES		YES	YES	
Ability Controls	YES	YES		YES	YES		YES	YES		YES	YES	
Background Variables	YES	YES		YES	YES		YES	YES		YES	YES	
GPA 1st Semester	NO	YES		NO	YES		NO	YES		NO	YES	

Note: Students who dropped out before the start of the third semester are excluded. The control group means of 3nd semester performance for different feedbacks can be similar because eligibility for feedback types is based on 1st semester performance. *Outcome variable:* credits 3rd semester; *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 15: LEE BOUNDS OF THE EFFECT ON CREDITS IN THIRD SEMESTER, BY FEEDBACK TYPE

	(1) Negative Feedback	(2) Ambiguous Feedback	(3) Positive Feedback, $(50, 80)$	(4) Positive Feedback, Top 20%
Experiment I				
Lower Bound	-1.267	0.148	1.939	-1.104
Treatment Effect	-1.196	0.990	2.809	-0.467
Upper Bound	-1.068	1.595	3.506	0.656
N	342	165	67	238
Selected N	257	162	63	231
Trimming Proportion	0.0038	0.0337	0.0597	0.0414
Experiment II				
Lower Bound	0.178	-1.086	-0.511	-1.928
Treatment Effect	1.607	0.382	0.101	-1.256
Upper Bound	4.007	2.326	1.260	-0.856
N	320	29	202	246
Selected N	202	28	192	239
Trimming Proportion	0.0964	0.0833	0.0449	0.0262
Strata FE	NO	NO	NO	NO
Ability Controls	NO	NO	NO	NO
Background Variables	NO	NO	NO	NO
GPA 1st Semester	NO	NO	NO	NO

Note: In the OLS estimates we exclude students that dropped out before the start of the third semester. For the Lee Bounds the group that suffers less from sample attrition is trimmed – either from below or from above – at the quantile of the outcome variable that corresponds to the share of excess observations in this group. The lower and upper bound of the treatment effect is calculated from the group differentials of the mean outcome. *Outcome variable:* credits 3rd semester.

Table 16: EFFECTS ON ATTEMPTED AND FAILED EXAMS, SECOND SEMESTER

	All Feedbacks Attempted Exams		Control Mean (SD)	All Feedbacks Failed Exams		Control Mean (SD)	Positive Feedback Attempted Exams		Control Mean (SD)	Positive Feedback Failed Exams		Control Mean (SD)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Experiment I	0.191	0.181	5.02	0.108	0.141	1.12	0.304**	0.280*	5.58	-0.118	-0.061	0.69
	(0.140)	(0.140)	(2.37)	(0.098)	(0.096)	(1.56)	(0.150)	(0.150)	(1.85)	(0.108)	(0.098)	(1.15)
N	812	812		758	758		305	305		300	300	
Experiment II: Replication	-0.115	-0.109	5.08	-0.041	-0.029	1.51	0.106	0.105	5.76	-0.150	-0.148	0.79
	(0.156)	(0.156)	(2.55)	(0.099)	(0.097)	(1.78)	(0.124)	(0.123)	(1.65)	(0.099)	(0.091)	(1.23)
N	797	797		738	738		448	448		443	443	
Pooled	0.044	0.045	5.05	0.033	0.054	1.31	0.183*	0.174*	5.69	-0.138*	-0.113*	0.75
	(0.105)	(0.105)	(2.46)	(0.070)	(0.068)	(1.69)	(0.096)	(0.095)	(1.74)	(0.073)	(0.067)	(1.20)
N	1609	1609		1496	1496		753	753		743	743	
Strata FE	YES	YES		YES	YES		YES	YES		YES	YES	
Ability Controls	YES	YES		YES	YES		YES	YES		YES	YES	
Background Variables	YES	YES		YES	YES		YES	YES		YES	YES	
GPA 1st Semester	NO	YES		NO	YES		NO	YES		NO	YES	

Note: Outcome variables: attempted exams 2nd semester and failed exams 2nd Semester; *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

-2cm-2cm

Table 17: EFFECTS ON GPA AND DROPOUT, SECOND SEMESTER

	All Feedbacks GPA		Control Mean (SD)	All Feedbacks Dropout		Control Mean (SD)	Positive Feedback GPA		Control Mean (SD)	Positive Feedback Dropout		Control Mean (SD)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Experiment I	-0.047	-0.016	2.58	-0.009	-0.006	0.13	-0.016	0.020	2.36	-0.038*	-0.035*	0.06
	(0.043)	(0.039)	(0.63)	(0.020)	(0.020)	(0.33)	(0.062)	(0.055)	(0.59)	(0.020)	(0.020)	(0.23)
N	707	707		812	812		295	295		305	305	
Experiment II: Replication	0.015	0.019	2.62	0.017	0.016	0.16	-0.010	-0.015	2.47	-0.006	-0.006	0.04
	(0.044)	(0.038)	(0.64)	(0.023)	(0.023)	(0.37)	(0.053)	(0.042)	(0.64)	(0.018)	(0.018)	(0.20)
N	666	666		797	797		434	434		448	448	
Pooled	-0.018	-0.000	2.60	0.004	0.004	0.15	-0.012	0.003	2.42	-0.018	-0.017	0.05
	(0.031)	(0.027)	(0.64)	(0.015)	(0.015)	(0.35)	(0.040)	(0.033)	(0.62)	(0.013)	(0.013)	(0.21)
N	1373	1373		1609	1609		729	729		753	753	
Strata FE	YES	YES		YES	YES		YES	YES		YES	YES	
Ability Controls	YES	YES		YES	YES		YES	YES		YES	YES	
Background Variables	YES	YES		YES	YES		YES	YES		YES	YES	
GPA 1st Semester	NO	YES		NO	YES		NO	YES		NO	YES	

Note: Outcome variables: GPA 2nd semester, dropout (during or before the 2nd semester); *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st Semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 18: SURVEY QUESTIONS ON ATTENDANCE AND WELL-BEING, 2ND SEMESTER OF EXPERIMENT I AND II

Question	
1	*Now we would like to ask you about your overall satisfaction with your life: How satisfied are you currently with your life, all things considered?*
	[0 - completely dissatisfied; 10 - completely satisfied]
2	*How frequently did you attend the lectures this semester?*
	[never; rarely; sometimes; often; very often; always]
3	*During the last weeks, how often did you feel stressed out by our studies?*
	[never; rarely; sometimes; often; very often; always]
4	*Please think about the current semester. To what extent do you agree with the following statements about your studies: When thinking about my studies, I think of...*
4.1	*- not having enough time*
4.2	*- interesting lectures and curriculum*
4.3	*- pressure to perform well*
4.4	*- freedom in organizing my studies*
4.5	*- competition among students*
4.6	*- personal development and growth*
	[0 - completely disagree; 7 - completely agree]
5	*Now we would like to ask you about your overall satisfaction with your studies: How satisfied are you currently with your studies, all things considered?*
	[0 - completely dissatisfied; 10 - completely satisfied]
6	*More specifically: How satisfied are you so far with your performance in your studies?*
6.1	*- With my grades, I am...*
6.2	*- With my attained credit points (ECTS), I am...*
	[0 - completely dissatisfied; 10 - completely satisfied]
Estimation Outcomes	
	For the outcomes in Table 19 we ran exploratory factor analyses to see if there are variables that load on a common factor. Afterwards we standardized all survey questions within cohort and study program. In the cases where multiple questions captured the same latent construct, we constructed our outcomes by averaging across the corresponding questions:
Lecture Visits	Question 2
Life Satisfaction	Question 1
Study Satisfaction	Questions 4.2, 4.6, and 5
Performance Satisfaction	Questions 6.1 and 6.2
Study Stress	Questions 3, 4.1, and 4.3

Note: In this table we only include questions on attendance and well-being that were asked in the same way in both experiments. All questions provide the option to give no answer. For this reason the number of observations in Table 19 can vary depending on the outcome of interest.

Table 19: EFFECT ON ATTENDANCE AND WELL-BEING, SECOND SEMESTER

	Lecture Visits		Life Satisfaction		Study Satisfaction		Perfomance Satisfaction		Study Stress	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Experiment I										
Negative Feedback	-0.010	0.023	0.159	0.195	-0.068	-0.043	-0.050	0.044	0.115	0.059
	(0.349)	(0.348)	(0.354)	(0.397)	(0.263)	(0.239)	(0.240)	(0.232)	(0.228)	(0.208)
N	57	57	57	57	58	58	57	57	58	58
Positive Feedback	0.105	0.078	0.181	0.188	0.177	0.216	0.124	0.026	-0.054	-0.041
	(0.167)	(0.174)	(0.164)	(0.168)	(0.140)	(0.144)	(0.112)	(0.086)	(0.150)	(0.159)
N	134	134	134	134	133	133	133	133	134	134
Experiment II: Replication										
Negative Feedback	0.695	-	-0.104	-	0.387	-	0.372	-	-0.182	-
	(0.540)	-	(0.668)	-	(0.462)	-	(0.345)	-	(0.358)	-
N	18	-	18	-	18	-	18	-	17	-
Positive Feedback	-0.039	-0.007	0.190	0.162	0.033	0.014	0.078	-0.033	0.029	0.015
	(0.211)	(0.235)	(0.210)	(0.218)	(0.162)	(0.162)	(0.150)	(0.138)	(0.195)	(0.204)
N	89	89	88	88	87	87	89	89	87	87
Pooled										
Negative Feedback	0.152	0.192	0.099	0.186	0.035	0.081	0.047	0.118	0.057	0.047
	(0.300)	(0.311)	(0.310)	(0.331)	(0.231)	(0.228)	(0.204)	(0.191)	(0.199)	(0.181)
N	75	75	75	75	76	76	75	75	75	75
Positive Feedback	0.051	0.039	0.184	0.155	0.123	0.118	0.107	0.000	-0.023	-0.005
	(0.131)	(0.133)	(0.129)	(0.134)	(0.107)	(0.107)	(0.090)	(0.073)	(0.119)	(0.123)
N	223	223	222	222	220	220	222	222	221	221
Strata FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Ability Controls	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
Background Variables	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES
GPA 1st Semester	NO	YES	NO	YES	NO	YES	NO	YES	NO	YES

Note: In the cases where no estimates are reported the number of observations was to low to run the corresponding specification. *Outcome variables:* see Table 11 for the outcome variables and the corresponding survey questions; *strata FE:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *ability controls:* HS GPA and credits 1st semester; *background variables:* age, female dummy, time since HS degree, and HS degree Abitur dummy. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 20: TRANSITION PROBABILITY MATRICES – MOVEMENT BETWEEN FEEDBACK TYPES FROM SECOND TO THIRD SEMESTER, POOLED SAMPLE

Treatment Group		Feedback Type in 2nd Semester		
		Negative Feedback	Ambiguous Feedback	Positive Feedback
Feedback Type in 3rd Semester	Dropout	32.31	2.83	2.67
	Negative Feedback	62.77	45.28	11.73
	Ambiguous Feedback	0.62	3.77	4.53
	Positive Feedback	4.31	48.11	81.07

Control Group		Feedback Type in 2nd Semester		
		Negative Feedback	Ambiguous Feedback	Positive Feedback
Feedback Type in 3rd Semester	Dropout	29.08	1.14	4.76
	Negative Feedback	63.50	38.64	15.61
	Ambiguous Feedback	1.48	3.41	4.23
	Positive Feedback	5.93	56.82	75.40

Differential		Feedback Type in 2nd Semester		
		Negative Feedback	Ambiguous Feedback	Positive Feedback
Feedback Type in 3rd Semester	Dropout	3.23	1.69	-2.09
	Negative Feedback	-0.73	6.64	-3.88
	Ambiguous Feedback	-0.86	0.36	0.3
	Positive Feedback	-1.62	-8.71	5.67

Note: The top two panels show the probabilities to receive a certain type of feedback in the third semester conditional on the type of feedback provided in the second semester for individuals in the treatment group and those eligible to receive a certain type of feedback in the control group. The bottom panel shows the difference between the transition probabilities of the treatment and the control group. Due to rounding errors in the top two panels columns may not sum up to 100 and in the bottom panel columns may not sum up to 0.

Figure 1: TIMELINE OF ORIGINAL EXPERIMENT AND REPLICATION

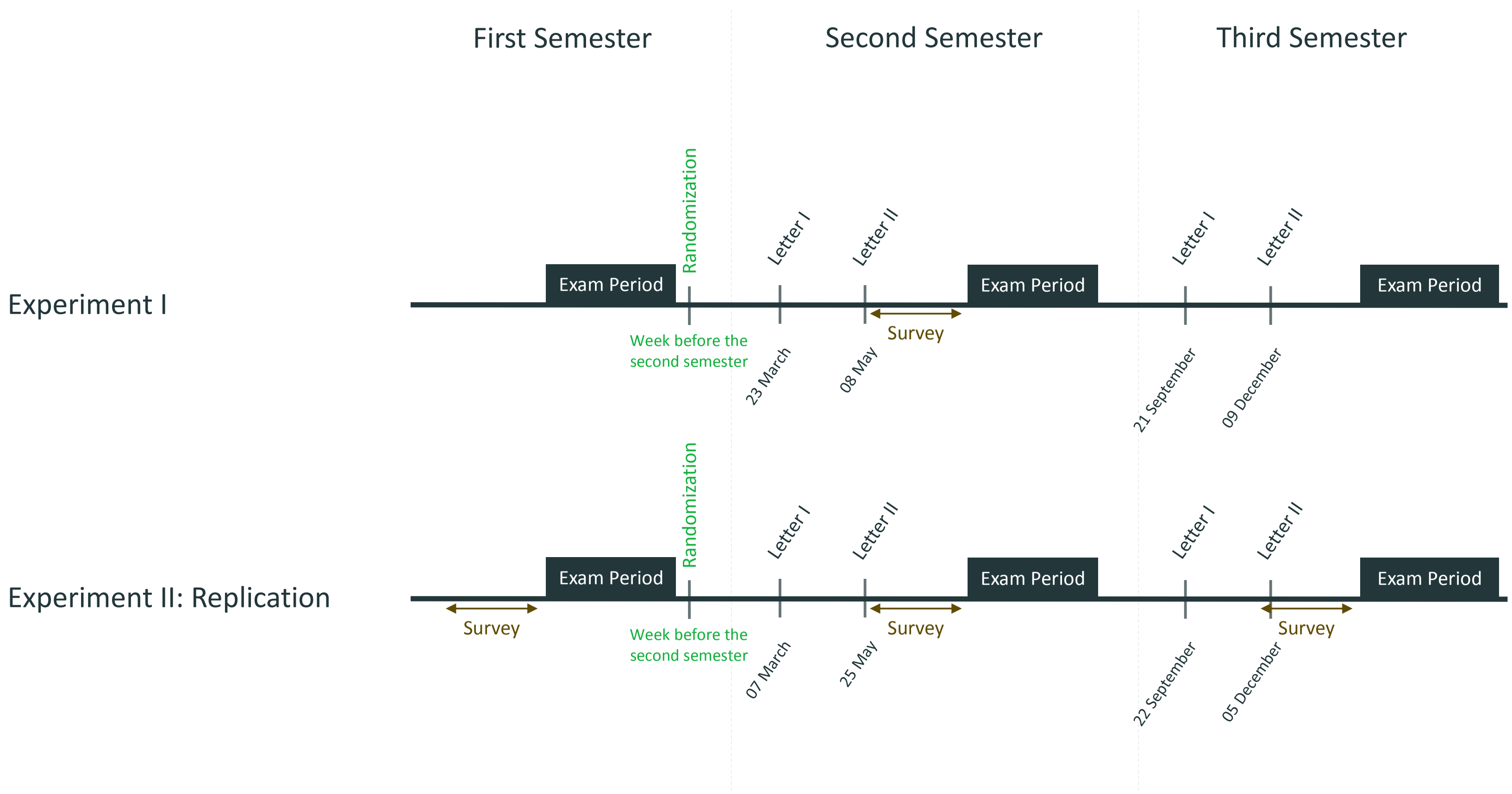
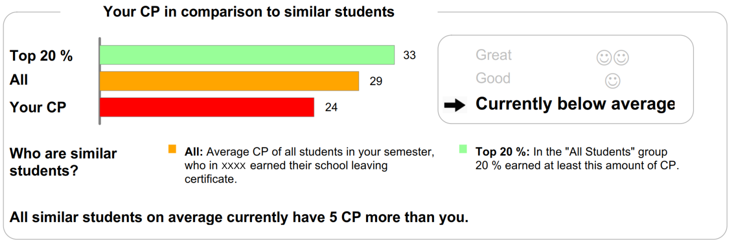
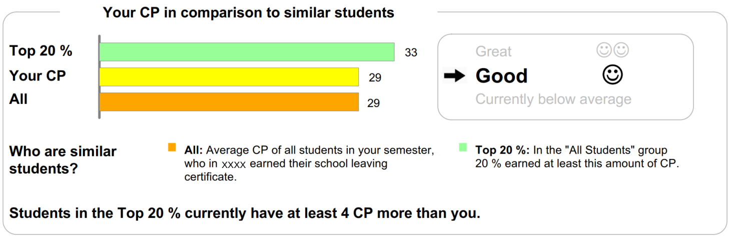


Figure 2: RELATIVE PERFORMANCE GRAPHS – TREATMENT GROUP (EXAMPLES)

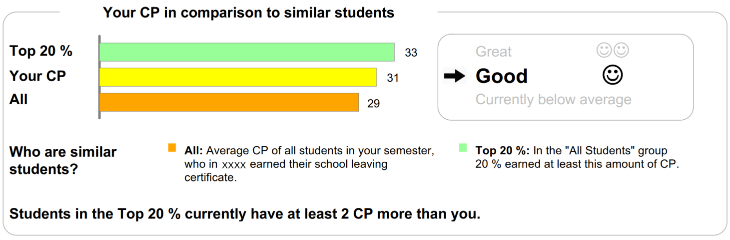
(a) Negative Feedback



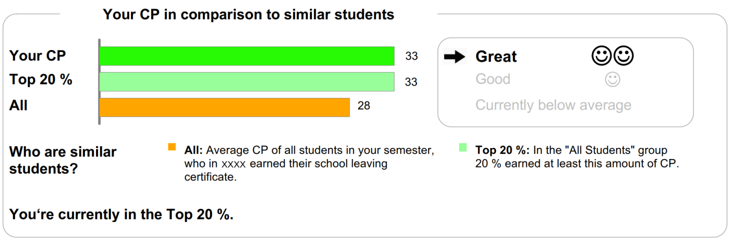
(b) Ambiguous Feedback



(c) Positive Feedback (50,80)



(d) Positive Feedback [80]



(e) Positive Feedback (80,100]

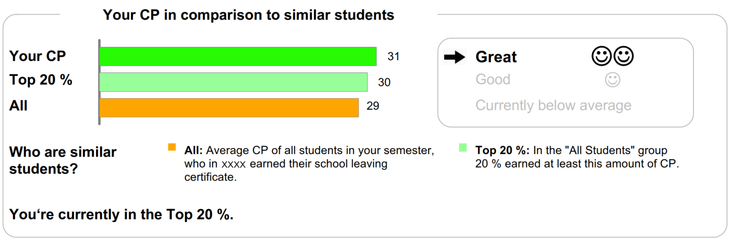


Figure 3: FEEDBACK LETTER I – TREATMENT GROUP

XXX
Postfach ▪ XXX XXX

Faculty of Business Administration

XXX XXX
XXX XXX
Access map at: XXX

Your reference:
Your message from:

Our reference:

Contact:
XXX XXX
xxx.xxx@xxx.de

Room: XXX

07/03/XXXX

Ms/Mr
XXX XXX
XXX XXX
XXX XXX

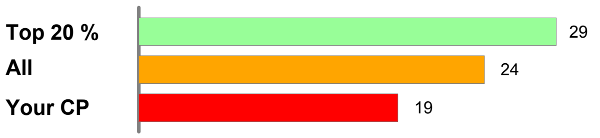
Feedback on your performance in the Bachelor's program International Business

Dear Ms/Mr XXX XXX,

the Department of Business Administration would like to assist you in the further organization and planning of your studies. To this end we provide you with feedback information about your current academic performance. So far you have earned **19 ECTS-Points (CP)** (as of 02/03/XXXX).

In order to allow you a better evaluation of your performance, the following figure compares you to students who are similar to you. Like you, they have been enrolled in International Business (Bachelor) at the XXX since the WS XXXX/XX.

Your CP in comparison to similar students



Great ☺☺
Good ☺
➡ **Currently below average**

Who are similar students?

All: Average CP of all students in your semester, who in XXXX earned their school leaving certificate.

Top 20 %: In the "All Students" group 20 % earned at least this amount of CP.

All similar students on average currently have 5 CP more than you.

Please also keep track of your grades when organizing and planning your studies. Your current grade point average is 2.55 (as of: 02/03/XXXX).

We wish you all the best for your studies and hope that you enjoy the time in XXX.

Yours sincerely

Prof. Dr. XXX XXX, Dean
Faculty of Business Administration

Figure 4: FEEDBACK LETTER I – CONTROL GROUP

XXX
Postfach ▪ XXX XXX

Ms/Mr
XXX XXX
XXX XXX
XXX XXX

Faculty of Business Administration

XXX XXX
XXX XXX
Access map at: XXX

Your reference:
Your message from:

Our reference:

Contact:
XXX XXX
xxx.xxx@xxx.de

Room: XXX

07/03/XXXX

Feedback on your performance in the Bachelor's program International Business

Dear Ms/Mr XXX XXX,

the Department of Business Administration would like to assist you in the further organization and planning of your studies. To this end we provide you with feedback information about your current academic performance. So far you have earned 23 ECTS-Points (CP), and your current grade point average is 3.43 (as of: 02/03/XXXX).

We wish you all the best for your studies and hope that you enjoy the time in XXX.

Yours sincerely

Prof. Dr. XXX XXX, Dean
Faculty of Business Administration

Figure 5: Feedback Letter II – Treatment Group

XXX
Postfach ▪ XXX XXX

Ms/Mr
XXX XXX
XXX XXX
XXX XXX

Faculty of Business Administration

XXX XXX
XXX XXX
Access map at: XXX

Your reference:
Your message from:

Our reference:

Contact:
XXX XXX
xxx.xxx@xxx.de

Room: XXX

25/05/XXXX

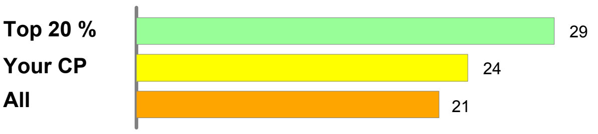
Your performance in the Bachelor's program International Business

Dear Ms/Mr XXX XXX,

The exam period is coming up soon and once more we would like to help you organize your studies. In order to assist you with the planning of this final phase of the semester, we provide you with feedback information about your current academic performance again. So far you have earned **24 ECTS-Points (CP)** (as of 11/05/XXXX).

In order to allow you a better evaluation of your performance, the following figure compares you to students who are similar to you. Like you, they have been enrolled in International Business (Bachelor) at the XXX since the WS XXXX/XX.

Your CP in comparison to similar students



Great ☺☺
➔ **Good** ☺
Currently below average

Who are similar students?

All: Average CP of all students in your semester, who before XXXX earned their school leaving certificate.

Top 20 %: In the "All Students" group 20 % earned at least this amount of CP.

Students in the Top 20 % currently have at least 5 CP more than you.

Please also keep track of your grades when organizing and planning your studies. Your current grade point average is 2.36 (as of: 11/05/XXXX).

We wish you all the best for the upcoming exams.

Yours sincerely

Prof. Dr. XXX XXX, Dean
Faculty of Business Administration

Figure 6: FEEDBACK LETTER II – CONTROL GROUP

XXX
Postfach ▪ XXX XXX

Ms/Mr
XXX XXX
XXX XXX
XXX XXX

Faculty of Business Administration

XXX XXX
XXX XXX
Access map at: XXX

Your reference:
Your message from:

Our reference:

Contact:
XXX XXX
xxx.xxx@xxx.de

Room: XXX

25/05/XXXX

Your performance in the Bachelor's program International Business

Dear Ms/Mr XXX XXX,

The exam period is coming up soon and once more we would like to help you organize your studies. In order to assist you with the planning of this final phase of the semester, we provide you with feedback information about your current academic performance again. So far you have earned 14 ECTS-Points (CP), and your current grade point average is 1.75 (as of: 11/05/XXXX).

We wish you all the best for the upcoming exams.

Yours sincerely

Prof. Dr. XXX XXX, Dean
Faculty of Business Administration

Figure 7: INFO PAGE (LETTER I AND II) – CONTROL AND TREATMENT GROUP

We can advise you on **effective learning**. Please contact XXX XXX (xxx.xxx@xxx.de) for further information.

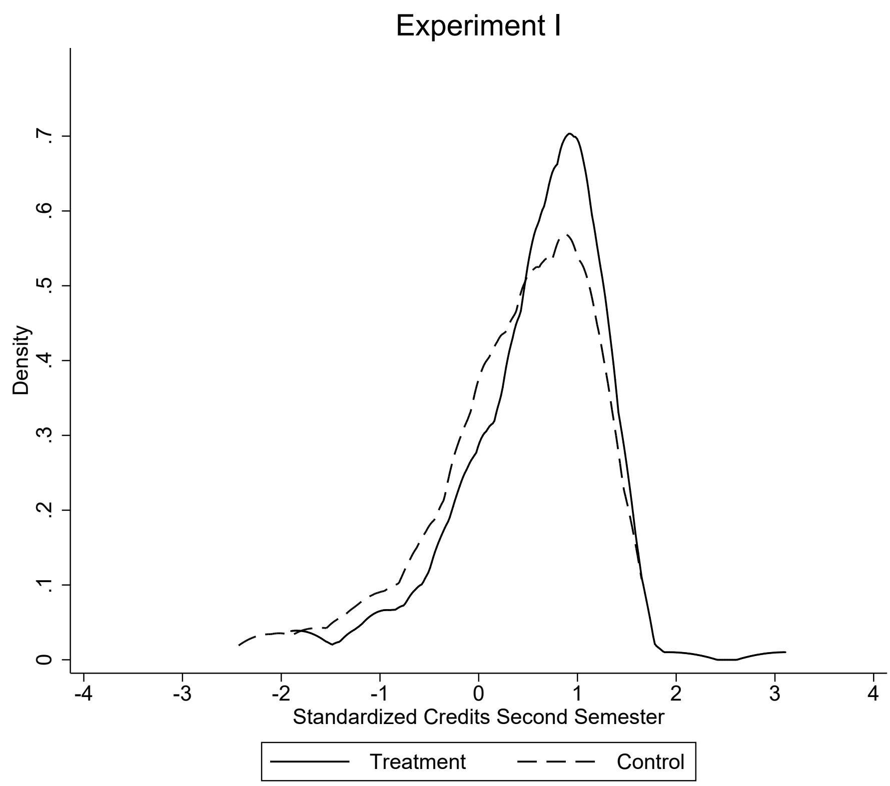
The **Service Lernen** provides further information and interesting programs concerning the broad issue of **learning** at: http://*xxx*

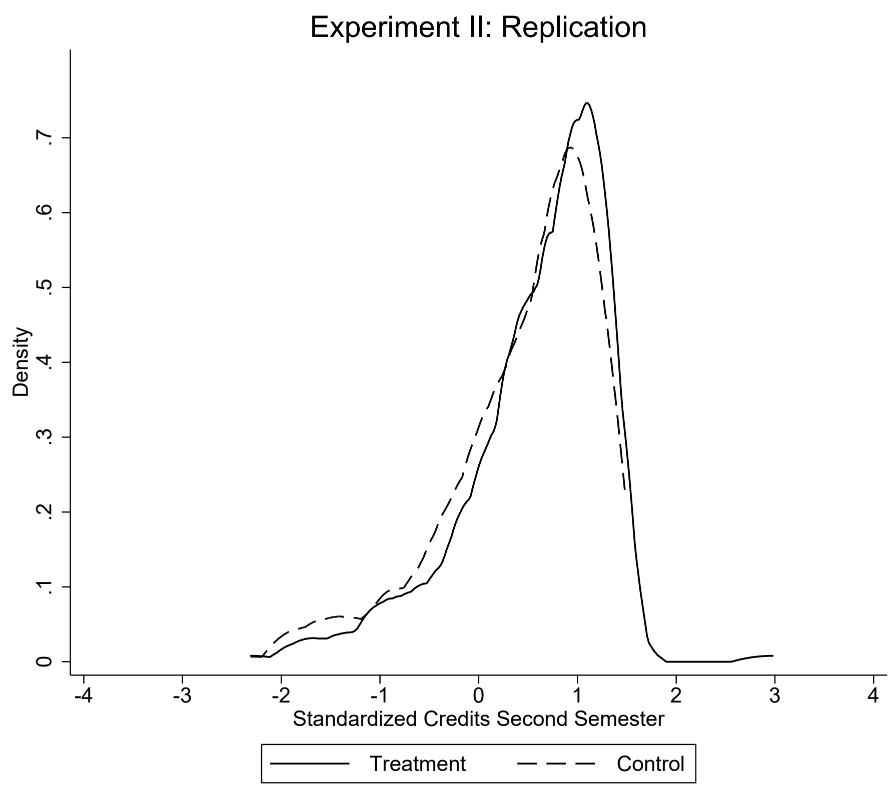
You can also check out the following links for more **information and counseling:**

- If you have general questions about your studies at XXX, please contact the **zentrale Studienberatung**: *http://xxx*
- Mentoring by students in the 3rd or higher semesters is provided by the **Studienberatungsportal**: *http://xxx*
- Find the **Studienfachberater/in** for your specific program of study program at: *http://xxx*
- The **psychologische Studienberatung** provides counseling for personal problems that are rooted in or related to your studies. Contact Prof. Dr. XXX XXX (xxx.xxx@xxx.de) or Prof. Dr. XXX XXX (xxx.xxx@xxx.de). *http://xxx*
- The Studentenwerk also offers **psychological advice**. Contact XXX XXX (xxx/xxx): *http://xxx*
- You can retrieve the **Studienprüfungsordnung** for your specific program of study at the department's web site: *http://xxx*

Please note: The information provided in this letter (CP and grades) is not legally binding. Legally binding information is only provided by the Prüfungsamt.

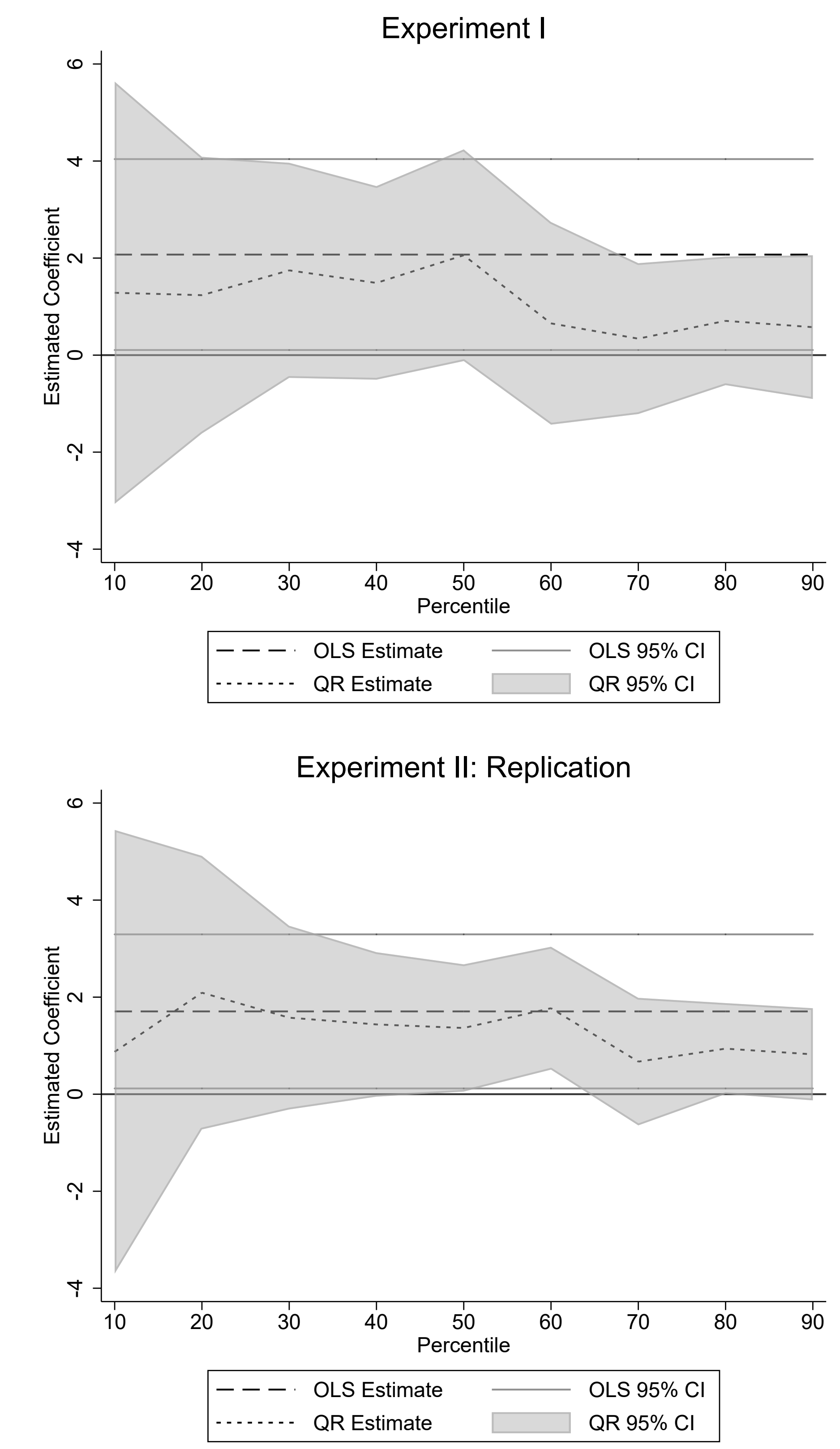
Figure 8: DENSITIES OF STANDARDIZED SECOND SEMESTER CREDITS (POSITIVE FEEDBACK)





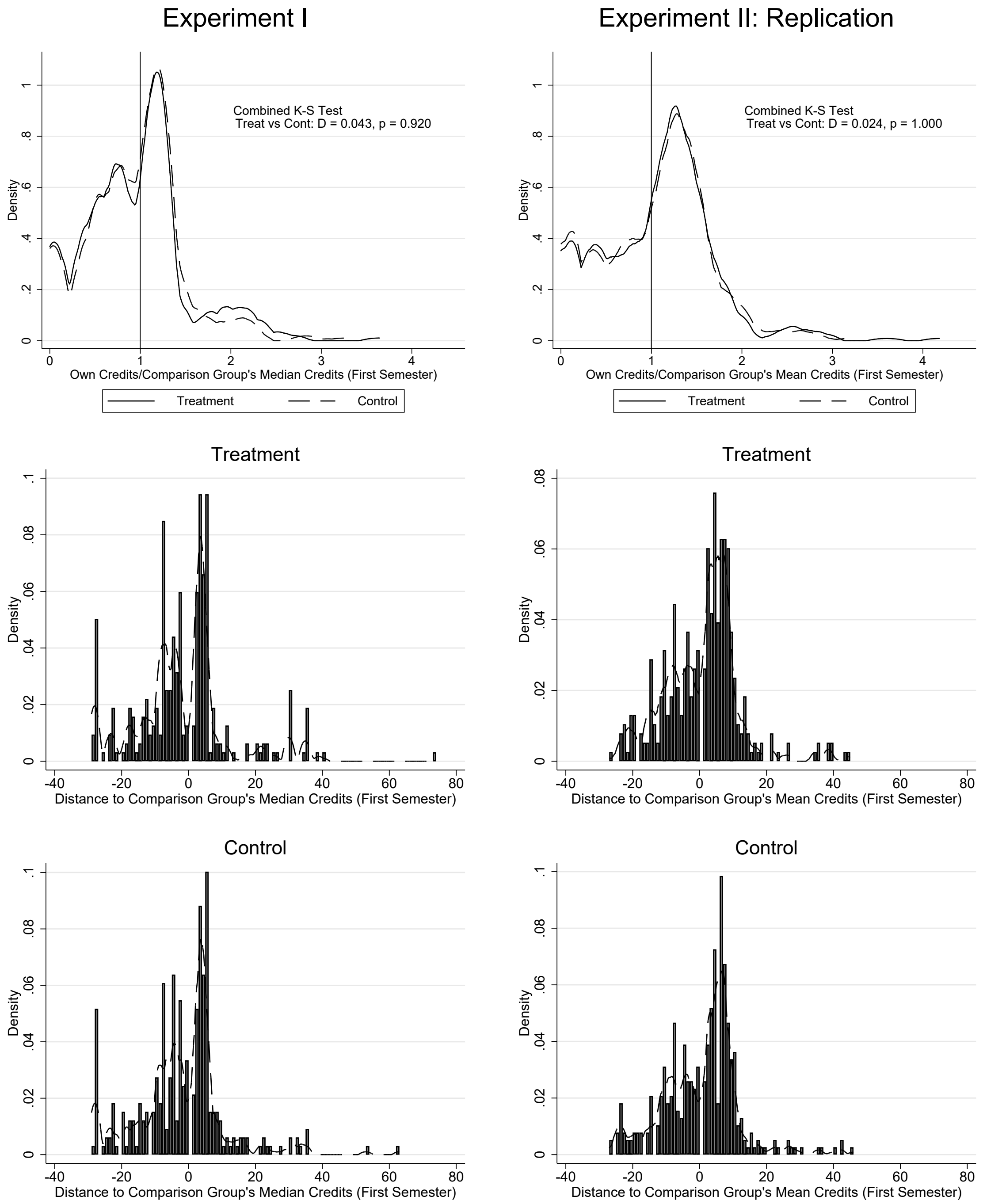
Note: Second semester credits are standardized within study program and cohort.

Figure 9: QUANTILE REGRESSION, EFFECT OF POSITIVE FEEDBACK ON 2ND SEMESTER CREDITS



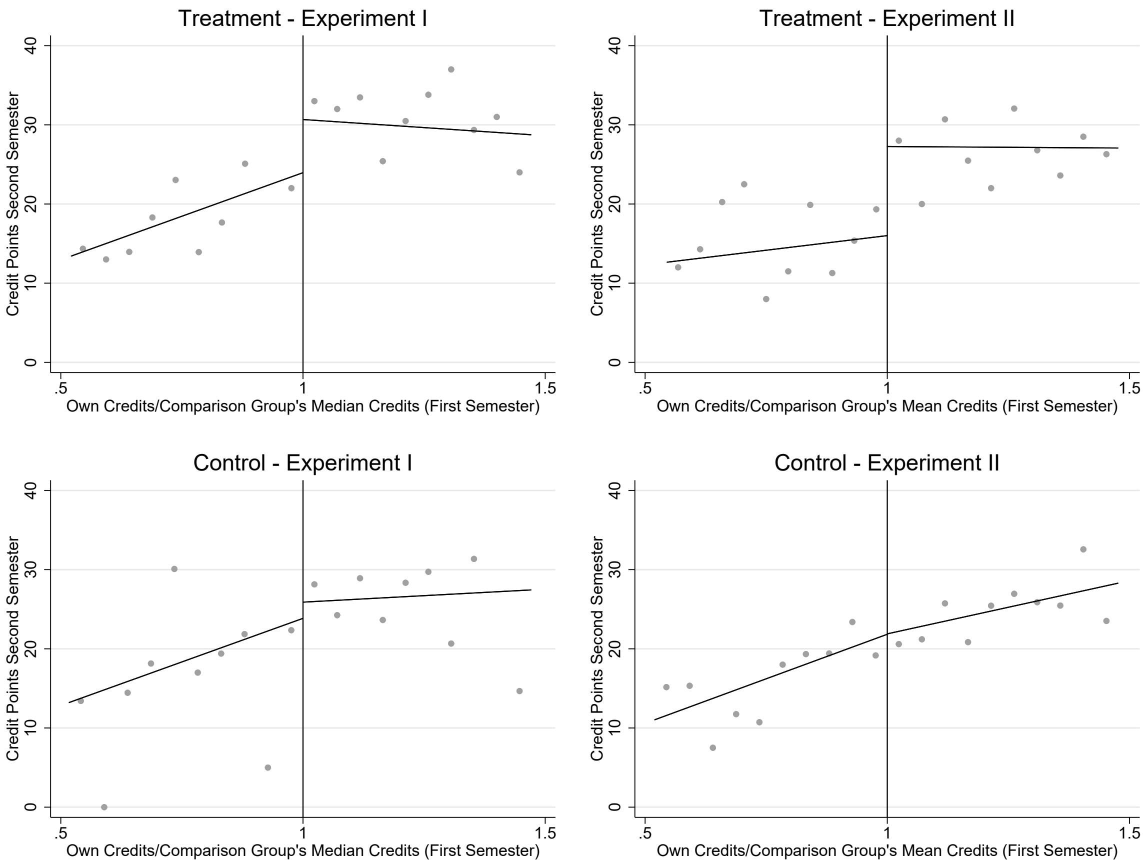
Note: Outcome variable: credits 2nd semester; *control variables* as specified in Equation 5. Standard errors of the simultaneous quantile regressions are bootstrapped using 1000 replications.

Figure 10: DISTRIBUTION OF THE RUNNING VARIABLE



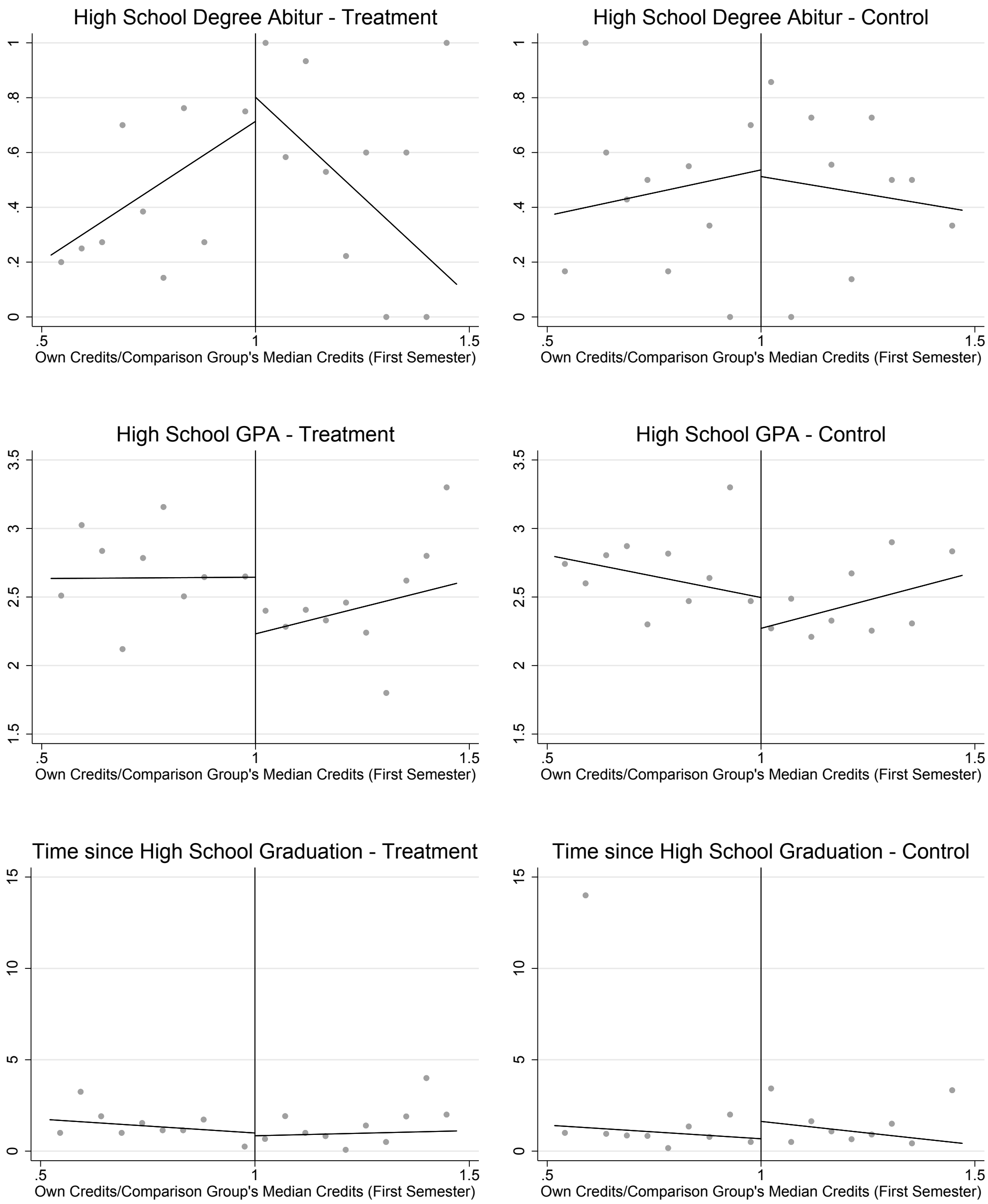
Note: The two top panels show the densities of the running variable used for the RDD. Observations with values lower than 1 received negative, and observations with values above 1 received positive feedback. The four bottom panels show the distribution of the distance to the comparison group's median (mean) in credit points. Observations with negative values received negative, and observations with positive values received positive feedback. Individuals that received ambiguous feedback are excluded in all panels.

Figure 11: REGRESSION DISCONTINUITY PLOT – 1ST ORDER POLYNOMIAL



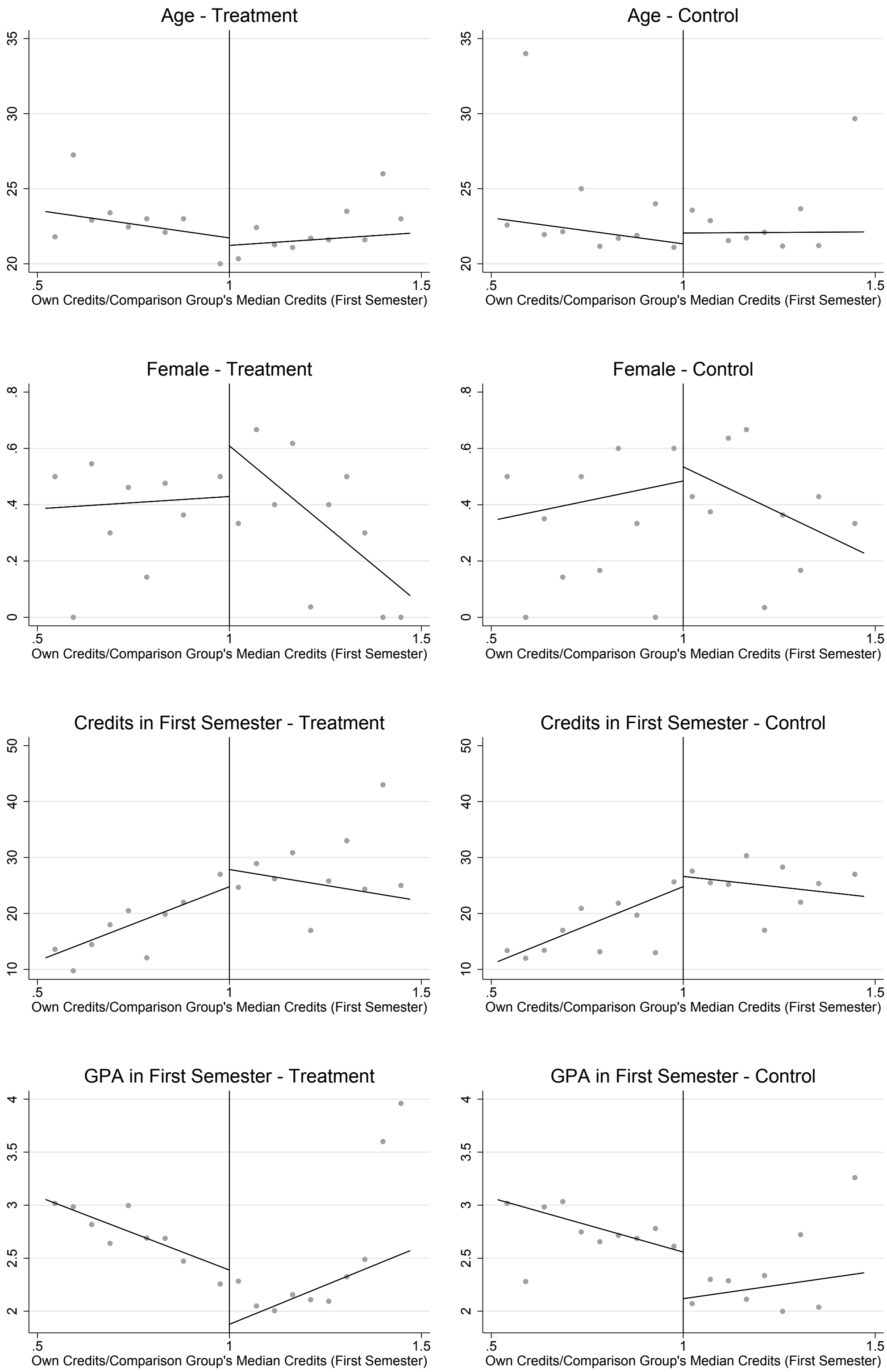
Note: Binned scatterplots using first order polynomials. Observations on the left side of the cut-off received negative feedback or were eligible to receive it in case of the control group. Observations on the right side received positive feedback or were eligible to receive it in case of the control group.

Figure 12: REGRESSION DISCONTINUITY PLOT – 1ST ORDER POLYNOMIAL, COVARIATES EXPERIMENT I



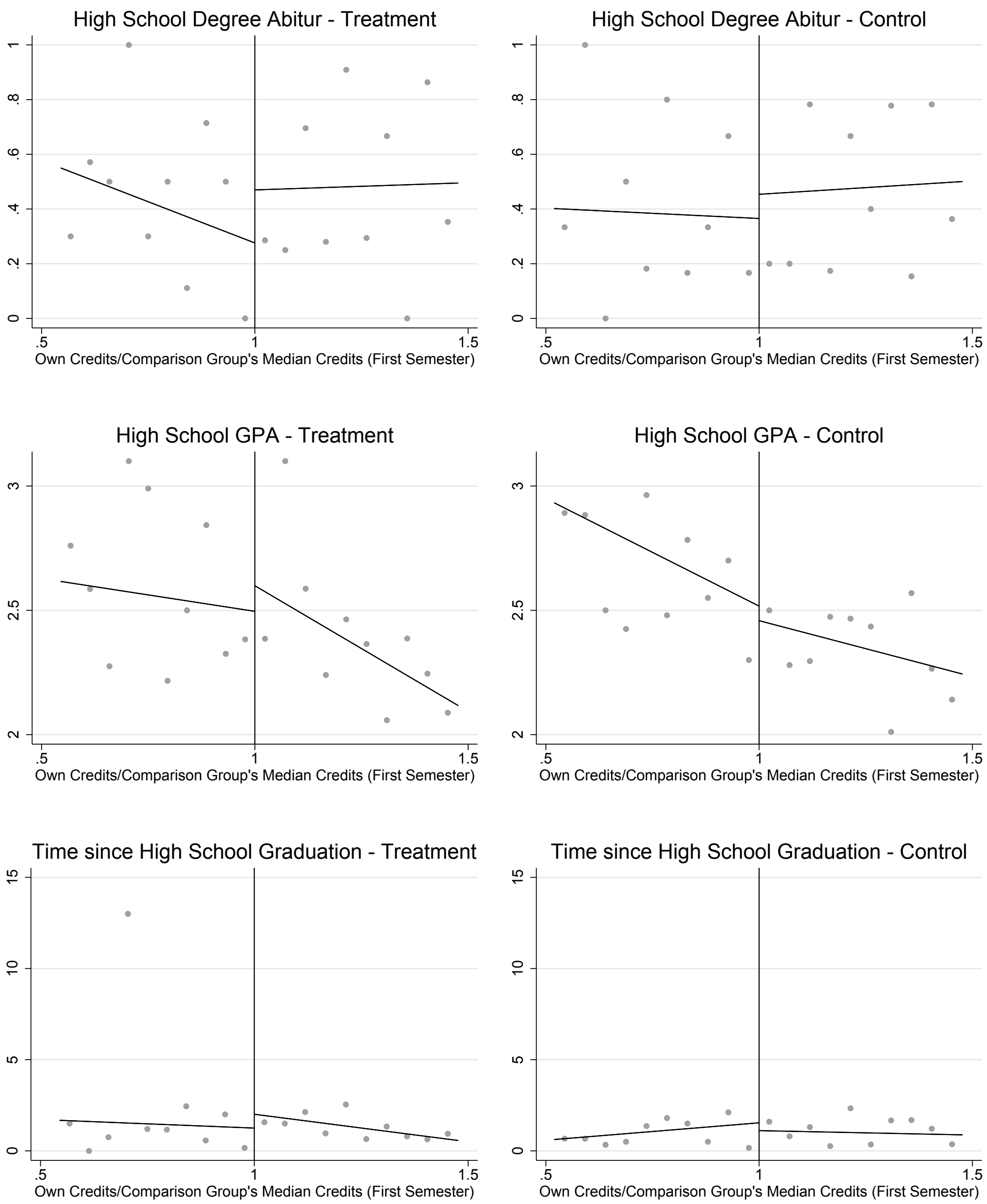
Note: Binned scatterplots using first order polynomials. Observations on the left side of the cut-off received negative feedback or were eligible to receive it in case of the control group. Observations on the right side received positive feedback or were eligible to receive it in case of the control group.

Figure 13: REGRESSION DISCONTINUITY PLOT – 1ST ORDER POLYNOMIAL, COVARIATES EXPERIMENT I (CONT.)



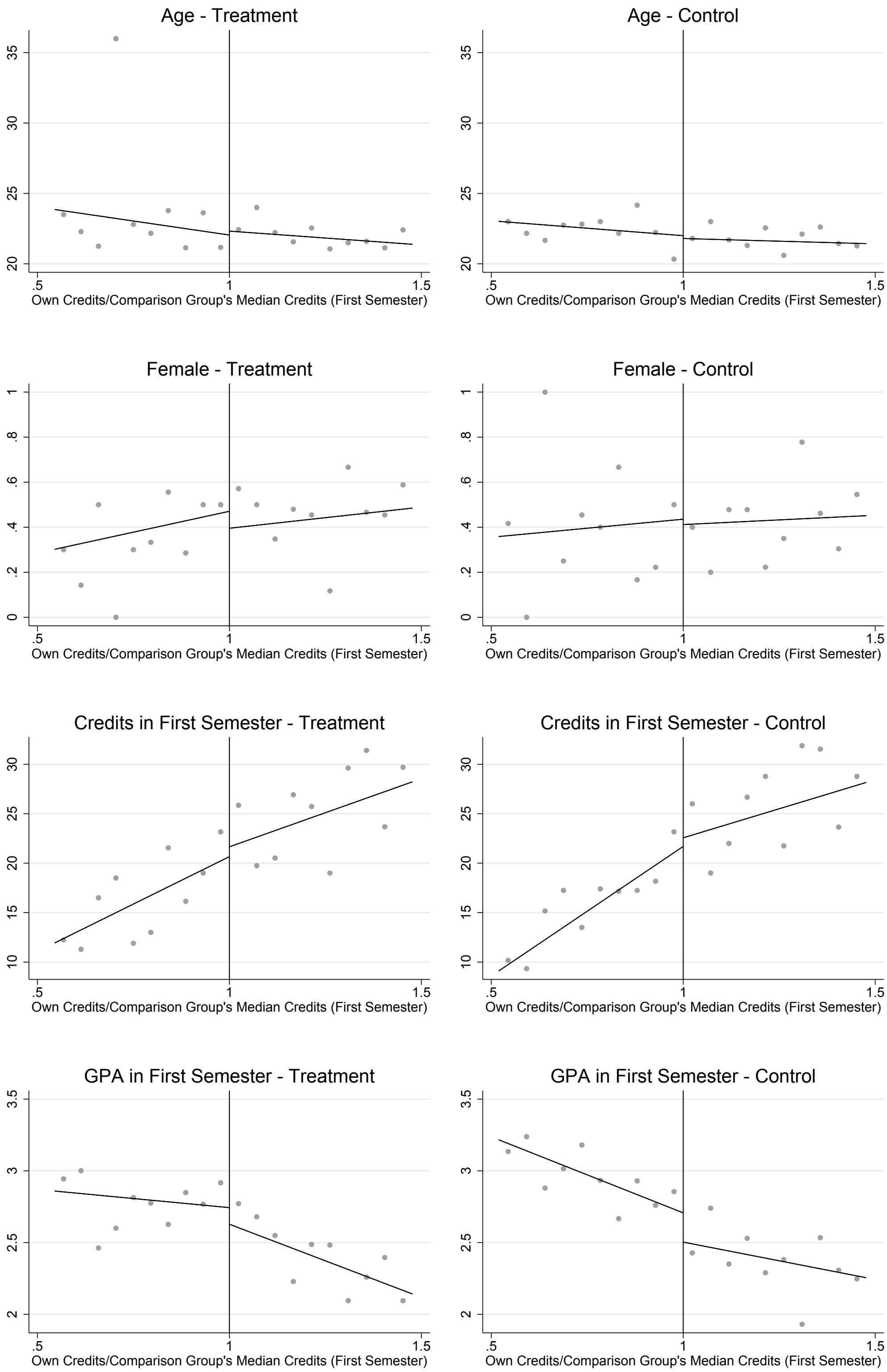
Note: Binned scatterplots using first order polynomials. Observations on the left side of the cut-off received negative feedback or were eligible to receive it in case of the control group. Observations on the right side received positive feedback or were eligible to receive it in case of the control group.

Figure 14: REGRESSION DISCONTINUITY PLOT – 1ST ORDER POLYNOMIAL, COVARIATES EXPERIMENT II: REPLICATION



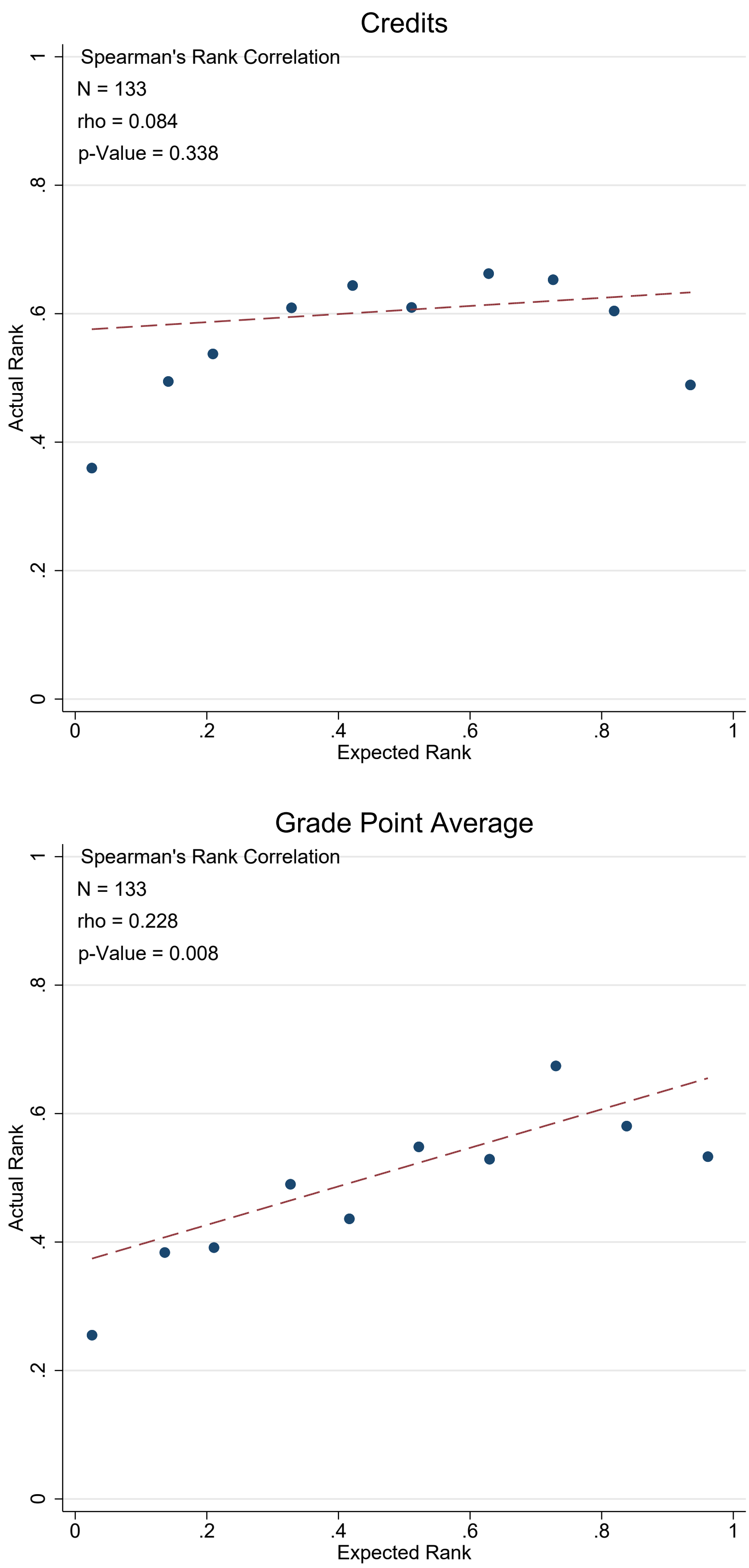
Note: Binned scatterplots using first order polynomials. Observations on the left side of the cut-off received negative feedback or were eligible to receive it in case of the control group. Observations on the right side received positive feedback or were eligible to receive it in case of the control group.

Figure 15: REGRESSION DISCONTINUITY PLOT – 1ST ORDER POLYNOMIAL, COVARIATES EXPERIMENT II: REPLICATION (CONT.)



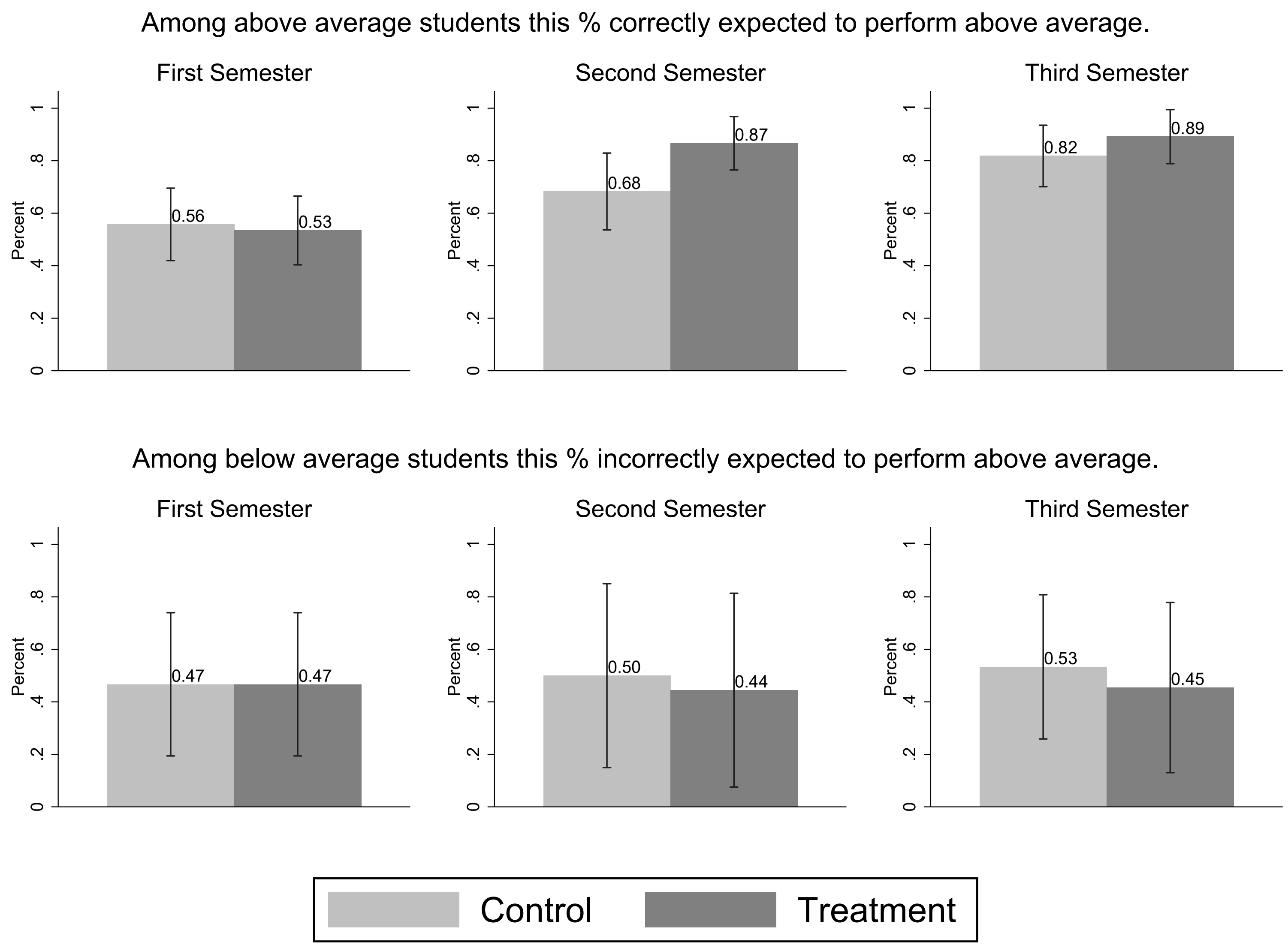
Note: Binned scatterplots using first order polynomials. Observations on the left side of the cut-off received negative feedback or were eligible to receive it in case of the control group. Observations on the right side received positive feedback or were eligible to receive it in case of the control group.

Figure 16: CORRELATION BETWEEN ACTUAL AND EXPECTED RANK IN FIRST SEMESTER, EXPERIMENT II



Note: Binned scatterplots with equally spaced bins and linear fit. The x-axis reports the expected rank, normalized between 0 (lowest rank) and 1 (highest rank) among students who enrolled in the same degree in the second cohort. See Table 11 for the survey questions on students' expectations. The y-axis provides information on the actual rank, normalized in the same way.

Figure 17: SHARES OF STUDENTS WHO EXPECTED TO PERFORM ABOVE AVERAGE, EXPERIMENT II



Note: See Table 11 for the survey questions on student expectations. We assume that expecting to belong to the top 50% (lower 50%) is equivalent to expecting an above (below) average performance. (In-) correct expectations: actual performance and expectations do (not) match. *Number of observations in each panel:* top left N = 110, top middle N = 86, top right N = 81, bottom left N = 30, bottom middle N = 19, and bottom right N = 26.

Data and Methods Appendix

We use anonymized administrative student-level data on performance at the exam-level, accumulated credit points (ACP) and accumulated grade point average (AGPA) at the time the feedback letters are composed, as well as demographic information. In this section we describe how this data was used in the feedback letters and the randomization, how outcome variables and covariates are defined, and how we augment the administrative data with four online surveys.

Feedback Data. We provide feedback on ACP and AGPA in two letters each semester (see Figure 1). For most students the information in the first and in the second letter was identical but for 133 (21) cases of Experiment I (II) the university updated the information on ACP during the first treatment semester.[41] These changes occurred if course results were not yet available at the time the first letter was composed, if grades were changed after students inspected their exams, or due to technical problems at the university.[42] As a consequence, a small number of students received different types of feedback in the two letters. For example, 15 (17) students in Experiment I (II) no longer received positive feedback in the second letter although they did so in the first letter, and 10 (12) no longer received negative feedback in the second letter. We define the feedback types based on the content of the second letter, as its provides the most accurate information and it is more likely to be salient when students start to prepare for their exams.

The university awards credit points and grades on a module level. Modules can consist of a single course or of several courses (sub-modules), all of which must be passed to complete the entire module. Official module-level grades are based on the credit-weighted grades of the courses which make up a module. To compute the AGPA the university sums up the product of the grades and credit points of all modules and divides by the ACP.[43] Failing grades (i.e. grade "5") do not enter into the AGPA. It is important to note that the university only considers completed modules for the ACP and the AGPA.[44] We refer to the university's approach of accumulating performance measures as aggregation on the *module-level*.

Students can access their personal ACP and AGPA online on a website maintained by the university. To ensure consistency between the official information students see online and the feedback information, we provided the ACP and AGPA in the letters. Although a feedback at the *course-level* would more accurately reflect performance (i.e., by including partly completed multiple-course-modules in the feedback), this was not an option. The reason is that this would have led to conflicting

[41] Updates during the semester also occur on a similar scale with respect to the AGPA. The changes in the ACP and AGPA do not necessarily coincide. The reasons for this are explained below.

[42] When the first letter of Experiment I was sent the university had not yet calculated the AGPA information for the study program Business Engineering (N=61).

[43] For the GPAs in the Business Administration program the university assigns double weights to every module that is scheduled after the first year.

[44] This procedure is in effect across all faculties for the AGPA, but not for the ACP. When calculating the ACP, the technical faculty also takes sub-modules into account, while the business faculty only counts completed modules. To give an example, the module Physical Basics of the study program Mechanical Engineering (technical faculty) consists of the sub-modules Physics (4 CP) and Electrical Engineering (3 CP). If a student only passes Physics (but not Electrical Engineering) in the first semester the ACP printed in the letter will include an additional 4 CP, however, her/his AGPA will not contain the Physics grade.

numbers between the official information on the web and the letters we send.

The university counts zero ACP and a missing AGPA if (i) students did not participate in an exam yet, (ii) students took exams but did not pass any of these, and (iii) students passed only sub-modules but did not yet complete a module. For example, at the beginning of the second semester 64 (79) students had zero ACP in Experiment I (II) and 210 (89) students had a missing AGPA.[45] In these cases, we replaced the AGPA with an asterisk which refers to a footnote stating that "Due to technical reasons the grade point average is currently not available. Individual grades can be checked on [the online study platform of the university]".

Randomization Data. In both cohorts, randomization was carried out in the week before the second semester started using demographic information and individual ACP and AGPA at the time of randomization.

We stratified on study program and ACP, and performed rerandomization (Morgan and Rubin 2012) based on AGPA, age, sex, high school grade, time since high school graduation, and (in Experiment II) type of high school degree. In Experiment I we defined five ACP strata for every study program ($ACP \leq 12$, $12 < ACP \leq 18$, $18 < ACP \leq 24$, $24 < ACP \leq 30$, $ACP > 30$) and rerandomized only in the largest bachelor's program (Business Administration (BA)). In Experiment II we defined ACP strata based on quantiles (Q); four ACP strata in the larger study programs BA and Mechanical Engineering ($ACP < Q_{0.25}$, $Q_{0.25} \leq ACP < Q_{0.5}$, $Q_{0.5} \leq ACP < Q_{0.75}$, $ACP \geq Q_{0.75}$) and two ACP strata in the other study programs ($ACP < Q_{0.5}$, $Q_{0.5} \leq ACP$). This approach allowed us to rerandomize in all study programs in Experiment II. For the randomization in Experiment II, we filled missing values on the variables high school GPA (N=30) and AGPA (N=89) with a constant in order to avoid losing units in the randomization and to balance on the full sample.[46] Tables 4 and 5 shows that missing data on both variables are balanced across the treatment and the control group.

Outcome Variables and Covariates. For the analysis in Section 4 we used student performance on the semester-exam-level, i.e. at the course level. We used credit points net of credits granted for internships, the number of attempted and failed exams (conditional on attempting at least one exam), dropout, student grades (excluding failing grades), and survey variables on students' well being as outcome variables.[47] In contrast to ACP and AGPA, the credit points, the number of attempted and failed exams, and grades are now measured on the course-level, i.e., if students only partly completed a multiple-course-module we still included the passed and failed sub-modules in our analyzes. Not only do these outcomes provide more accurate information on the students' performance in each

[45] As already stated in footnote 44, cumulative figures are calculated differently across faculties. This results in unequal numbers of zeros and missing values on the ACP and AGPA variables. Furthermore, the university had not yet calculated the AGPA information for the study program Business Engineering (N=61) when the first letter of Experiment I was composed. Thus there are 61 missing values on the AGPA variable for the study program Business Engineering, which means that there are more missing values on the AGPA variable than there are zeros on the ACP variable in Experiment I (see footnote 42).

[46] In Experiment I we balanced only for the study program Business Administration and only for observations without missing values.

[47] Internships are scheduled later in the study program (4th/5th semester). Some students are awarded these credits at the start of their studies because they completed an apprenticeship and have work experience. As we are interested in the effect of treatment on academic performance, we do not count these internship credits.

semester, but using the ACP and the AGP as outcome variables could also result in an overstated treatment coefficient.[48]

In the regressions we included stratification fixed effects (study program dummies, ACP strata dummies and a binary variable indicating Experiment I or II for pooled estimations), balancing variables (age, sex, high school grades,[49] and time since high school graduation), and further control variables (type of HS degree, sub-module-level credits of the first semester) as covariates. To keep the number of observations constant across specifications we do not include the AGPA at randomization (210 (89) missing values in Experiment I (II)) in the vector of balancing variables. Instead, in the specifications using further control variables we complement the vector of ability controls by adding the individual GPA on the course-level (= baseline GPA in Equation 5). The course-level GPA still has missing values for students who attempted no exams or failed all exams they attempted (55 (66) in Experiment I (II) in the overall sample; 1 (0) missing values for individuals who were eligible to receive positive feedback). We therefore predict the GPA of these students by running linear regressions of the first semester GPA on study program fixed effects, age dummies, gender, time since high school graduation, type of high school degree, and high school GPA to impute these missing values. The imputation allows us to keep the sample size comparable across estimations with and without further control variables.

Survey Data. We also use data from four online surveys. They were conducted in the second half of the semesters, approximately at the time when we usually sent the second letter (see Figure 1 and Table 21). Three of the surveys were carried out after the treatment but in Experiment II we also conducted an additional survey prior to the treatment. The questionnaires included questions on outcome variables such as: how often students attend the lectures, how satisfied they are with their life, the degree to which they are satisfied with their study program, the degree to which they are satisfied with their performance, and how stressful they find their studies (see Table 18 for the survey questions and the variables used in the estimations). We only considered questions as potential outcomes of interest if they were asked the same way in the surveys of both experiments. Because some questions cover similar topics and to reduce the number of outcomes we ran exploratory factor analyses to see which questions load on a common factor. We then standardized all survey questions within cohort and course of study and in the cases where multiple questions captured the same latent construct, we constructed our outcomes by averaging across the corresponding questions. Furthermore, in Experiment II we also gathered pre- and post-treatment information on students' expectations about their relative performance (Table 11).

---

[48] The upward bias occurs when a module consists of several courses which are taken in different semesters. To calculate the ACP and the AGPA the university records the credits and grades awarded for a module in the semester in which the last sub-module has been passed. Let's consider two sub-modules each worth five credits that constitute a composite module running over the first and second semester. Now compare two otherwise identical students – one in treatment and in the control group – both have already passed the first sub-module. If we assume that the feedback causes the treatment student to pass the second exam, the treatment effect in the cumulative data would be 10 CP. However, the actual performance difference between the two individuals in the treatment semester is only five credits.

[49] For some students, the university has no information on high school GPA. We therefore predict 11 (15) missing values on high school grades in Experiment I (II) from a linear regression of the HS GPA on study program fixed effects, age dummies, gender, time since high school graduation and type of high school degree.

Table 21: SURVEY TIME LINE AND PARTICIPATION

	Experiment I		Experiment II: Replication	
	Dates	Participation	Dates	Participation
First Semester	-	-	23/11 - 08/12	20.5%
Second Semester	25/06 - 09/07	31.7%	25/05 - 04/06	15.1%
Third Semester	-	-	29/11 - 20/12	17.4%

Table 22: DESCRIPTIVE STATISTICS AND BALANCING PROPERTIES, STUDENTS WHO WERE ELIGIBLE TO RECEIVE POSITIVE FEEDBACK AND PARTICIPATED IN 1ST SEMESTER SURVEY, EXPERIMENT II

	(1) Control Group Mean (Std. Deviation)	(2) Treatment Coefficient[a)] (Robust SE)	(3) p-Value
Age	22.019	-0.230	0.667
	(3.196)	(0.534)	
Female	0.404	-0.015	0.848
	(0.495)	(0.076)	
HS Degree Abitur	0.596	-0.131	0.155
	(0.495)	(0.092)	
Time since HS Degree	1.308	-0.240	0.509
	(2.119)	(0.362)	
HS GPA	2.292	-0.029	0.781
	(0.610)	(0.103)	
% HS GPA Imputed[b)]	0.000	0.000	.
	(0.000)	(0.000)	
GPA 1st Semester	2.358	-0.013	0.909
	(0.608)	(0.114)	
% GPA 1st Semester Imputed[b)]	0.000	0.000	.
	(0.000)	(0.000)	
Credits 1st Semester	24.413	1.037	0.299
	(7.227)	(0.994)	
Attempted Exams 1st Semester	5.654	-0.328*	0.098
	(1.235)	(0.196)	
Failed Exams 1st Semester	0.346	-0.125	0.160
	(0.789)	(0.089)	
AGPA at Randomization	2.368	-0.005	0.966
	(0.635)	(0.120)	
% AGPA at Randomization NA[b)]	0.000	0.000	.
	(0.000)	(0.000)	
N		110	

Note: Column (1) presents the unadjusted control group means and standard deviations of the covariates. For details on the variables see Table 1 and the Data and Methods Appendix. *a)* Column (2) presents the estimated coefficients of regressing the covariates on the treatment indicator using Equation 2. Column (3) tests the null hypothesis of no treatment effect. *b)* See the Data and Methods Appendix for details on the missing values and the imputation. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 23: DESCRIPTIVE STATISTICS AND BALANCING PROPERTIES 3RD SEMESTER, EXCLUDING DROPOUTS

	Experiment I			Experiment II: Replication		
	(1) Control Group Mean (Std. Deviation)	(2) Treatment Coefficient[a)] (Robust SE)	(3) p-Value	(4) Control Group Mean (Std. Deviation)	(5) Treatment Coefficient[a)] (Robust SE)	(6) p-Value
Age	22.419	0.036	0.871	22.303	0.132	0.556
	(3.082)	(0.220)		(2.999)	(0.224)	
Female	0.414	-0.010	0.746	0.372	-0.013	0.682
	(0.493)	(0.032)		(0.484)	(0.032)	
HS Degree Abitur	0.422	-0.004	0.907	0.405	0.007	0.858
	(0.495)	(0.035)		(0.492)	(0.037)	
Time since HS Degree	1.258	0.022	0.884	1.180	0.006	0.967
	(2.132)	(0.149)		(1.875)	(0.145)	
HS GPA	2.532	0.019	0.618	2.516	-0.042	0.276
	(0.557)	(0.038)		(0.613)	(0.038)	
% HS GPA Imputed[b)]	0.011	0.003	0.698	0.009	0.004	0.593
	(0.106)	(0.008)		(0.095)	(0.008)	
GPA 1st Semester	2.457	-0.046	0.288	2.563	-0.022	0.619
	(0.622)	(0.043)		(0.644)	(0.043)	
% GPA 1st Semester Imputed[b)]	0.023	-0.007	0.500	0.042	-0.020*	0.094
	(0.149)	(0.010)		(0.201)	(0.012)	
Credits 1st Semester	21.827	0.566*	0.076	20.727	0.335	0.368
	(9.153)	(0.318)		(10.233)	(0.372)	
Attempted Exams 1st Semester	4.912	0.034	0.641	5.015	0.062	0.481
	(1.325)	(0.073)		(1.451)	(0.088)	
Failed Exams 1st Semester[c)]	0.626	0.070	0.163	0.863	-0.028	0.647
	(0.963)	(0.050)		(1.183)	(0.061)	
AGPA at Randomization	2.462	-0.039	0.487	2.549	-0.012	0.788
	(0.704)	(0.055)		(0.670)	(0.046)	
% AGPA at Randomization NA[b)]	0.204	-0.018	0.290	0.066	-0.021	0.170
	(0.404)	(0.017)		(0.249)	(0.015)	
N		713			661	

Note: Columns (1) and (4) present the unadjusted control group means and standard deviations of the covariates. For details on the variables see Table 1 and the Data and Methods Appendix. *a)* Columns (2) and (5) present the estimated coefficients of regressing the covariates on the treatment indicator using Equation 2. Columns (3) and (6) test the null hypothesis of no treatment effect. *b)* See the Data and Methods Appendix for details on the missing values and the imputation. *c)* Failed Exams excludes students who attempted no exams (8 in Experiment I and 8 in Experiment II). * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 24: Descriptive Statistics and Balancing Properties 3rd semester, Students eligible for Positive Feedback, excluding dropouts

	Experiment I			Experiment II: Replication		
	(1) Control Group Mean (Std. Deviation)	(2) Treatment Coefficient[a)] (Robust SE)	(3) p-Value	(4) Control Group Mean (Std. Deviation)	(5) Treatment Coefficient[a)] (Robust SE)	(6) p-Value
Age	22.575	-0.382	0.219	22.089	-0.106	0.671
	(2.929)	(0.310)		(2.841)	(0.249)	
Female	0.411	-0.031	0.532	0.355	-0.001	0.969
	(0.494)	(0.050)		(0.480)	(0.039)	
HS Degree Abitur	0.432	0.073	0.172	0.411	0.014	0.753
	(0.497)	(0.053)		(0.493)	(0.046)	
Time since HS Degree	1.466	-0.236	0.336	1.061	-0.043	0.780
	(2.489)	(0.244)		(1.656)	(0.152)	
HS GPA	2.442	-0.016	0.791	2.379	-0.039	0.422
	(0.518)	(0.060)		(0.587)	(0.049)	
% HS GPA Imputed[b)]	0.014	-0.002	0.830	0.005	-0.005	0.321
	(0.117)	(0.011)		(0.068)	(0.005)	
GPA 1st Semester	2.234	-0.080	0.228	2.371	-0.014	0.792
	(0.586)	(0.066)		(0.603)	(0.054)	
% GPA 1st Semester Imputed[b)]	0.007	-0.007	0.320	0.000	0.005	0.321
	(0.083)	(0.007)		(0.000)	(0.005)	
Credits 1st Semester	26.288	0.933	0.137	25.624	0.085	0.849
	(8.070)	(0.625)		(7.731)	(0.444)	
Attempted Exams 1st Semester	5.295	-0.096	0.348	5.407	-0.026	0.787
	(1.249)	(0.102)		(1.174)	(0.095)	
Failed Exams 1st Semester[c)]	0.123	0.030	0.498	0.313	-0.073	0.170
	(0.369)	(0.044)		(0.725)	(0.053)	
AGPA at Randomization	2.266	-0.021	0.801	2.381	-0.012	0.829
	(0.703)	(0.082)		(0.627)	(0.056)	
% AGPA at Randomization NA[b)]	0.068	-0.004	0.814	0.000	0.008	0.167
	(0.253)	(0.017)		(0.000)	(0.006)	
N		294			431	

Note: Columns (1) and (4) present the unadjusted control group means and standard deviations of the covariates. For details on the variables see Table 1 and the Data and Methods Appendix. *a)* Columns (2) and (5) present the estimated coefficients of regressing the covariates on the treatment indicator using Equation 2. Columns (3) and (6) test the null hypothesis of no treatment effect. *b)* See the Data and Methods Appendix for details on the missing values and the imputation. *c)* Failed Exams excludes students who attempted no exams (1 in Experiment II). * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 25: DESCRIPTIVE STATISTICS AND BALANCING PROPERTIES, STUDENTS WHO PARTICIPATED IN 2ND SEMESTER SURVEY

	Experiment I			Experiment II: Replication		
	(1) Control Group Mean (Std. Deviation)	(2) Treatment Coefficient[a)] (Robust SE)	(3) p-Value	(4) Control Group Mean (Std. Deviation)	(5) Treatment Coefficient[a)] (Robust SE)	(6) p-Value
Age	22.675	-0.321	0.436	22.466	-0.886	0.136
	(3.480)	(0.412)		(3.465)	(0.590)	
Female	0.421	0.019	0.717	0.534	0.021	0.810
	(0.496)	(0.053)		(0.503)	(0.086)	
HS Degree Abitur	0.357	0.054	0.357	0.534	-0.071	0.454
	(0.481)	(0.058)		(0.503)	(0.094)	
Time since HS Degree	1.341	-0.055	0.847	1.328	-0.520	0.131
	(2.549)	(0.285)		(1.923)	(0.342)	
HS GPA	2.433	-0.014	0.825	2.243	0.036	0.711
	(0.509)	(0.065)		(0.584)	(0.097)	
% HS GPA Imputed[b)]	0.008	0.010	0.493	0.000	0.000	.
	(0.089)	(0.015)		(0.000)	(0.000)	
GPA 1st Semester	2.353	-0.117*	0.094	2.419	-0.139	0.187
	(0.618)	(0.070)		(0.636)	(0.105)	
% GPA 1st Semester Imputed[b)]	0.016	-0.007	0.591	0.000	0.000	.
	(0.125)	(0.013)		(0.000)	(0.000)	
Credits 1st Semester	24.083	-0.003	0.996	24.284	1.641	0.126
	(8.614)	(0.592)		(8.615)	(1.063)	
Attempted Exams 1st Semester	5.127	0.131	0.190	5.328	0.169	0.329
	(1.200)	(0.100)		(1.114)	(0.172)	
Failed Exams 1st Semester	0.452	0.144*	0.065	0.534	-0.132	0.292
	(0.891)	(0.078)		(1.030)	(0.125)	
AGPA at Randomization	2.310	-0.114	0.227	2.424	-0.135	0.222
	(0.689)	(0.094)		(0.663)	(0.110)	
% AGPA at Randomization NA[b)]	0.230	-0.034	0.239	0.017	-0.027	0.272
	(0.423)	(0.029)		(0.131)	(0.024)	
N		256			116	

Note: Columns (1) and (4) presents the unadjusted control group means and standard deviations of the covariates. For details on the variables see Table 1 and the Data and Methods Appendix. *a)* Columns (2) and (5) present the estimated coefficients of regressing the covariates on the treatment indicator using Equation 2. Columns (3) and (6) test the null hypothesis of no treatment effect. *b)* See the Data and Methods Appendix for details on the missing values and the imputation. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.