Relative Performance Feedback and the Effects of Being Above Average - Field Experiment and Replication

Brade, Raphael and Himmler, Oliver and Jäckle, Robert

14 April 2022

Relative Performance Feedback and the Effects of Being Above Average – Field Experiment and Replication

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

April 14, 2022

In a randomized field experiment, we give first-year students in higher education feedback on their relative performance and show that the type of feedback matters, as feedback increases performance only if it informs the student that they placed above average in the past. We reproduce the results in a replication experiment and investigate mechanisms: The effects are not driven by above-average students reacting particularly well to feedback due to individual characteristics; rather, the information about being above average makes feedback effective. We present evidence that individuals focus on good news to adjust their beliefs, and that feedback can offset disadvantages faced by individuals who are held back by their own underestimation of relative abilities. Once beliefs between controls and the treated converge, repeated treatment does not add to the effects.

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

JEL Classification: I23, C93

[a]University of Erfurt, Faculty of Economics, Law and Social Sciences, Nordhäuser Strasse 63, 99089 Erfurt, Germany. email: raphael.brade@uni-erfurt.de.

[b]University of Erfurt, Faculty of Economics, Law and Social Sciences, Nordhäuser Strasse 63, 99089 Erfurt, Germany. email: oliver.himmler@uni-erfurt.de.

[c]Nuremberg Institute of Technology, Faculty of Business Administration, Bahnhofstrasse 87, 90402 Nürnberg, Germany. email: robert.jaeckle@th-nuernberg.de.

A previous version of this paper was circulated under the title "Normatively Framed Relative Performance Feedback – Field Experiment and Replication". We gratefully acknowledge financial support from the German Federal Ministry of Education and Research under grant 01PX16003A and 01PX16003B, the Staedtler Stiftung, and administrative support from the Nuremberg Institute of Technology. Declarations of interest: none.

1 Introduction

It has long been established that the behavior and performance of others can provide an important benchmark against which individuals can compare their performance and gauge their abilities (see, e.g., Bandura 1991, Corcoran, Crusius and Mussweiler 2011, Festinger 1954, and Taylor, Wayment and Carrillo 1996). However, in many situations, the information necessary for comparison may be imprecise or incomplete, leaving individuals with no appropriate frame of reference. Under such conditions, providing feedback relative to a suitable peer group may facilitate social comparison, thereby enhancing decision-making, motivation, and ultimately performance.

An economically and socially important setting where feedback may improve outcomes is higher education. Students at the start of their university careers face new tasks which are complex and challenging, and they are surrounded by new peers, making it difficult to assess their relative abilities. Providing relative performance feedback early on could therefore present a low-cost and easily scalable tool for universities to help students improve their performance.

To test this, we implement a relative performance feedback intervention at one of the largest universities of applied sciences in Germany. After the first semester, control group students receive letters from the university, informing them about how many credit points they managed to obtain in the previous semester.[1] Students in the treatment group receive the same information, but the letters also inform them about how well, in terms of credits, they performed relative to the average student and the student on the 80th percentile. To make individuals aware of what type of behavior is approved and to prevent potential boomerang effects, i.e., negative effects for high performers that are sometimes implied by social comparison theory, we draw from the social norms literature and add an approving normative message (*Good* or *Great* plus an emoticon) for students who obtained at least the average amount of credits.[2]

Our approach complements the few existing studies that consider the effects of relative

[1] Obtained credits are a standardized measure of academic progress that reflects the number of passed exams, where exams are weighted by the required workload for each class. Europe-wide, universities use a standardized system (European Credit Transfer and Accumulation System, ECTS), under which a full-time academic year consists of 60 credits, with the typical workload for one credit equaling 25-30 study hours. See also `https://ec.europa.eu/education/resources-and-tools/european-credit-transfer-and-accumulation-system-ects_en`, retrieved on March 08, 2022.

[2] Normative frames are popular in social comparison contexts that are not concerned with performance. They are frequently employed in the social norms literature to make individuals aware of what type of behavior is approved. Applications include fostering, e.g., tax compliance (Hallsworth et al., 2017; Slemrod, 2016) or environmentally friendly behavior (Allcott and Rogers, 2014; Costa and Kahn, 2013; Goldstein, Cialdini and Griskevicius, 2008).

feedback on the overall performance of students: While we send students postal letters that include relative feedback on their accumulated credits (i.e., quantity), Azmat et al. (2019) and Cabrera and Cid (2017) supply students with information regarding the quality of their study progress by giving access to their relative grade point average (GPA) via the universities' online services and find null or negative effects on performance. (Dobrescu et al., 2021) give feedback on inputs: they provide real-time ranks during a semester-long online assignment and show that this improves exam performance in the intervention course and generates positive spillovers to other courses.

The initial implementation of our feedback intervention has a positive, but statistically insignificant treatment effect on subsequent performance in the full sample. Based on the design of the feedback, we explore heterogeneous responses. Students whom the feedback informs that they have performed above average in the first semester increase their performance by a statistically significant 2 credits points (.16 standard deviations) in the second semester, relative to controls. For students at or below the average we do not find effects on behavior.

These results thus give rise to the hypothesis that above-average feedback enhances performance. We test this hypothesis with a replication experiment one year later with a new cohort of students and reproduce the full pattern of results. Most importantly, we replicate the result that above-average feedback significantly increases performance (1.6 credit points or .12 standard deviations). Across both experiments the effects are roughly equivalent to one third of an additional passed exam and correspond to an effect size of .14 standard deviations.

While our interventions target credits, we also show that those who benefit from the feedback on credit points maintain the same GPA as the individuals in the control group. This shows that the increase in earned credits can be interpreted as a net performance gain, as students do not buy gains on the (treated) credit points dimension with losses on the grade dimension. In light of research showing that performance incentives on one domain can lead to negative effects on other domains (Altmann, Grunewald and Radbruch, in press; Eriksson, Poulsen and Villeval, 2009), this result is not self-evident. In addition, we also survey students on potential negative side effects of the intervention on their well-being, as for example Celik Katreniak (2018) shows that stronger incentives can come at the cost of increased stress and reduced happiness. We find no such evidence: treated students are no different from controls in any of the well-being domains we observe.

We investigate potential mechanisms behind this causal reaction to above-average feedback: Employing a regression discontinuity design (RDD) based on the sharp cutoff at the average allows us to assess the causal effect of being informed about an above-average per-

formance versus being informed about an average or below-average performance. Receiving above-average feedback increases subsequent performance by about six credits in comparison to receiving another type of feedback. This indicates that it is the information about being above average itself and not any differences in underlying characteristics (ability, motivation, learning technology, etc.) that leads to the increase in performance.[3]

In the replication experiment, we further explore mechanisms by analyzing the role of pre- and post-treatment expectations about relative performance. We find that students have inaccurate expectations about their relative performance pre-treatment and that students who receive above-average feedback subsequently update their beliefs. In the absence of treatment, those whose beliefs underestimate their actual relative performance obtain fewer credits. Feedback is able to offset this disadvantage associated with inaccurate beliefs. There is no evidence of updating for students at or below the average. We argue that the pattern of effects and the evidence that students selectively process feedback is consistent with theories on the management of self-confidence (see, e.g., Bénabou and Tirole 2002).

Finally, we show that repeated treatment in the third semester does not elicit additional performance gains. One reason may be that in the third semester, beliefs about relative performance held by the controls are almost as accurate as those of the treatment group.

Overall, our results show that providing relative performance feedback is beneficial for a large share of individuals entering university. When beliefs are inaccurate and the feedback is encouraging, performance can increase even in complex and challenging tasks such as passing exams in higher education. At the same time, the intervention is also inexpensive at a total cost of less than €2.5 per student and semester (see Table A.1). The results also highlight that feedback schemes can have distributional implications. While the effects may be considered pareto-improving (no negative effects), it is mostly students in the upper middle of the performance distribution who profit. The intervention may thus decrease performance equality in the education system. Inequality-averse policy-makers should take this into account and be aware that for the most precarious students, making their shortcomings salient via feedback may not generate any effects.

Relation to the literature. Our study contributes to the research on the effects of relative performance feedback in higher education.[4] This setting is highly relevant, given that a large

[3] Given the design of our intervention, it is not possible to discern whether a neutral feedback would generate the same results as our feedback, which includes a normative frame. As we discuss in Section 3.1, there is no evidence that the approving normative frame alone is behind the effects we find.

[4] 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, 2018; Goulas and Megalokonomou, 2021; Hermes et al., 2021). However, it is not clear whether these results extend to an adult population and an environment that is much less structured than primary and secondary education.

share of students takes much longer than the prescribed time to obtain their degree (see, e.g., Bound, Lovenheim and Turner 2012).[5] Increasing obtained credits is the only means to achieve a more timely graduation (or graduation at all), and there is still a dearth of low-cost and scalable interventions that can elicit changes in academic performance (see, e.g., Oreopoulos et al. 2022 and Oreopoulos and Petronijevic 2019).

It is thus surprising that research on relative feedback in higher education is still sparse. Most closely related to our paper, Azmat et al. (2019) and Cabrera and Cid (2017) also study the effects of relative performance feedback based on overall performance considering the two relevant outcome dimensions, grades and number of courses passed – a setup which allows to detect both substitution effects across courses and between outcome dimensions. Contrary to the results in this paper, both studies find null or negative effects on performance, which may be related to several differences in the feedback schemes used: i) Rather than on GPA, we give feedback on obtained course credits. While the GPA aims at measuring the quality of students' overall performance, credits track students' progress toward their degree and are thus a quantity-oriented measure of performance.[6] Recent studies suggest that quality- and quantity-based feedback can indeed induce different behavioral responses (Gardner, 2020; Zhang et al., 2021).[7] ii) We provide coarse (two reference levels) instead of precise (total percentile distribution or rank) relative performance information. Coarse feedback has proven effective in secondary education (Azmat and Iriberri, 2010). iii) We combine descriptive information with approving normative frames in order to clearly convey to students that they should be encouraged by an at least average performance and to prevent potential "boomerang effects" for high performers (= falling back to the lower reference level), and iv) students in our experiments receive feedback by personalized physical letters and not via an online service portal of the university.[8]

Although their experimental design is less similar to ours, Dobrescu et al. (2021) conduct

[5] E.g., in OECD countries, only 39% of full-time students graduate within the planned duration of the program (OECD, 2019).

[6] Because faster credit accumulation is a prerequisite for earlier graduation, it can have considerable payoffs: first, for the individual, it reduces forgone income by circumventing a longer study duration – and according to the literature on academic momentum, especially the initial progress in college is positively related to degree completion (see Doyle 2011; Attewell, Heil and Reisel 2012; Attewell and Monaghan 2016); second, from a social perspective, (faster) graduation can be expected to translate into, e.g., longer duration of contributions, in terms of taxes and payments into the social security systems.

[7] In the context of higher education, one possible explanation might be that it is easier for students to produce more of the same output rather than the same output at a higher quality level. While the first means that students need to apply the same learning technology to additional exams, the latter might require a change in learning strategies.

[8] Letters may generate more attention, as students nowadays often receive many emails but few letters from the university. This notion is consistent with interventions which are unable to increase academic performance often using digital formats (e.g., Oreopoulos et al. 2022). Moreover, DellaVigna and Linos (2022) also provide evidence that interventions using physical letters may be more effective than interventions that rely on emails.

the only other feedback intervention in higher education that also produces positive effects on the overall performance of students – in their case this also holds true in the full sample.[9] They provide students with real-time intermediate feedback on their rank during a semester-long computerized (online) assignment. This improves the grade in the intervention course by .21 standard deviations and creates positive spillovers to the overall performance in the following semester. Our setup differs in two important ways. First, the feedback in Dobrescu et al. (2021) can be thought of as directly targeting input factors like effort, while our study targets overall performance, an educational output. This may be contributing to the different pattern of effects: Some work on goal setting (Clark et al., 2020) and financial incentives (Fryer Jr., 2011; Hirshleifer, 2021) suggests that incentivizing inputs instead of outputs may be particularly effective when individuals are present-biased or have little knowledge of the education production function. Second, our feedback is based on the overall performance of all courses taken, which should be readily available in existing administrative data. In contrast, the feedback system in Dobrescu et al. (2021) requires the systematic measurement and analysis of data on intermediate performance in one or even all courses. Therefore, in institutions and courses where there is no concept or infrastructure to collect intermediate performance in place, additional resources are needed to implement this type of feedback system.

Ours and the other papers on relative feedback in higher education also make an important contribution to the general literature on relative performance feedback. First, this research strand focuses on the transition into a new environment, a time when individuals face higher uncertainty about relative ability. This contrasts with existing research from the field, which often introduces relative performance feedback when individuals are already familiar with their tasks and peers (Ashraf, 2022; Barankay, 2012; Blanes i Vidal and Nossol, 2011; Delfgaauw et al., 2013). Second, while university presents a complex working environment with challenging, high-stakes tasks, much of the previous literature relies on real effort tasks in the lab (Azmat and Iriberri, 2016; Charness, Masclet and Villeval, 2014; Eriksson, Poulsen and Villeval, 2009; Gill et al., 2019; Kuhnen and Tymula, 2012) or on rather repetitive tasks in the field (Ashraf, 2022; Bandiera, Barankay and Rasul, 2013; Barankay, 2012; Blanes i Vidal and Nossol, 2011; Delfgaauw et al., 2013).

The results of our study also bear significance for the literature that studies the link between confidence and performance. Our findings are consistent with the idea that individuals try to maintain a positive self-assessment of their abilities by processing positive feedback while discarding negative feedback, which in turn leads to higher confidence in ability and

[9] Kajitani, Morimoto and Suzuki (2020) and Tran and Zeckhauser (2012) provide relative feedback on the performance in a mid-term or practice exam and study effects on performance in the final exam of the intervention course but do not account for potential spillovers to other courses or the overall performance.

motivates individuals to work harder and take beneficial risks (see, e.g., Bénabou and Tirole 2002 and Compte and Postlewaite 2004).[10] Empirical evidence from the lab shows that individuals who receive bad news indeed have little willingness to update their self-concept, whereas people who receive positive information are willing to incorporate the good news in their beliefs (Eil and Rao, 2011; Möbius et al., in press). Our study thus complements the existing literature on motivated beliefs, by showing first tentative evidence for this type of behavior in the context of a relative performance feedback intervention.

Finally, our paper is also related to studies on ordinal rank and academic achievements. Elsner and Isphording (2017) and Murphy and Weinhardt (2020) show that a higher rank in secondary and primary school has positive effects on later outcomes such as the probability of finishing high school and attending college, test scores as well as subject choice. In the context of university, Elsner, Isphording and Zölitz (2021) find that a higher ordinal rank increases performance and affects major choice. This literature thus provides evidence that relative position in the performance distribution matters. Relative feedback can both make rank information more precise as well as make rank more salient and thus lead to behavioral responses. Consistent with our suggested theoretical mechanism, the literature on achievement rank also argues that self-confidence is an important mechanism, and provides evidence for positive effects of rank on, e.g., subject-specific confidence (Murphy and Weinhardt, 2020), perceived relative intelligence (Elsner and Isphording, 2017), and expected grades (Elsner, Isphording and Zölitz, 2021).

The remainder of the paper is structured as follows. Section 2 describes the institutional background, data, and design of our intervention as well as the empirical approach. Section 3 reports on the main results of our two field experiments. In Sections 4 to 7, we explore the drivers and mechanisms behind the main results, investigate effects of repeated treatment, present the effects on auxiliary outcomes, and discuss potential spillovers from treatment to control. Section 8 concludes.

2 Institutional background and research design

We conduct our field experiments at one of the largest universities of applied sciences (UAS) in Germany. Our interventions consist of two cohorts of students who enrolled in five bachelor's degree programs at the faculties of Business Administration (BuA) and Mechanical Engineering (ME). All interventions were implemented and outcomes realized before any Corona-related restrictions.

[10] A strand of the psychological literature also 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; Pearce and Porter, 1986).

2.1 Institutional background

A substantial part of the German student population is enrolled at UAS and the study programs in our field experiments are among the most popular in Germany: in the winter term of 2019, about 38.4% of freshman students started studying at a UAS and 8.5% and 3.5% of freshman students enrolled in BuA and ME, making them the first and fourth most popular study programs in Germany. About 39.3% of BuA and ME freshman students in 2019 were women, while 37% of the students in our sample are female (Bundesamt, 2020). Importantly, and in contrast to Azmat et al. (2019) and (Dobrescu et al., 2021), we do not conduct our field experiments at a selective institution. Rather, students in our sample are somewhat negatively selected in terms of ability: About 41.6% of them hold a degree from the highest secondary education track – the so-called "Abitur", while this was the case for 51.1% of freshman students at UAS in Germany in 2018.[11] Additionally, the average high school GPA of students in our sample is 2.56, while it was 2.41 among all German high school graduates in 2019.[12]

The bachelor's programs at the UAS where we conduct our interventions are organized according to the European Credit Transfer System (ECTS, see Footnote 1) and have a scheduled study duration of seven semesters. Given that each program requires students to accumulate a total of 210 credits, they are thus expected to pass courses worth 30 credits per semester to graduate within the scheduled duration. In practice, students in the five programs take on average about 8.5 semesters to graduate (standard deviation of about 1 semester).

Students can at all times access information on their individual study progress via a web portal maintained by the university. The portal provides data on absolute performance – credits and GPA. In the absence of our treatment, the university does not provide any information on a student's relative performance.

2.2 Field Experiment I

The initial field experiment is conducted with a cohort of first-year students who enrolled in five bachelor's programs offered by the BuA and ME faculties (Table A.2 provides an overview of all degree programs and the number of students in our intervention). Treatment commences in the second semester, i.e., when information on first semester performance is

[11] See `https://www.datenportal.bmbf.de/portal/en/Tabelle-2.5.106.html`, retrieved on January 26, 2022.

[12] See `https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik/abiturnoten.html`, retrieved on March 08, 2022. In the German system 1.0 is the best and 4.0 the worst final high school GPA.

available. Our sample consists of all 812 students who are still enrolled in their study program at the start of the second semester. There is no selection into or out of the sample of the field experiment.

Randomization. Randomization was carried out after the first semester, using stratification and balancing (Morgan and Rubin, 2012). We built strata along bachelor's programs and obtained first-semester credits. Within these blocks we balanced on age, sex, high school GPA, time since high school graduation, pre-treatment GPA, and type of high school degree.[13] Table 1 shows that in the full sample all covariates are balanced. As we will explain later, an important subgroup in our paper are above-average students. Table A.4 shows that all variables are balanced in this group as well.

Feedback letter I. In the week before the second semester lectures start, students in the control and treatment group receive an unannounced letter in the mail, providing them with information on their accumulated credits and their cumulative GPA (see timeline in Figure 1). The letters thus include the same information that is available on the web portal of the university. The treatment group additionally receives a graphical illustration that provides relative performance feedback on accumulated credit points. This feedback is shown in Figure 2, and we explain the design in detail below (the full letters are shown in Figures A.1 and A.2).

Feedback letter II. About four to five weeks before the exam period, students of both groups receive a second letter (see Figure 1). The letter design is identical to the first one, and for most students the contained information will also be identical to the first letter. In some cases the university updated the information on grades and credits (e.g., because first semester course results were not yet available when the first letter was composed), which can lead to different feedback compared to the first letter.[14] 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 will use the content of the graph in the second letter when studying heterogeneity across the different types of feedback that students receive.

Relative performance feedback. The relative performance information in the treatment group is shown in Figure 2. It closely follows social comparison approaches which have re-

[13] See Appendix C for more details on the randomization procedure.

[14] See Appendix C for details on the reasons and the number of observations that are affected.

peatedly been shown to reduce energy consumption, e.g., Allcott (2011), Allcott and Rogers (2014), and Schultz et al. (2007). A bar chart compares the individual student's earned credits to the "Top 20%" and to "All" students who are enrolled in the same bachelor's program and are in the same cohort as the student receiving the letter. The feedback explains that "All" represents the average number of credits of students in their comparison group (see next paragraph), and that the "Top 20%" of students in the comparison group earned at least the displayed amount of credits. The specific values are given by the 80th percentile and the median in the comparison group.[15]

To further personalize the performance feedback, we define several comparison groups for each program. This increases perceived similarity and minimizes the psychological distance to the reference group (Festinger 1954 and Trope and Liberman 2010). In smaller programs the comparison group consists of students "who in/before ⟨*year*⟩ earned their school leaving certificate.", where *year* is the year in which the addressee of the letter received their school leaving certificate. In the two large bachelor's programs, we use more fine-grained comparison groups by additionally distinguishing between the school-leaving certificates "vocational track degree (or below)" and "general track degree".

According to the focus theory of normative conduct (Cialdini, Reno and Kallgren, 1990; Cialdini, 2011), this relative feedback can represent a descriptive norm, i.e., a reference point against which students can compare their performance. To prevent those above a norm from focusing downwards and reducing their effort (the so-called "boomerang effect"), the theory proposes to approve their performance using an injunctive norm. We thus follow the literature and add the following approving normative frames (Allcott, 2011; Cialdini, 2003; Schultz et al., 2007): performance that is at least average is categorized as *Good* (plus one "smiley" emoticon) and performance that is equal to or better than the 20% percentile is called *Great* (plus two "smiley" emoticons). Students who perform below the average do not receive an approving norm. Instead, the feedback includes the statement "currently below average" (and no emoticon).

Based on this design, in the heterogeneity analyses in Section 3, we consider the following feedback types: below-average, on-average, above-average (*Good*), and above-average (*Great*) feedback; see Table 2. Although the feedback for on-average students also includes a *Good* frame, we study this feedback type separately from the above-average type, because the literature suggests that normative frames may not have the intended effect if they are not aligned with the descriptive information (Cialdini et al., 2006). As we will discuss further in Sections 3 and 4, we also analyze the treatment effects for those who receive above-average

[15]The letters use the term "average" – which does not clearly denote one specific statistic – since students at the beginning of their studies may not be familiar with the term median. We have no information about whether students interpret the term as the mean or the median.

feedback versus those who do not.

2.3 Field Experiment II: Replication

Using the same design as in the original experiment, we repeat the intervention one year later in the same programs and faculties. This time the sample consists of 797 students who were enrolled for the second semester (Tables 1 and A.4 show the balancing properties[16]). The aim is to establish with a new cohort of students whether the results are replicable. In the taxonomy of Hunter (2001), Hamermesh (2007), and Czibor, Jimenez-Gomez and List (2019), ours is a statistical replication, as it is based on a different sample (new cohort) but uses an identical model (same protocol as in the initial experiment) and the same underlying population.[17] Czibor, Jimenez-Gomez and List (2019) argue that early statistical replications are "crucial" in experimental economics. Levitt and List (2009) call the difficulty of replication a "potential shortcoming" of field experiments, and so providing credible evidence that results can be reproduced is an important part of our study, especially given the inconclusive results in the literature on relative performance feedback and, more generally, the recent debate about replicability in economics and other fields (Camerer et al., 2016; Duvendack, Palmer-Jones and Reed, 2017; Open Science Collaboration, 2015).

The only small change in the replication experiment is that we use the mean instead of the median, when computing the average performance for the "All" bar in the feedback letters. This decision was made based on the result from the initial experiment that above-average feedback improves performance, as we will discuss in Sections 3 and 4. Table 2 shows that this tweak increases the share of students that receive above-average feedback from 37.5% in the initial experiment to 56.2% in the replication, and the share of students which exactly match the average is reduced from 20.3% to 3.6%. The wording and graphical representation of the feedback are kept identical across experiments. In the feedback letters the mean is referred to as the "average". This ensures that across the two experiments students interpret the feedback information in the same way and that the design of the original intervention remains unaltered.

2.4 Data and estimation

We use student-level data provided by the university's examination office and augment it with online surveys. Our main outcome of interest is the number of obtained credits, but

[16] We made some small adjustments to the randomization procedure in Experiment II. See Appendix C for details.

[17] According to Hunter (2001) four features are essential for a statistical replication: same independent variable, same dependent variable, same procedures, and sampling from the same population.

as already mentioned, we will check for potential negative side effects on other domains, especially GPA and dropout behavior. We use demographic information and pre-treatment outcomes (first semester credits and first semester GPA) as covariates in our estimations. A detailed description of the data collection and processing as well as the use of the data for the randomization, feedback letters, and estimations can be found in Appendix C. Table A.3 provides a description of all variables.

Unless otherwise specified, we provide intention-to-treat effects from OLS estimations that compare the outcomes of the control and the treatment group. In the baseline specification, we follow the recommendations of Bruhn and McKenzie (2009) and control for the method of randomization:

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \mathbf{s_i}\boldsymbol{\alpha_2} + \varepsilon_i, \tag{1}$$

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. The vector $\mathbf{s_i}$ includes strata fixed effects which control for the random assignment of treatment and control units within blocks. In estimations with pooled data from both experiments, we also include a cohort dummy and its interaction with the strata variables.

In the second specification, we add a vector capturing baseline performance and further control variables:

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \mathbf{s_i}\boldsymbol{\alpha_2} + \mathbf{x_i}\boldsymbol{\alpha_3} + \varepsilon_i. \tag{2}$$

$\mathbf{x_i}$ includes high school GPA, first semester credits, first semester GPA, age at randomization, an indicator for being female, time since high school graduation, and an indicator for the type of high school degree. 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.[18]

As discussed in Section 2.2 and shown in Table 2, building on the feedback design, we analyze heterogeneity across different feedback types:

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \mathbf{F_i}\boldsymbol{\alpha_2} + Treatment_i\mathbf{F_i}\boldsymbol{\alpha_{12}} + \mathbf{s_i}\boldsymbol{\alpha_3} + \varepsilon_i, \tag{3}$$

where $\mathbf{F_i}$ is a vector including either i) indicators for below-average (reference category), on-average, above-average (*Good*), and above-average (*Great*) feedback; or ii) an indicator for receiving above-average feedback (*Good* or *Great*). In a second specification, we again include the vector $\mathbf{x_i}$ with additional controls.

[18] See Appendix C for details.

3 Main results

3.1 Field Experiment I

Main effect. Table 3 shows the main effect of treatment on credits obtained in the second semester. Column (1) indicates that treated students obtain .665 additional credits when controlling only for the method of randomization (.054 control group standard deviations; henceforth abbreviated SD). Adding further control variables in Column (2) reduces the estimated effect to .287 credits (.023 SD). Although the parameters are positive, neither estimate is statistically significant and we therefore conclude that across the entire cohort, feedback does not increase performance.

Heterogeneity by feedback type. A characteristic of our experiment is that while all students in the treatment group receive feedback, the feedback slightly differs, depending on the position in the performance distribution (see Section 2.2). An important question to ask is thus whether students benefit from none of the feedback types or whether the treatment effects are, in fact, heterogeneous (see Bryan, Tipton and Yeager (2021) for an overview of the treatment effect heterogeneity in earlier feedback experiments on energy conservation, and the authors' call for a "heterogeneity revolution" in the behavioral sciences).

Panel (a) of Figure 3 visualizes the raw treatment effects with no control variables for the four different feedback types and Columns (1) and (2) in Panel (a) of Table 4 report estimates based on Equation 3 that control for the method of randomization and add covariates.

Figure and table show positive treatment effects for the two above-average feedback types. The raw treatment effect for above-average students who receive a *Good* frame is 4.874 credits (4.741 when controlling for the method of randomization and 4.150 with further controls) and significant at the 10%-level. Students in the top 20% received a *Great* frame, and we estimate a raw treatment effect of 1.718 credits (1.859 with strata controls and 1.393 with all controls) that is, however, not statistically significant (we discuss potential reasons for this smaller effect in Section 4.2). For both feedback types among students who are not above the average, on the other hand, we find no statistically significant effect on subsequent performance – though the point estimates are slightly negative. Still, there is an interesting distinction. For those below the average, receiving no approving normative frame is aligned with the information of having performed below the average. Yet, the students with an average performance receive an approving frame *(Good)*, so that frame and information can be considered misaligned in that the approving normative frame is not backed by factual information of being better than average. The lack of a (positive) treatment effect for this group is a tentative indication that simply attaching an approving normative frame to the informa-

tion about a merely average performance is not able to raise subsequent performance. This is in line with research, which suggests that normative frames will not have the intended effect if they are not aligned with the descriptive information (Cialdini et al., 2006).

Overall, it appears that those who receive information of being above the average respond positively to feedback, while those who receive other types of feedback show no behavioral response. For all above-average-students, we find a significant treatment effect of 2.411 credits (2.526 with strata controls and 1.988 with all controls); see Figure 3 (dark green and dark orange bars) and Panel (b) of Table 4. This corresponds to an effect size of roughly .16 to .20 SD. For those who do not receive above-average feedback we estimate effects between −0.198 (Figure 3) and −0.625 credits that are far from being statistically significant at any conventional level. The difference in treatment effects between these two feedback types, above versus not above average, is statistically significant (interaction T*Above-average).

We use these insights from the original experiment when setting up the replication experiment. Given the evidence of performance gains for those receiving above-average feedback, our main hypothesis for the second experiment is that above-average feedback has performance enhancing effects (Section 4 shows more evidence from RDDs that this feedback type is driving the effects). It therefore seems reasonable to find ways of extending the above-average feedback to a larger share of students. This can be achieved by defining the average as the mean instead of the median, since the mean number of credits is lower than the median number of credits. As a consequence, in the replication experiment a much larger share of students receives above-average feedback (see Table 2 and Figure 3). While this small change leads to a much higher prevalence of above-average feedback, it preserves all design features of the original experiment.

3.2 Replication experiment and pooled results

Experiment II: Replication. In the replication experiment with a new cohort of students, we test the main hypothesis derived from the first experiment: receiving above-average feedback leads to improved performance. In addition, we investigate whether the entire pattern of results from the earlier experiment replicates. As can be seen in Columns (3) and (4) of Table 3 the full sample effects in the replication are very similar to the original experiment, at 0.312 to 0.617 credits (.023 to .046 SD), and not statistically significant. Looking at the effects for the different feedback types in Panel (b) of Figure 3 and Columns (3) and (4) of Table 4, we very closely replicate the results from the original experiment.

In line with our main hypothesis derived from Experiment I, we find that students increase their subsequent performance by 1.837 credits in response to feedback of being above

the average (1.853 with strata controls and 1.632 with all controls). This corresponds to an effect size of roughly .12 to .14 SD. As in the original experiment, the treatment effects for students who placed above-average are significantly different from those who did not (for whom we again find no statistically significant effects).

Pooled results. Since the two experiments share the same design, we also report results based on pooling the observations to increase the power of our statistical analysis. In the spirit of Camerer et al. (2016) and Open Science Collaboration (2015), these results can also be interpreted as meta-analytic estimates. They are shown in panel (c) of Figure 3 and Columns (5) and (6) of Tables 3 and 4. Both experiments produced the same pattern of results and, as expected, the estimated treatment effects fall between the original and the replication experiment and are more precise due to the larger sample size. Just as in the two separate experiments, in the pooled sample, above-average feedback increases subsequent performance (by 1.787 to 2.134 credits; .14 to .17 SD) and the estimates are statistically significant at the 1%-level. Following the benchmark for effect sizes in education proposed by Kraft (2020), those are medium effects.[19] The effects can be interpreted as roughly one in three of these students passing an additional exam due to the treatment (on average, in our data an exam is worth 5.75 credits). Overall, the two experiments provide very robust evidence that above-average feedback increases performance. Students who are informed that they did not place above average show no behavioral response.

4 Mechanisms (i): regression discontinuity designs

In this section, we use regression discontinuity designs to show: (i) the behavioral response of above-average students is due to the information of being above average, and not driven by those students being more capable of responding to feedback than other students; (ii) the smaller treatment effect among above-average students who place in the top 20% is probably due to ceiling effects.

4.1 Characteristics of above-average students do not explain the response to above-average feedback

Given that the different feedback types are based on performance in the first semester, it is conceivable that the causal effect of above-average feedback is due to characteristics of above-average students and not due to the specific type of the feedback itself. For example,

[19] Based on 747 randomized controlled trials evaluating education interventions, Kraft (2020) proposes the following classification of effect sizes: $< .05$: small, $.05 - 0.2$: medium, and ≥ 0.2: large.

these students might have higher ability, a better learning technology, or they may be more motivated, enabling not only a higher first-semester performance, but also a better response to relative feedback.

Similar to the approach in Allcott (2011), we investigate this by employing a sharp regression discontinuity design (RDD) among treated students. We compare treated students just above the average to those on the average or just below, as they should not differ in their underlying characteristics. Instead, the only difference is the type of feedback that students receive: above-average feedback or not.

When implementing the RDD, we follow the suggestions of Lee and Lemieux (2010). If the usual RDD assumption holds, i.e., if there are no other discontinuities around the cutoff, it provides a causal local average treatment effect. To gather some intuition if this assumption is likely to hold, we study the control group, for which we should find no behavioral changes at the cutoff since they receive no relative performance feedback.[20] The running variable is the accumulated credits a student obtained in the first semester as depicted in the feedback letter, divided by the average credits of their respective comparison group (the corresponding distribution is shown in Figure B.1 in the Appendix).[21] Besides providing graphical depictions of the behavior of the outcome variable around the cutoff, we estimate the size of the jump by implementing a parametric RDD, using the following equation:

$$Y_i^k = \alpha_0 + \alpha_1 P_i + f(r_i) + f(r_i)P_i + \mathbf{s_i}\boldsymbol{\alpha_2} + \varepsilon_i, \tag{4}$$

where P_i indicates if a person placed on the right side of the cutoff, i.e., above the average. $f(r_i)$ is any smooth function of the running variable r_i that we allow to vary between the left and the right side of the cutoff and $\mathbf{s_i}$ is a vector including study program fixed effects and, in the pooled sample, a cohort fixed effect and its interaction with the study program fixed effects.[22]

The existence of a control group that does not receive relative performance feedback allows us to account for any potential jump in our outcome variable at the cutoff that is due to unobserved discontinuities that are the same in the treatment and the control group. We do this by estimating the following regression discontinuity difference-in-difference (RD-

[20] Another assumption is that individuals have no or only imprecise control over the running 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.

[21] This provides smoother distributions around the cutoff than using the raw distance to the cutoffs, because of differences in the credit point distributions within the different comparison groups.

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

DID) specification:[23]

$$
\begin{aligned}
Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \alpha_2 P_i + \alpha_{12} Treatment_i P_i \\
+ \alpha_3 r_i + \alpha_{13} Treatment_i r_i + \alpha_{23} P_i r_i + \alpha_{123} Treatment_i P_i r_i + \mathbf{s_i \alpha_4} + \varepsilon_i,
\end{aligned}
\tag{5}
$$

where we are interested in the parameter α_{12}.

Figure 4 visualizes the behavior of the outcome variable around the cutoff for the pooled sample. For the treatment group we observe a large jump of about seven credits when students receive above-average feedback, while we find no jump for the control group. Estimates of the respective coefficients based on two different discontinuity samples and a first order polynomial are shown in Table 5. Across the two experiments and the pooled sample, we estimate effects of 5.492 to 8.745 credits for students in the treatment group, all significant at the 1%-level. For students in the control group, on the other hand, we find no evidence for changes in second semester performance at the cutoff. Accordingly, coefficients based on the RD-DID specification shown in the bottom row confirm the effects estimated for the treatment group.[24]

Overall, these RD estimates thus deliver robust evidence that relative performance feedback increases the subsequent performance of a student around the cutoff by roughly six credits if it informs them about an above-average instead of an at or below-average performance. This finding provides evidence that the large treatment effects for above average students are not due to their underlying characteristics. Rather, it suggests that the content of the relative performance feedback matters for the increase in subsequent performance. It is important to reiterate that the content of the feedback includes being above average or not as well as the differing normative frames (see Section 2.2). The design of our experiment does not allow us to cleanly disentangle whether the information about being above average itself or the combination with the approving normative frame is driving the effect of above-average feedback. As noted in Section 3.1, the fact that we do not observe positive treatment effects for students with an average performance provides evidence that the inclusion of an approving normative frame alone is not sufficient to elicit performance gains.

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

[24] In Table B.1 in the Appendix, we show that the pooled sample estimates are robust to different polynomial specifications and further discontinuity samples. Estimated coefficients only become imprecise and unreliable when we increase the order of the polynomials while at the same time using small discontinuity samples. We are not concerned about this, as higher order polynomials can lead to overfitting, especially when the number of observations is low. As an additional robustness check, Figures B.2 to B.3 visualize the behavior of pre-treatment covariates around the cutoff. Most of them behave smoothly and in the few cases where we do observe small discontinuities, they behave similarly in the treatment and the control group. Any effects of those discontinuities should therefore be captured by our RD-DID specification.

4.2 Why are the positive effects smaller for students in the top 20 percent?

While the treatment effects are positive for all students above the average, the top 20% of students react less to feedback. We want to understand what is behind this heterogeneity – specifically, we are interested in whether it is driven by the slight differences in feedback, or by the characteristics and specific circumstances of the top 20% students.

This can be assessed with an RDD following the same approach as in the last Section. We now use the ratio of first semester credits to the 80th percentile of credits in the comparison group as our running variable and study the sharp cutoff at the 80th percentile.[25] If the attenuated treatment effects among students in the top 20% are indeed caused by the slightly different feedback type, we should find a negative jump in the outcome variable at the 80th percentile among treated students.

Figure 5 and the estimates in Table 6, show no significant jump at the cutoff.[26] We therefore conclude that there is no evidence of the differences in feedback type being behind the observed heterogeneity among above-average students.

Instead, we argue that the heterogeneity is rooted in the characteristics and circumstances of those in the top 20%. Specifically, ceiling effects are a likely explanation. In Figure 3 it can be seen that treatment raises second semester performance for the above-average groups to roughly 29 credits, but no higher than that. Figure 6 provides additional evidence. It shows the cumulative distributions of total credit points at the end of the second semester for those above and those not above the average, separated by treatment status. Clearly, the treatment effects among above-average students are concentrated in the lower parts of the distribution, i.e., below 60 first-year credits.

Several mechanisms provide potential explanations for these observations. First, it could be the case that even top performing students do not have the capacity, cognitive or otherwise, to obtain more than roughly 30 credits per semester. Second, and perhaps more plausible, in the standardized European ECTS system, students are advised to collect 30 credits per semester, and curricula are set accordingly. This number may set an "artificial" ceiling or reference point that most students do not exceed. Finally, in Section 6.1, we will show that the effect of relative feedback is larger for students who did not expect to perform above average. It is therefore also possible that the best performing students are less surprised and, accordingly, less affected by the above-average feedback.

[25] The distribution of the running variable is shown in Figure B.4 in the Appendix.

[26] Table B.2 in the Appendix shows that this result is robust to various polynomial specifications and discontinuity samples.

5 Repeated treatment

In this section, we report results for the effects of repeated treatment, i.e., effects on performance in the third semester (= second treatment semester). We have to take into account that the treatment effects on second semester performance can affect the type of feedback that students receive in the third semester. To avoid this possible endogeneity problem, we therefore analyze the effects of receiving a certain type of feedback in the second semester on performance in the third semester.

Results based on estimations with all control variables are provided in Table 7. Panel (a) shows estimates for the treatment effects on credits obtained in the third semester. As in the second semester, we do not find evidence for effects on performance among all students (Columns 1 to 3) or for those who did not place above average (Columns 4 to 6) in either cohort. We also do not find additional positive effects for students who received above-average feedback in the second semester (Columns 4 to 6). In Panel (b), we report effects on the accumulated credits at the end of the third semester. In line with the previous results, the full sample effect is close to zero. Regarding the effects on above-average students, we find that the estimated coefficients are close to the corresponding effects on second semester credits shown in Panel (b) of Table 4. For example, in the pooled sample, treated above-average students have accumulated 1.658 credits more by the end of the third semester, which is close to the second-semester treatment effect of 1.787 credits reported in Column (6) of Table 4. However, due to the higher variance of the accumulated credits at the end of the third semester – see the bottom row of Panel (b) in Table 7 – the coefficient is no longer statistically significant ($p = 0.148$).

These findings suggest that there are no additional effects of repeated treatment on top of those that we found for the second semester. In the next section, we provide evidence that the significant second semester effects, as well as the lack of additional effects of repeated treatment may plausibly be driven by beliefs about relative performance.[27]

[27] The institutional setup at the university may also play a role for the dynamics of the effects, and prevent treatment effects in the third semester. For more than half of the students in our sample, the mandatory internship semester is scheduled for the fourth semester. Therefore, students who wish to take additional exams in the third semester may have to choose from courses that are scheduled for the period after the internship semester. These courses are mostly electives, and because the internship period is supposed to help students figure out which electives to choose, they may decide against taking these exams early – thus precluding further effects.

6 Mechanisms (ii): beliefs about relative performance and theoretical considerations

6.1 Expectations and treatment effects

The goal of the second experiment was not only replication, but also to further investigate the mechanisms through which feedback changes behavior. One important condition for the effectiveness of feedback is that it provides new information and that individuals actually process it. To shed light on this, in the replication experiment, we conducted pre- and post-treatment surveys asking students about their expected relative performance in terms of credits; see Figure 1 for the timing of the surveys and Table A.5 for the questions.[28][29] We can then address the following questions. First, how do students' expectations line up with actual relative performance? Second, are the beliefs about relative performance influenced by the feedback? Third, does the treatment effect depend on the accuracy of the initial beliefs?

To address the first two questions, we use the survey questions to create a variable that indicates whether a student expects to place above average in the performance distribution at the end of the semester. Figure 7 visualizes the pattern of belief updating across the three semesters separately for students who performed above-average (top panel) and students who did not (bottom panel). The two left panels provide evidence that pre-treatment students have very little intuition about their actual relative performance. Irrespective of actually placing above average or not, around 50% of treatment and control students expect to perform above average. In the second semester (middle panels), we find that those who actually placed above average updated their beliefs: 87% in the treatment group and 68% in the control group now expect to be better than the average; the difference of 19 percentage points is significant at the 5%-level. For students who did not receive above-average feedback in the second semester, on the other hand, neither the control nor the treatment group appears to update their expected relative performance. The right panels show that in the third semester a large majority of above-average students, both in the treatment and the control group, expects to perform better than the average. For those who did not receive above-average feedback, we still find no evidence for updates in beliefs.

[28] The wording in the pre- and post-treatment survey is not exactly 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.

[29] One caveat applies when considering the data from the surveys: in line with the general development in survey nonresponse rates (see, e.g., Leeper 2019), the response rates in our surveys are between 15 and 30% (see Figure 1). Accordingly, our sample is rather small, and we find evidence that the respondents are a positively selected subpopulation.

These results suggest the following: First and in line with the positive effects on performance, in the second semester above-average feedback leads to more accurate beliefs about relative performance compared to controls. Second, over time control students who performed above average learn about their relative performance, even in the absence of relative feedback. As a result, in the third semester, beliefs of students in the control group are almost as accurate as the beliefs of students in the treatment group. This disappearing informational gap between the two groups can plausibly explain the lack of additional treatment effects in the third semester. Last, we cautiously take the fact that those who did not place above average do not update beliefs as evidence that students discard or discount the feedback if they do not perform above average. This could provide an explanation why in the second semester we only observe a behavioral response for above-average students.

We can also study whether above-average students who received new or unexpected information respond more strongly – in contrast to students who already expected that they would perform above average.[30] To test this, we create a dummy U_i that is 1 if a student underestimated their performance in the first semester, i.e., they did not expect an above average performance although they then actually performed better than the average. We estimate the following equation among students that placed above average:

$$Y_i^k = \alpha_0 + \alpha_1 Treatment_i + \alpha_2 U_i + \alpha_{12} Treatment_i U_i + \mathbf{s_i}\boldsymbol{\alpha_3} + \varepsilon_i, \tag{6}$$

where α_1 gives the treatment effect for those who correctly expected to be above average, and α_{12} gives the difference in the treatment effect for those who underestimated their relative performance. Table 8 presents the results. Column (3) shows that the treatment effect for those who correctly estimated their position is roughly 1.426 credits, which is about 1.3 points smaller than the 2.678 credit treatment effect that we find in the entire survey sample (Column 1). Control group students who underestimated their relative performance obtain on average 2.841 credits less in the second semester than students who did not underestimate their actual relative performance. 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 Columns (2) and (4) we add control variables; especially the pre-treatment performance should be accounted for, as it is correlated with both the post-treatment performance and with the first semester expectations. We find that the negative effect of underestimating performance now becomes stronger. Given the covariates, this indicates that at equal ability, lower confidence can be detrimental to performance. Again, this negative effect is

[30] The number of survey respondents who perform (below-)average is too low to study their behavior in such detail.

completely offset by receiving feedback. Overall, the treatment generates a significant effect of 5.713 credits for students who underestimated their relative performance (fourth row in Column 4), indicating that relative performance feedback will be especially helpful for this group.

6.2 Theoretical considerations

How does our pattern of results and the evidence on students' beliefs about their relative performance line up with some of the most common theories on the effects of feedback and social comparison?

First, social comparison theory and competitive preferences do not fit our results particularly well: The focus theory of normative conduct suggests that individuals try to comply with descriptive norms, e.g., the average performance level, predicting positive treatment effects for those below the descriptive norm (Cialdini, Reno and Kallgren, 1990; Cialdini, 2011). For those above the norm, the focus is on preventing negative effects by adding an approving message, i.e., an injunctive norm (Allcott, 2011; Cialdini, 2003; Schultz et al., 2007). If individuals have competitive preferences, relative feedback is usually predicted to increase performance across the entire distribution (Azmat and Iriberri, 2010; Dobrescu et al., 2021). Given that we find no evidence for an increase in performance after below-average feedback, these theories do not provide a convincing explanation for our pattern of results.

Second, "self-perception theory" proposes that feedback can influence behavior by changing beliefs over ability (Azmat and Iriberri, 2010; Dobrescu et al., 2021; Ertac, 2005): performance is a function of both effort and ability, and ability and effort are assumed to be complements. Feedback will then affect beliefs about ability and thus the optimal choice of effort, leading to increased (decreased) effort and performance if the feedback signals a higher (lower) ability than individuals previously believed. In our context, above-average feedback provides a favorable signal about ability, explaining the positive treatment effects for this group of students – in particular among those who did not expect to perform above average. However, our results for students who do not place above average and who arguably receive an unfavorable signal about their ability, are in contrast with the theoretical predictions, as we find no evidence for negative effects.

Instead, our pattern of results and belief updating can best be reconciled with theories on the management of self-confidence and the selective processing of information that often accompanies it (Villeval 2020 provides an overview on how relative performance feedback and confidence are connected). Similar to the mechanism in "self-perception theory", relative feedback provides a signal about ability. Because the signal can affect the confidence of individuals in their ability, it creates incentives to process feedback in a confidence-

preserving way. For example, in the model by Bénabou and Tirole (2002) favorable signals serve individuals to form beliefs about their ability and maintain a positive self-image; positive effects on performance are expected, because effort and ability are assumed to be complements. To maintain confidence in one's own ability, adverse signals, on the other hand, do not adequately enter into beliefs: individuals selectively process good information. In Compte and Postlewaite (2004) the mechanism is similar: here, positive outcomes are attributed to own abilities or efforts whereas 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.[31]

The notion that individuals will update beliefs in ego-relevant domains such as relative performance asymmetrically is supported by evidence from laboratory experiments: individuals who receive good news about their rank in an IQ test are willing to incorporate this information in their beliefs, while individuals who receive negative news have little willingness to update their self-concept (Eil and Rao, 2011; Möbius et al., in press).

Our result pattern is in accordance with the theoretical ideas in Bénabou and Tirole (2002), if we assume that students' beliefs about relative performance are linked to their beliefs about ability. In the second semester, above-average students who receive feedback have more favorable beliefs about their relative performance than controls and increase their subsequent performance. By the third semester, our results on students' beliefs suggest that beliefs about ability in the control and the treatment group have converged, which explains why there are no additional effects on performance. Students who did not receive above-average feedback show no sign of correcting their beliefs downward. Consistent with Bénabou and Tirole (2002), these students may hold optimistic beliefs about their abilities to overcome lack of willpower and stay motivated.[32] Feedback would inform these students that their ability is lower than initially believed, which would lead to a decline in motivation and effort and potentially worse outcomes. Disregarding feedback may then prevent this by preserving confidence in ability.

[31] 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; Köszegi, 2006). This may also be a motivational factor for effort allocation.

[32] Estimates suggest that up to 95% of college students may be subject to self-control problems (Ellis and Knaus, 1977; O'Brien, 2002); König, Schweighofer-Kodritsch and Weizsäcker (2019) provide evidence that university students manage beliefs about return to effort in ways consistent with Bénabou and Tirole (2002).

7 Spillovers

7.1 Negative spillovers to other domains?

We have found robust effects of above-average feedback on achieved second semester credits. An important question is whether students generate these gains in performance at the cost of losses in other domains.

First, a concern could be that encouraging students to obtain more credits may come at the expense of worse grades because students may shift attention away from them. In Panel (a) of Table 9, we report treatment effects on students' GPA at the end of the second semester. We do not find any significant effects, neither in the full sample (Columns 1 to 3) nor among above-average students (Columns 4 to 6). Table A.6 in the Appendix shows that this also holds true in the third semester. It thus appears that feedback can raise performance in terms of obtained credits, without negatively affecting the other major performance dimension.

Second, it could be the case that relative performance feedback affects the dropout decision of students. Panel (b) of Table 9 reports effects on having dropped out of the study program by the end of the second semester. As with GPA, we find no evidence for statically significant effects on students' dropout behavior; this is also the case in the third semester (see Table A.6 in the Appendix).

One might also worry that the feedback affects other dimensions that are indirectly related to performance. For both our experiments we conducted a post-treatment survey in which we asked students 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 Figure 1 for the timing of the surveys and Table A.7 in the Appendix for the questions and the variables used in the estimations). Table 10 shows the corresponding treatment effects for the pooled sample.[33] We find no statistically significant effects on any of the well-being dimensions we observe.

Taken together, these results show that the increase in the number of obtained credits does not come at the cost of negative effects on other outcomes, and can therefore be interpreted as a net positive effect of above-average feedback.

7.2 Spillovers to the control group?

Spillovers from the treatment to the control group might arise from treated students sharing the feedback information with the control group. Several observations suggest that the

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

performance enhancing effects of above-average feedback are unlikely due to spillovers.

If there are positive spillovers, and controls also benefit from the above-average feedback, our treatment effects are downward biased, presenting a lower bound for the true effects of above-average feedback. In theory, there might also be negative spillovers. For example, students in the control group may have felt disadvantaged by the fact that they supposedly did not get the same attention from the university as the treatment group, which then might result in reduced performance. However, this is unlikely as i) control students also receive a letter from the faculty including absolute feedback, and ii) the contact person named on the control group letters did not receive any complaints regarding missing relative feedback information.

When designing the intervention, we also incorporated measures to make sharing the feedback among students more difficult. First, as described in Section 2.2, we provide feedback not on the level of the degree program but by using smaller comparison groups (only students in the same program, and with the same year of school leaving certificate, and same type of university entrance qualification). This "tailored" information in the letters should be shared less frequently because it may not appear to be of interest to other students who do not share these characteristics. Second, the feedback graphic necessarily includes the individual's obtained credits, and the feedback letter additionally shows the student's GPA. Both are information most students do not want to make public.

We also have some information tentatively indicating the absence of substantial spillover effects. First, anecdotal feedback from students suggests that they did not observe any sharing of this information, e.g., on social media. Second, the results on student beliefs about relative performance presented in Section 6.1 show that within the subgroup of students who received – or in the control group qualified for – above-average feedback, the beliefs of treated students are significantly more accurate in the first treatment semester compared to control.[34] Finally, as mentioned in Section 4.1, the results of the RDD also provide evidence against spillovers. In the case of widespread spillovers, we would expect controls to also be aware of whether they placed above average or not. Contrary to what we actually observe, we would then expect to find a discrete jump at the average in the control group, too.

[34] However, the finding that the control group also appears to learn about relative rank over time may be at least partly driven by sharing of the feedback information, and we thus cannot rule out that we underestimate the effects of the treatment, particularly in the third semester.

8 Conclusion

In a field experiment and a direct replication, we investigate the effects of relative feedback on academic performance. Our results show that students increase their subsequent performance when the feedback informs them about an above-average performance. With the help of a regression discontinuity design we show that this is true irrespective of the underlying characteristics of the students.

In order to investigate the mechanism behind the behavioral reaction to above-average feedback, we survey individuals about their pre- and post-treatment expectations concerning relative performance and find suggestive evidence that the information about a below-average or average performance is not processed in the same way as the information about an above-average performance – which can explain the difference in the behavioral responses. In addition, our findings suggest that relative performance feedback is especially effective for those individuals with an above-average performance who initially underestimate their relative performance and this underestimation is linked to worse performance in the absence of relative feedback. The pattern of results is consistent with theoretical ideas suggesting that a higher confidence in ability motivates individuals, and that individuals try to maintain a positive self-assessment of their abilities by selectively processing information (see, e.g., Bénabou and Tirole 2002).

Our results have important implications from a policy perspective. The intervention presents a low-cost and easy to implement tool which can increase the performance of a large share of students at a time that is crucial for habit-formation and getting on track to graduation. In addition, our findings suggest that the feedback is especially helpful to individuals who are held back by underestimation of their relative abilities. We also present tentative evidence that selective information processing may prevent undesirable effects of relative feedback when it threatens confidence in ability. While we find no negative effects of feedback on those in the lower parts of the performance distribution, policy-makers should be aware that this still implies that relative performance feedback can have distributional implications to the effect of widening achievement gaps.

Future feedback schemes aimed at preventing this should take into account that feedback for weaker students may need to be designed in a way that does not jeopardize confidence but is able to enhance it. Future studies might also want to explicitly test whether the inclusion of normative frames can make (otherwise neutral) relative performance feedback more effective and whether normative frames can be crafted such that they benefit weaker students. Another interesting avenue for future research is to investigate if and under which conditions feedback can generate beneficial effects in the long-run.

References

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics,* 95(9–10): 1082–1095.

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

Altmann, Steffen, Andreas Grunewald, and Jonas Radbruch. in press. "Interventions and Cognitive Spillovers." *Review of Economic Studies.*

Ashraf, Anik. 2022. "Performance Ranks, Conformity, and Cooperation: Evidence from a Sweater Factory." *CESifo Working Paper No. 9591.*

Attewell, Paul, and David Monaghan. 2016. "How Many Credits Should an Undergraduate Take?" *Research in Higher Education,* 57(6): 682–713.

Attewell, Paul, Scott Heil, and Liza Reisel. 2012. "What is Academic Momentum? And Does it Matter?" *Educational Evaluation and Policy Analysis*, 34(1): 27–44.

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

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

Azmat, Ghazala, Manuel Bagues, Antonio Cabrales, and Nagore Iriberri. 2019. "What You Don't Know... Can't Hurt You? A Natural Field Experiment on Relative Performance Feedback in Higher Education." *Management Science*, 65(8): 3449–3947.

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

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

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

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

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

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

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

Bryan, Christopher J., Elizabeth Tipton, and David S. Yeager. 2021. "Behavioural Science Is Unlikely to Change the World Without a Heterogeneity Revolution." *Nature Human Behaviour*, 5(8): 1–10.

Bundesamt, Statistisches. 2020. "Bildung und Kultur: Studierende an Hochschulen Wintersemester 2019/2020."

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

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science*, 351(6280): 1433–1436.

Celik Katreniak, Dagmara. 2018. "Dark Side of Incentives: Evidence From a Randomized Control Trial in Uganda." *mimeo*.

Charness, Gary, David Masclet, and Marie Claire Villeval. 2014. "The Dark Side of Competition for Status." *Management Science*, 60(1): 38–55.

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

Cialdini, Robert B. 2011. "The Focus Theory of Normative Conduct." In *Handbook of Theories of Social Psychology*. Vol. 2, , ed. Paul A. M. Van Lage, Arie W. Kruglanski and E. Tory Higgins, 295–312. Sage Thousand Oaks, CA.

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

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

Clark, Damon, David Gill, Victoria Prowse, and Mark Rush. 2020. "Using Goals to Motivate College Students: Theory and Evidence From Field Experiments." *Review of Economics and Statistics*, 102(4): 648–663.

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

Corcoran, Katja, Jan Crusius, and Thomas Mussweiler. 2011. "Social Comparison: Motives, Standards, and Mechanisms." In *Theories in Social Psychology*. , ed. Derek Chadee, 119–139. Wiley-Blackwell.

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

Czibor, Eszter, David Jimenez-Gomez, and John A. List. 2019. "The Dozen Things Experimental Economists Should Do (More of)." *Southern Economic Journal*, 86(2): 371–432.

Danzer, Natalia, and Victor Lavy. 2018. "Paid Parental Leave and Children's Schooling Outcomes." *Economic Journal*, 128(608): 81–117.

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

DellaVigna, Stefano, and Elizabeth Linos. 2022. "RCTs to Scale: Comprehensive Evidence from Two Nudge Units." *Econometrica*, 90(1): 81–116.

Dobrescu, Loretti Isabella, Marco Faravelli, Rigissa Megalokonomou, and Alberto Motta. 2021. "Relative Performance Feedback in Education: Evidence from a Randomised Controlled Trial." *Economic Journal*, 131(640): 3145–3181.

Doyle, William R. 2011. "Effect of Increased Academic Momentum on Transfer Rates: An Application of the Generalized Propensity Score." *Economics of Education Review*, 30(1): 191–200.

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

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

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

Ellis, Albert, and William J. Knaus. 1977. *Overcoming Procrastination.* New York:Signet Books.

Elsner, Benjamin, and Ingo E. Isphording. 2017. "A Big Fish in a Small Pond: Ability Rank and Human Capital Investment." *Journal of Labor Economics*, 35(3): 787–828.

Elsner, Benjamin, Ingo E Isphording, and Ulf Zölitz. 2021. "Achievement Rank Affects Performance and Major Choices in College." *Economic Journal*, 131(640): 3182–3206.

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

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

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

Fischer, Mira, and Valentin Wagner. 2018. "Effects of Timing and Reference Frame of Feedback: Evidence from a Field Experiment." *IZA Discussion Paper No. 11970*.

Fryer Jr., Roland G. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics*, 126(4): 1755–1798.

Gardner, John W. 2020. "Managing Production Yields and Rework through Feedback on Speed, Quality, and Quantity." *Production and Operations Management*, 29(9): 2182–2209.

Gill, David, Zdenka Kissová, Jaesun Lee, and Victoria Prowse. 2019. "First-Place Loving and Last-Place Loathing: How Rank in the Distribution of Performance Affects Effort Provision." *Management Science*, 65(2): 494–507.

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

Goulas, Sofoklis, and Rigissa Megalokonomou. 2021. "Knowing Who You Actually Are: The Effect of Feedback on Short- and Long Term Outcomes." *Journal of Economic Behavior & Organization*, 183: 589–615.

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

Hamermesh, Daniel S. 2007. "Replication in Economics." *Canadian Journal of Economics/Revue canadienne d'économique*, 40(3): 715–733.

Hermes, Henning, Martin Huschens, Franz Rothlauf, and Daniel Schunk. 2021. "Motivating Low-Achievers – Relative Performance Feedback in Primary Schools." *Journal of Economic Behavior & Organization*, 187: 45–59.

Hirshleifer, Sarojini R. 2021. "Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance." *CEGA Working Paper Series No. WPS-182*.

Hunter, John E. 2001. "The Desperate Need for Replications." *Journal of Consumer Research*, 28(1): 149–158.

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

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

Kajitani, Shinya, Keiichi Morimoto, and Shiba Suzuki. 2020. "Information Feedback in Relative Grading: Evidence from a Field Experiment." *PLoS ONE*, 15(4): e0231548.

König, Tobias, Sebastian Schweighofer-Kodritsch, and Georg Weizsäcker. 2019. "Beliefs as a Means of Self-Control? Evidence from a Dynamic Student Survey." *WZB Discussion Paper SP II 2019–204*.

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

Kraft, Matthew A. 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher*, 49(4): 241–253.

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

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

Leeper, Thomas J. 2019. "Where Have the Respondents Gone? Perhaps We Ate Them All." *Public Opinion Quarterly*, 83(S1): 280–288.

Levitt, Steven D., and John A. List. 2009. "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review*, 53(1): 1–18.

Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. in press. "Managing Self-Confidence: Theory and Experimental Evidence." *Management Science*.

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

Murphy, Richard, and Felix Weinhardt. 2020. "Top of the Class: The Importance of Ordinal Rank." *Review of Economic Studies*, 87(6): 2777–2826.

O'Brien, William K. 2002. "Applying the Transtheoretical Model to Academic Procrastination." *Doctoral Dissertation*.

OECD. 2019. "Education at a Glance 2019: OECD Indicators."

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

Oreopoulos, Philip, and Uros Petronijevic. 2019. "The Remarkable Unresponsiveness of College Students to Nudging And What We Can Learn from It." *NBER Working Papers No. 26059*.

Oreopoulos, Philip, Richard W. Patterson, Uros Petronijevic, and Nolan G. Pope. 2022. "Low-Touch Attempts to Improve Time Management among Traditional and Online College Students." *Journal of Human Resources*, 57(1): 1–43.

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

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

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

Taylor, Shelley E., Heidi A. Wayment, and Mary Carrillo. 1996. "Social Comparison, Self-Regulation, and Motivation." In *Handbook of Motivation and Cognition, Vol. 3. The Interpersonal Context.* , ed. Richard M. Sorrentino and E. Tory Higgins, 3–27. Guilford Press.

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

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

Villeval, Marie Claire. 2020. "Performance Feedback and Peer Effects." In *Handbook of Labor, Human Resources and Population Economics.* , ed. Klaus F. Zimmermann. Springer.

Zhang, XiaoLi, Jelle de Vries, René de Koster, and ChenGuang Liu. 2021. "Fast and Faultless? Quantity and Quality Feedback in Order Picking." *Production and Operations Management.*

Tables and figures

Table 1: Descriptive statistics and balancing properties

	Experiment I			Experiment II: Replication		
	(1) Control Mean (Std. Dev.)	(2) Treatment Coefficient (Robust SE)	(3) p-Value	(4) Control Mean (Std. Dev.)	(5) Treatment Coefficient (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[a)]	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[a)]	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)	
GPA at Randomization	2.491	-0.039	0.479	2.584	-0.031	0.493
	(0.713)	(0.054)		(0.683)	(0.046)	
% GPA at Randomization NA[a)]	0.264	-0.014	0.426	0.116	-0.007	0.707
	(0.441)	(0.017)		(0.320)	(0.018)	
N	405	407		398	399	

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

Table 2: Experimental design and number of observations by feedback type

Desc. Perf. Information	Not above average		Above average		
	Below average	On average	Below top 20%	Top 20%	
Normative Frame	No frame		*Good*	*Great*	*N* - Total
Experiment I	342	165	67	238	812
Control	174	76	34	121	405
Treatment	168	89	33	117	407
Experiment II: Replication	320	29	202	246	797
Control	163	12	97	126	398
Treatment	157	17	105	120	399

Note: Feedback type refers to the feedback students received in the first treatment semester.

Table 3: Effect of feedback on credits

			Replication			
	Experiment I		Experiment II		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.665	0.287	0.617	0.312	0.641	0.327
	(0.730)	(0.695)	(0.736)	(0.702)	(0.518)	(0.493)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes
N	812	812	797	797	1609	1609
Control Mean	21.07	21.07	19.75	19.75	20.42	20.42
(Std. Dev.)	(12.34)	(12.34)	(13.32)	(13.32)	(12.84)	(12.84)

Note: Outcome variable: credits second semester; *strata:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *controls:* HS GPA, credits first semester, GPA first semester, 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 4: Effect of feedback on credits – by feedback type

	Experiment I		Replication Experiment II		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel (a)						
Treatment (T)	-0.169	-0.459	-0.930	-1.412	-0.542	-0.893
	(1.196)	(1.196)	(1.147)	(1.146)	(0.829)	(0.826)
T*On-average	-1.088	-0.882	-3.470	-1.789	-1.145	-0.519
	(1.831)	(1.702)	(4.263)	(4.179)	(1.567)	(1.466)
T*Above-average *(Good)*	4.910*	4.609*	4.406**	4.553***	4.376***	4.245***
	(2.748)	(2.544)	(1.770)	(1.738)	(1.450)	(1.403)
T*Above-average *(Great)*	2.029	1.852	1.404	1.711	1.698	1.792
	(1.711)	(1.667)	(1.733)	(1.639)	(1.217)	(1.169)
T+T*On-average	-1.257	-1.341	-4.400	-3.200	-1.687	-1.411
	(1.390)	(1.238)	(4.110)	(4.030)	(1.333)	(1.219)
T+T*Above-average *(Good)*	4.741*	4.150*	3.475**	3.141**	3.834***	3.352***
	(2.471)	(2.237)	(1.346)	(1.299)	(1.188)	(1.130)
T+T*Above-average *(Great)*	1.859	1.393	0.474	0.299	1.156	0.900
	(1.222)	(1.155)	(1.300)	(1.179)	(0.891)	(0.828)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes
N	812	812	797	797	1609	1609
P-value of F-test	0.110	0.119	0.051	0.054	0.006	0.009
Panel (b)						
Treatment (T)	-0.334	-0.625	-0.968	-1.352	-0.587	-0.857
	(0.942)	(0.921)	(1.139)	(1.126)	(0.724)	(0.707)
T*Above-average	2.859**	2.613*	2.821*	2.984**	2.721***	2.644***
	(1.449)	(1.385)	(1.475)	(1.423)	(1.018)	(0.977)
T+T*Above-average	2.526**	1.988*	1.853**	1.632*	2.134***	1.787***
	(1.101)	(1.026)	(0.936)	(0.873)	(0.715)	(0.671)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes
N	812	812	797	797	1609	1609
P-value of F-test	0.049	0.059	0.056	0.036	0.008	0.007

Note: Reference category in Panel (a) is below-average feedback. Reference category in Panel (b) are not-above-average students. P-values in the bottom row of the two panels are from F-tests that test the hypothesis that all interaction terms of treatment with the feedback types, i.e., $\boldsymbol{\alpha_{12}}$ in Equation 3, are equal to zero. *Outcome variable:* credits second semester; *strata:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *controls:* HS GPA, credits first semester, GPA first semester, 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 5: RD estimates at average – first order polynomial

	Replication					
	Experiment I		Experiment II		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
Discontinuity Sample	$0.25 < r < 1.75$	$0.5 < r < 1.5$	$0.25 < r < 1.75$	$0.5 < r < 1.5$	$0.25 < r < 1.75$	$0.5 < r < 1.5$
Treatment Group	7.874***	6.675***	6.346***	8.745***	5.492***	6.418***
	(1.776)	(1.991)	(2.028)	(2.531)	(1.243)	(1.450)
N	336	295	320	238	656	533
Control Group	0.515	-0.144	-0.689	-0.787	-0.887	-1.246
	(2.073)	(2.633)	(2.415)	(2.972)	(1.465)	(1.743)
N	344	302	313	238	657	540
Diff-in-Diff	7.163***	6.377**	6.488**	9.035**	5.742***	7.126***
	(2.704)	(3.225)	(3.153)	(3.878)	(1.889)	(2.230)
N	680	597	633	476	1313	1073
Study Program FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: Outcome variable: credits second semester; *study program FE:* study program FE and in the pooled estimations a cohort dummy and its interaction with the study program FE; *running variable (r):* ratio of first semester credits as depicted in the feedback letter to the comparison group's average credits. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 6: RD estimates at 80th percentile – first order polynomial

	Replication					
	Experiment I		Experiment II		Pooled	
	(1)	(2)	(3)	(4)	(5)	(6)
Discontinuity Sample	$0.25 < r < 1.75$	$0.5 < r < 1.5$	$0.25 < r < 1.75$	$0.5 < r < 1.5$	$0.25 < r < 1.75$	$0.5 < r < 1.5$
Treatment Group	2.271	0.836	1.681	0.555	2.632	0.891
	(2.584)	(2.332)	(2.350)	(3.532)	(1.720)	(2.038)
N	348	285	319	267	667	552
Control Group	-0.108	0.926	1.135	5.094**	0.621	3.256
	(2.808)	(2.918)	(2.562)	(2.575)	(1.894)	(2.113)
N	355	291	314	267	669	558
Diff-in-Diff	2.485	0.012	0.734	-4.330	2.097	-2.233
	(3.778)	(3.639)	(3.368)	(4.200)	(2.501)	(2.848)
N	703	576	633	534	1336	1110
Study Program FE	Yes	Yes	Yes	Yes	Yes	Yes

Note: Outcome variable: credits second semester; *study program FE:* study program FE and in the pooled estimations a cohort dummy and its interaction with the study program FE; *running variable (r):* ratio of first semester credits as depicted in the feedback letter to the 80th percentile of credits in the comparison group. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Table 7: Effect of feedback on credits – repeated treatment

	(1) Exp. I	(2) Exp. II	(3) Pooled	(4) Exp. I	(5) Exp. II	(6) Pooled
(a) Credits in 3rd sem.						
Treatment (T)	0.032	-0.239	-0.061	-0.321	0.318	0.020
	(0.686)	(0.717)	(0.496)	(0.918)	(1.167)	(0.720)
T*Above-average				0.948	-0.974	-0.150
				(1.355)	(1.467)	(0.983)
T+T*Above-average				0.626	-0.656	-0.129
				(0.998)	(0.891)	(0.671)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	812	797	1609	812	797	1609
Control Mean	20.35	17.93	19.15			
(Std. Dev.)	(12.54)	(13.23)	(12.93)			
(b) Total credits 3rd sem.						
Treatment (T)	0.319	0.074	0.266	-0.946	-1.034	-0.836
	(1.179)	(1.238)	(0.855)	(1.580)	(2.045)	(1.245)
T*Above-average				3.561	2.010	2.494
				(2.340)	(2.536)	(1.694)
T+T*Above-average				2.614	0.976	1.658
				(1.712)	(1.503)	(1.145)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	812	797	1609	812	797	1609
Control Mean	61.66	56.35	59.03			
(Std. Dev.)	(29.70)	(32.29)	(31.10)			

Note: Above-average and not-above-average (the reference category) refers to the type of feedback students received in the second semester. *Outcome variables:* credits obtained in third semester (Panel a), sum of credits obtained in the first three semesters (Panel b); *strata:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *controls:* HS GPA, credits first semester, GPA first semester, 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: Effect of feedback on credits by pre-treatment expectations – above-average students, Experiment II: Replication

	(1)	(2)	(3)	(4)
Treatment	2.678*	2.952**	1.426	0.701
	(1.528)	(1.274)	(1.634)	(1.147)
Underestimated Performance			-2.841	-3.457
			(2.704)	(2.133)
Treatment*Underestimated			2.797	5.012*
			(3.271)	(2.694)
Treatment+(Treatment*Underestimated)			4.223	5.713**
			(2.794)	(2.410)
Study Program FE	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes
N	110	110	110	110

Note: Underestimated performance indicates students who expected a not-above-average performance in the first semester but then received above-average feedback. *Outcome variable:* credits second semester; *controls:* HS GPA, credits first semester, GPA first semester, 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 9: Effect of feedback on GPA and dropout

	(1) Exp. I	(2) Exp. II	(3) Pooled	(4) Exp. I	(5) Exp. II	(6) Pooled
(a) GPA						
Treatment (T)	-0.012	0.016	0.002	-0.021	0.043	0.002
	(0.021)	(0.020)	(0.014)	(0.029)	(0.038)	(0.023)
T*Above-average				0.025	-0.044	0.000
				(0.041)	(0.044)	(0.029)
T+T*Above-average				0.004	-0.001	0.002
				(0.029)	(0.021)	(0.017)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	767	744	1511	767	744	1511
Control Mean	2.57	2.63	2.60			
(Std. Dev.)	(0.59)	(0.60)	(0.60)			
(b) Dropout						
Treatment (T)	-0.006	0.016	0.004	0.010	0.041	0.022
	(0.020)	(0.023)	(0.015)	(0.030)	(0.047)	(0.026)
T*Above-average				-0.041	-0.046	-0.039
				(0.037)	(0.051)	(0.030)
T+T*Above-average				-0.031	-0.005	-0.017
				(0.022)	(0.020)	(0.015)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	812	797	1609	812	797	1609
Control Mean	0.13	0.16	0.15			

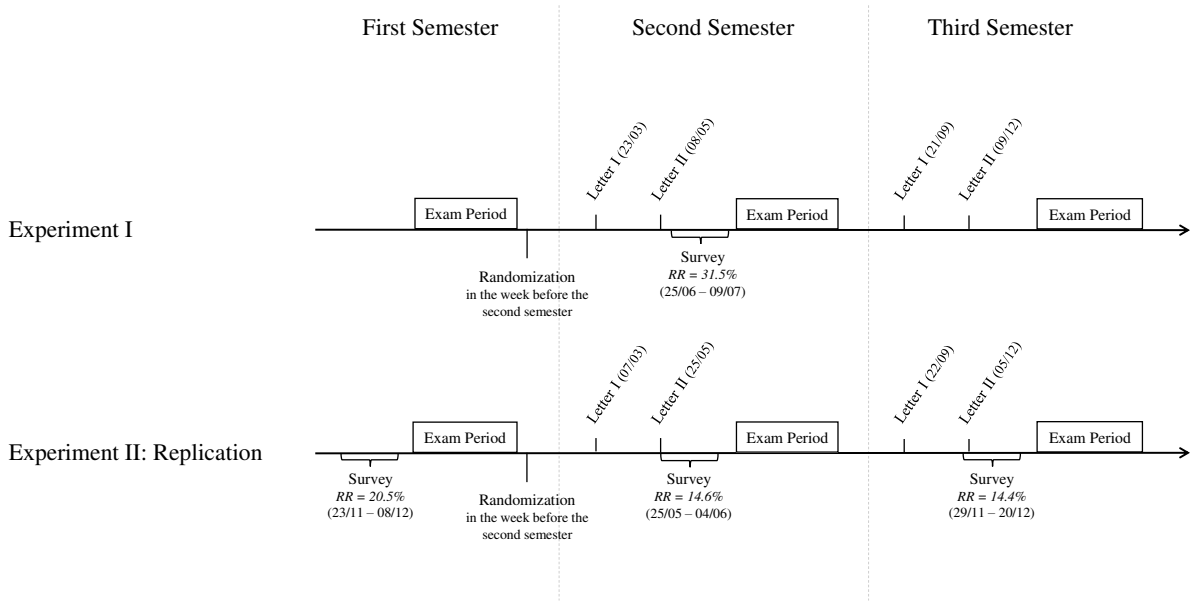
Note: Outcome variables: GPA by the end of the second semester (Panel a; passing grades only; highest passing grade is 1.0, lowest passing grade is 4.0), dropout during or before the second semester (Panel b); *strata:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *controls:* HS GPA, credits first semester, GPA first semester, 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 10: Effect of feedback on well-being – pooled sample

	Satisfaction with				Satisfaction with			
	Life	Studies	Perform.	Stress	Life	Studies	Perform.	Stress
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment (T)	0.058	0.070	0.020	-0.017	-0.061	0.037	-0.015	-0.018
	(0.105)	(0.085)	(0.066)	(0.084)	(0.179)	(0.143)	(0.126)	(0.123)
T*Above-average					0.193	0.054	0.054	0.004
					(0.218)	(0.181)	(0.149)	(0.170)
T+T*Above-average					0.132	0.091	0.038	-0.014
					(0.128)	(0.108)	(0.075)	(0.115)
Study Program FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	362	360	361	361	362	360	361	361

Note: Outcome variables: Outcomes are standardized to have mean zero and standard deviation one; see Table A.7 for the survey questions that are used for the construction of the outcomes; *study program FE:* study program FE, a cohort dummy, and the interaction of the cohort dummy with the study program FE; *controls:* HS GPA, credits first semester, GPA first semester, age, female dummy, time since HS degree, and HS degree Abitur dummy. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

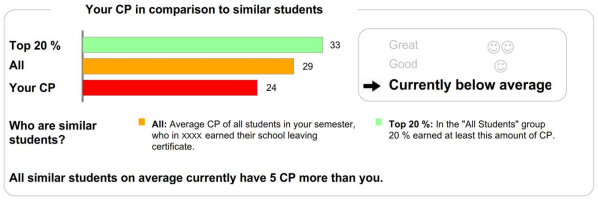
Figure 1: Timeline of experiments



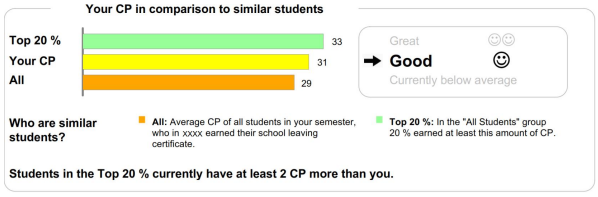
Note: RR = response rate, i.e., the share of our sample that participated in the respective survey.

Figure 2: Relative feedback graphs – treatment group (examples)

(a) No approving frame



(b) Approving frame *Good*



(c) Approving frame *Great*

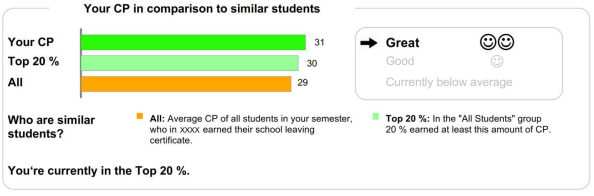
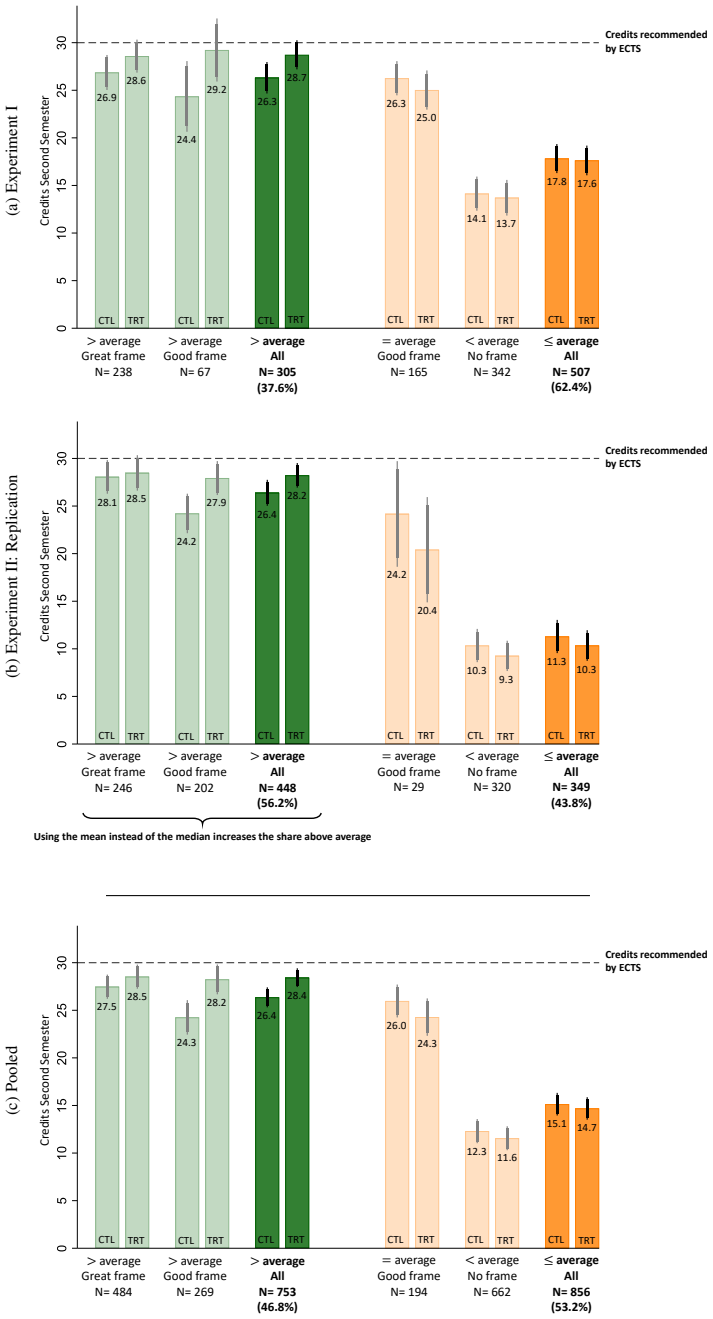
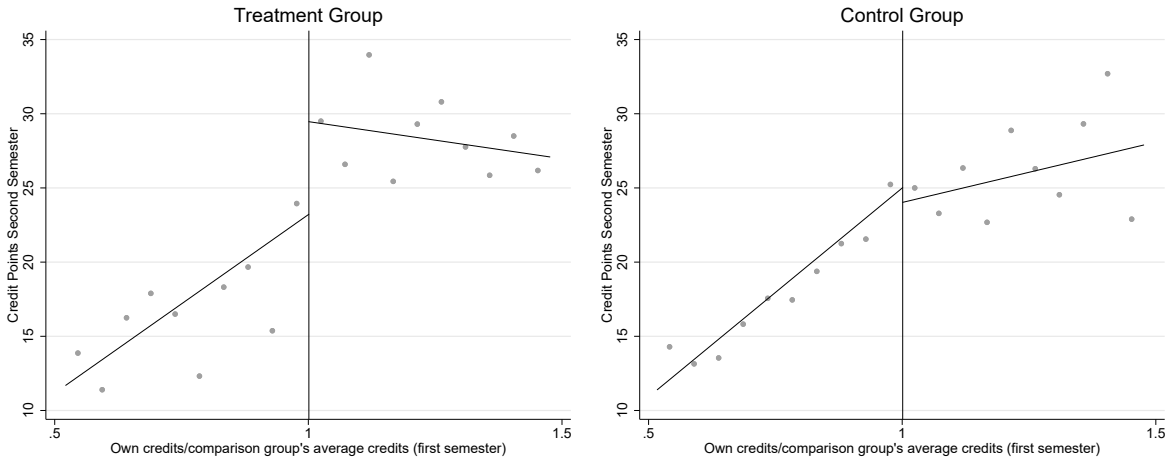


Figure 3: Effect of feedback on credits across feedback types and experiments



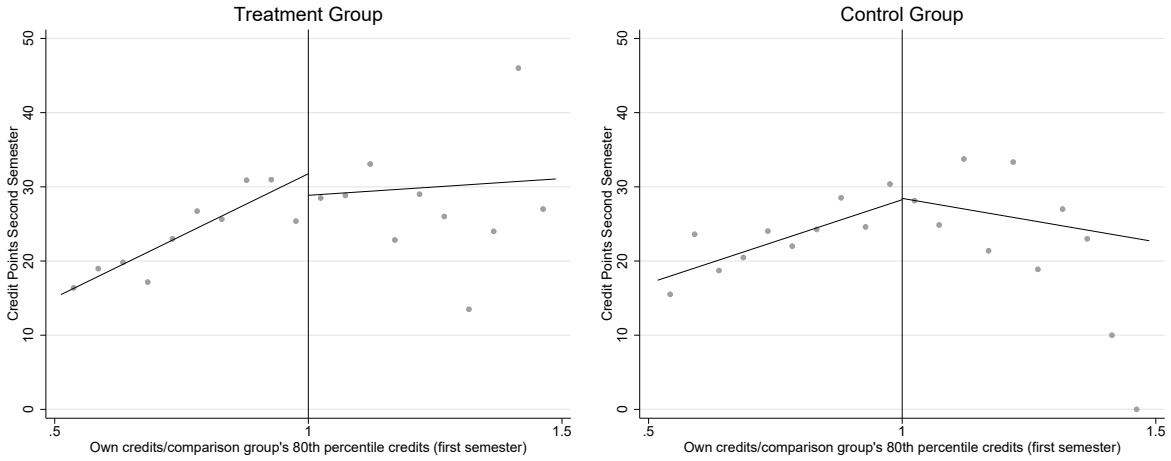
Note: The figures show the raw treatment effects without control variables across the four different treatment types in lighter shading. The bold print and darker shaded treatment effects for *>average* and *≤average* combine the two *above-average* categories and *on-average* and *below-average*, respectively. Panel (a) shows treatment effects for the original experiment, Panel (b) for the replication experiment one year later, and Panel (c) for the pooled sample. In accordance with the ECTS (see Footnote 1) on average students are supposed to pass exams worth 30 credits per semester to finish in the scheduled study duration. 90% (thick) and 95% (thin) confidence intervals are shown.

Figure 4: RD plot at average – first order polynomial, pooled sample



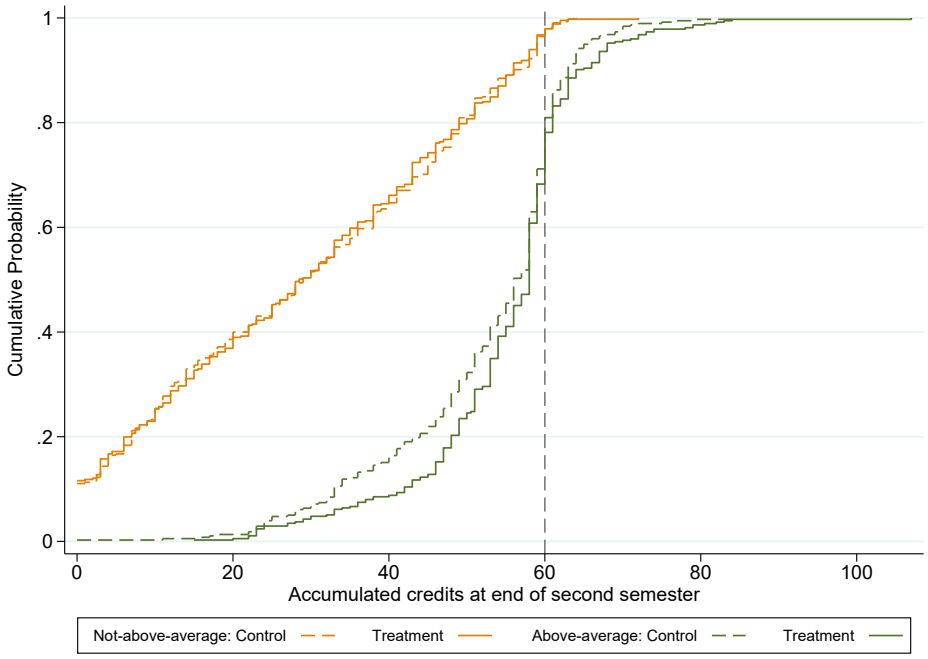
Note: Binned scatterplots using first order polynomials. Running variable is the ratio of first semester credits as depicted in the feedback letter to the comparison group's average credits. Observations on the left side of the cutoff did not place above average. Observations on the right side placed above average.

Figure 5: RD plot at 80th percentile – first order polynomial, pooled sample



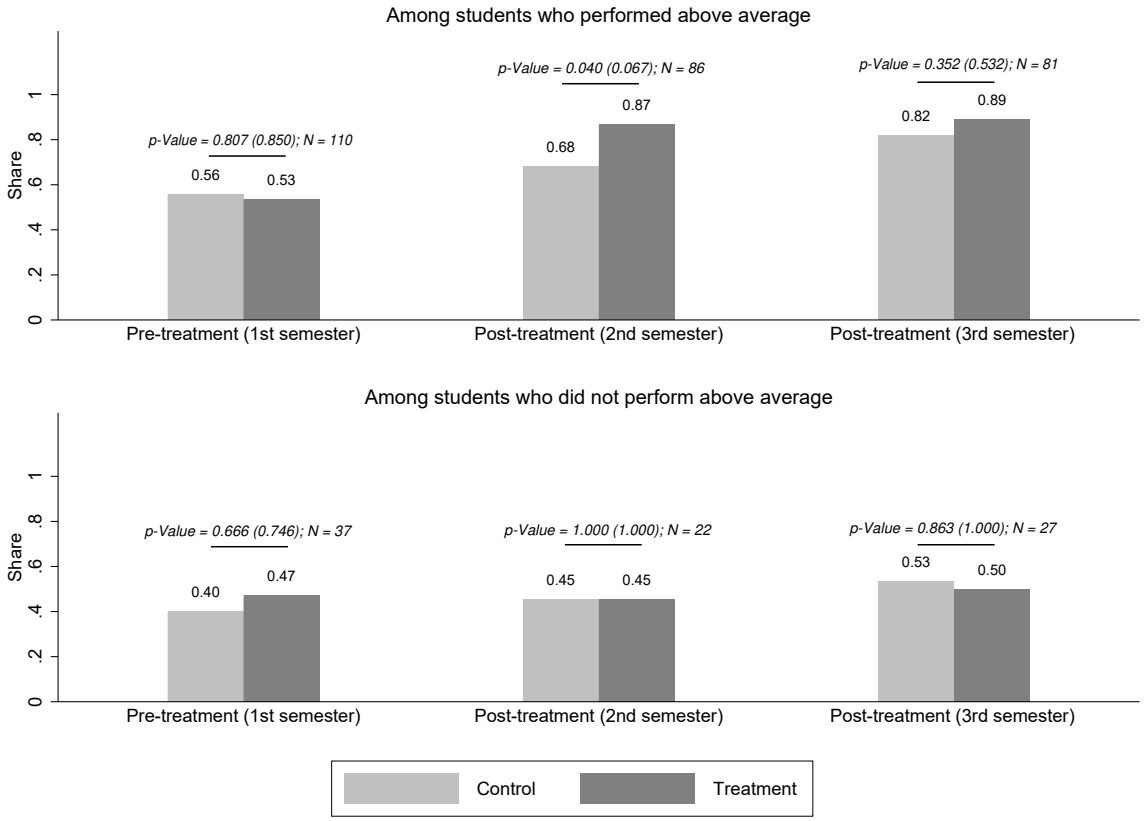
Note: Binned scatterplots using first order polynomials. Running variable is the ratio of first semester credits as depicted in the feedback letter to the 80th percentile of credits in the comparison group. Observations on the left side of the cutoff placed below the top 20%. Observations on the right side of placed in the top 20%.

Figure 6: Cumulative distribution of accumulated credit points at the end of the second semester – pooled sample



Note: The figure plots the cumulative distribution of the accumulated credit points at the end of the second semester by treatment status for students who did not receive above-average feedback and students who received above-average feedback. The vertical dashed line indicates the number of accumulated credit points that students should have obtained at the end of the second semester in accordance with the ECTS (see Footnote 1).

Figure 7: Shares of students who expected to perform above average – Experiment II: Replication



Note: For the 1st and 2nd semester expectations depicted in the figure, above- and not above-average performance refer to the type of feedback students received in the second semester. For the 3rd semester expectations, above- and not above-average performance refer to the type of feedback students received in the third semester. See Table A.5 for the survey questions on students' expectations and Figure 1 for the exact timing of the surveys. p-Values based on Pearson's chi-squared tests. p-Values in parenthesis based on Fisher's exact test.

Appendix

A Additional tables and figures

Table A.1: 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 A.2: Study programs, number of students, and treatment rates

		Observations		Fraction in Treatment	
Study program	Faculty	Experiment I	Experiment II Replication	Experiment I	Experiment II Replication
Business Administration (BuA)	BuA	402	333	50.25%	50.15%
International Business (IB)	BuA	63	59	49.21%	50.85%
Business Engineering[a)] (BE)	BuA	61	63	50.82%	50.79%
Mechanical Engineering (ME)	ME	235	298	50.21%	49.66%
Energy and Building Services Engineering (EBSE)	ME	51	44	49.02%	50.00%
N – Overall		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 A.3: 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 A.2.
Credit strata	Indicating strata based on first-semester credit points.[a)]
Control Variables	
Age	Age in years at randomization.
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 first semester	First semester grade point average (exam-level[b)]); (1=best, 4=worst); failed exams are not included in calculation. Missing values imputed.[a)]
Credits first semester	Number of credit points (exam-level[b)]) obtained in the first semester net of credits granted for an internship.[a)]
GPA at randomization	First semester grade point average provided to us by the university at the time of randomization (module-level[b)]); (1=best, 4=worst); only used in the randomization procedure.
Outcome Variables[c)]	
Credits	Credit points obtained in the respective semester net of credits granted for an internship.
Accumulated Credits	Total credit points accumulated until the end of the respective semester net of credits granted for an internship.
GPA	Grade point average at the end of the respective semester (1=best, 4=worst); failed exams are not included in calculation.
Dropout	Indicator for having dropped out of the study program before or in the respective semester.
Well-being	See Table A.7.

Note: [a)] For details see C. [b)] Exam-level: includes partly completed multiple-exam-modules (= passed sub-modules). Module-level: considers only fully completed modules. For more details see C. [c)] All outcome variables are measured on the exam-level.

Table A.4: Descriptive statistics and balancing properties – above-average students

	Experiment I			Experiment II: Replication		
	(1) Control Mean (Std. Dev.)	(2) Treatment Coefficient (Robust SE)	(3) p-Value	(4) Control Mean (Std. Dev.)	(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[a)]	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[a)]	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)	
GPA at Randomization	2.286	-0.030	0.711	2.392	-0.017	0.753
	(0.711)	(0.081)		(0.624)	(0.054)	
% GPA at Randomization NA[a)]	0.071	-0.001	0.966	0.000	0.008	0.160
	(0.258)	(0.017)		(0.000)	(0.006)	
N	155	150		223	225	

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

Table A.5: Survey questions on students' expectations – Experiment II: Replication

First semester	*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?*
Second/third semester	*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?*

Note: Questions provide the option to give no answer.

Table A.6: Effect of feedback on GPA and dropout – third semester

	(1) Exp. I	(2) Exp. II	(3) Pooled	(4) Exp. I	(5) Exp. II	(6) Pooled
(a) GPA						
Treatment (T)	-0.012	-0.023	-0.016	-0.030	0.003	-0.017
	(0.023)	(0.021)	(0.016)	(0.032)	(0.039)	(0.024)
T*Above-average				0.047	-0.043	0.001
				(0.045)	(0.046)	(0.031)
T+T*Above-average				0.017	-0.040	-0.016
				(0.032)	(0.023)	(0.019)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	771	746	1517	771	746	1517
Control Mean	2.59	2.68	2.64			
(Std. Dev.)	(0.60)	(0.58)	(0.59)			
(b) Dropout						
Treatment (T)	0.003	-0.000	0.001	0.012	0.010	0.010
	(0.022)	(0.025)	(0.017)	(0.032)	(0.048)	(0.027)
T*Above-average				-0.022	-0.018	-0.019
				(0.042)	(0.054)	(0.033)
T+T*Above-average				-0.010	-0.009	-0.009
				(0.027)	(0.024)	(0.018)
Strata	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	812	797	1609	812	797	1609
Control Mean	0.18	0.26	0.22			

Note: Above-average and not-above-average (the reference category) refers to the type of feedback students received in the second semester. *Outcome variables:* GPA by the end of the third semester (Panel a; passing grades only; highest passing grade is 1.0, lowest passing grade is 4.0), dropout during or before the third semester (Panel b); *strata:* credit strata FE, study program FE, and in the pooled estimations a cohort dummy and its interaction with the study program FE; *controls:* HS GPA, credits first semester, GPA first semester, 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 A.7: Survey questions on well-being – second 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	*During the last weeks, how often did you feel stressed out by our studies?*
	[never; rarely; sometimes; often; very often; always]
3	*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...*
3.1	*- not having enough time*
3.2	*- interesting lectures and curriculum*
3.3	*- pressure to perform well*
3.4	*- freedom in organizing my studies*
3.5	*- competition among students*
3.6	*- personal development and growth*
	[1 - completely disagree; 7 - completely agree]
4	*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]
5	*More specifically: How satisfied are you so far with your performance in your studies?*
5.1	*- With my grades, I am...*
5.2	*- With my attained credit points (ECTS), I am...*
	[0 - completely dissatisfied; 10 - completely satisfied]
Estimation Outcomes	
	For the outcomes in Table 10 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:
Life Satisfaction	Question 1
Study Satisfaction	Questions 3.2, 3.6, and 4
Performance Satisfaction	Questions 5.1 and 5.2
Study Stress	Questions 2, 3.1, and 3.3

Note: All questions provide the option to give no answer. For this reason the number of observations in Table 10 can vary depending on the outcome of interest.

Figure A.1: Feedback letter I – treatment group (example)

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

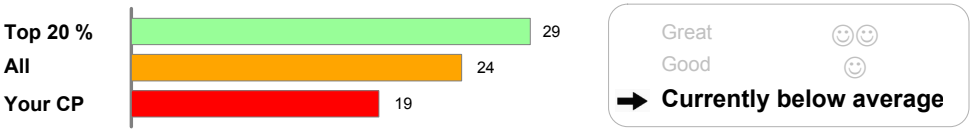
Feedback on your performance in the Bachelor's program International Business

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 **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



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 A.2: Feedback letter I – control group (example)

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

B Robustness of regression discontinuity designs

B.1 RDD at average

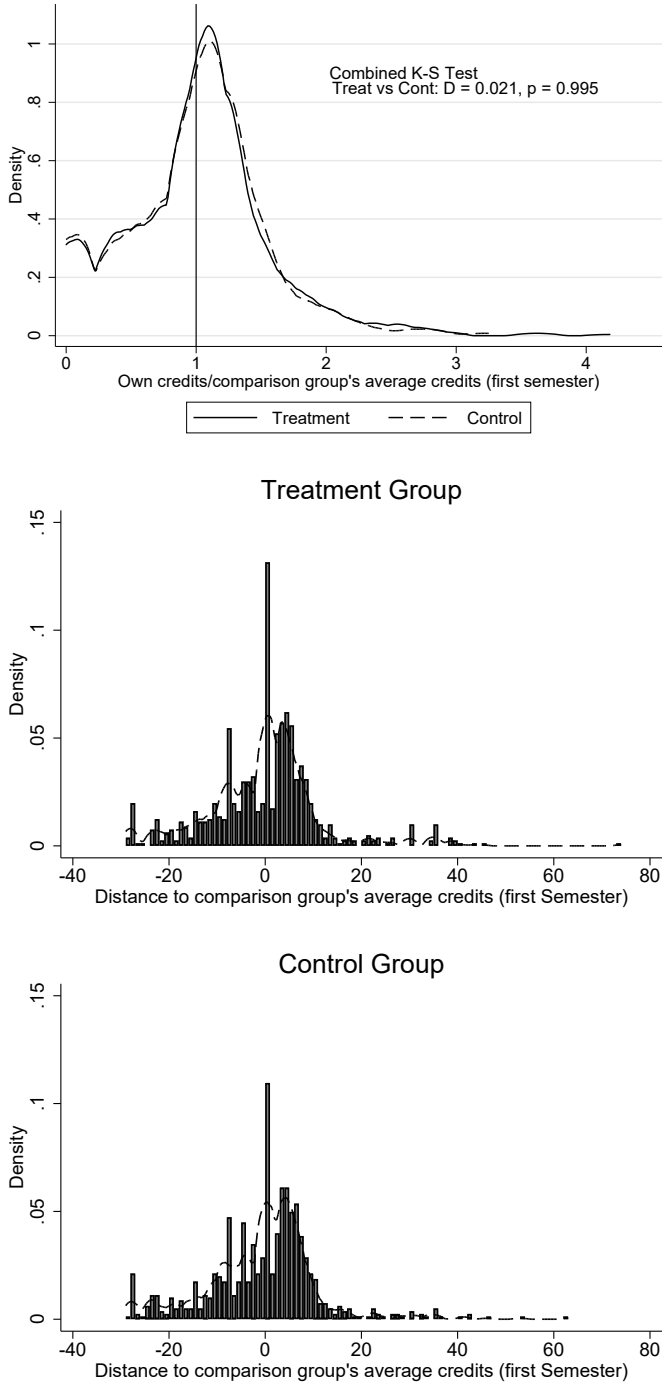
Table B.1: RD estimates at average – different polynomials and discontinuity samples, pooled sample

(a) Treatment Group				
	(1)	(2)	(3)	(4)
	$0 < r < 2$	$0.25 < r < 1.75$	$0.5 < r < 1.5$	$0.75 < r < 1.25$
1st Order Polynomial	6.096***	5.492***	6.418***	8.570***
	(1.214)	(1.243)	(1.450)	(2.063)
2nd Order Polynomial	5.974***	7.379***	7.375***	10.517***
	(1.772)	(1.828)	(2.282)	(3.653)
3rd Order Polynomial	7.263***	6.138**	11.247***	-8.681
	(2.465)	(2.728)	(3.304)	(8.176)
4th Order Polynomial	7.504**	10.563***	12.308**	39.245**
	(3.223)	(3.555)	(5.550)	(18.682)
Study Program FE	Yes	Yes	Yes	Yes
N	700	656	533	352

(b) Control Group				
	(1)	(2)	(3)	(4)
	$0 < r < 2$	$0.25 < r < 1.75$	$0.5 < r < 1.5$	$0.75 < r < 1.25$
1st Order Polynomial	0.189	-0.887	-1.246	-3.599
	(1.354)	(1.465)	(1.743)	(3.304)
2nd Order Polynomial	-2.499	-2.698	-5.206	0.214
	(2.023)	(2.502)	(3.431)	(5.650)
3rd Order Polynomial	-3.790	-4.535	1.476	-14.008
	(3.335)	(4.008)	(5.243)	(10.211)
4th Order Polynomial	-4.245	-2.311	-7.905	26.981
	(4.666)	(5.491)	(7.927)	(23.156)
Study Program FE	Yes	Yes	Yes	Yes
N	699	657	540	342

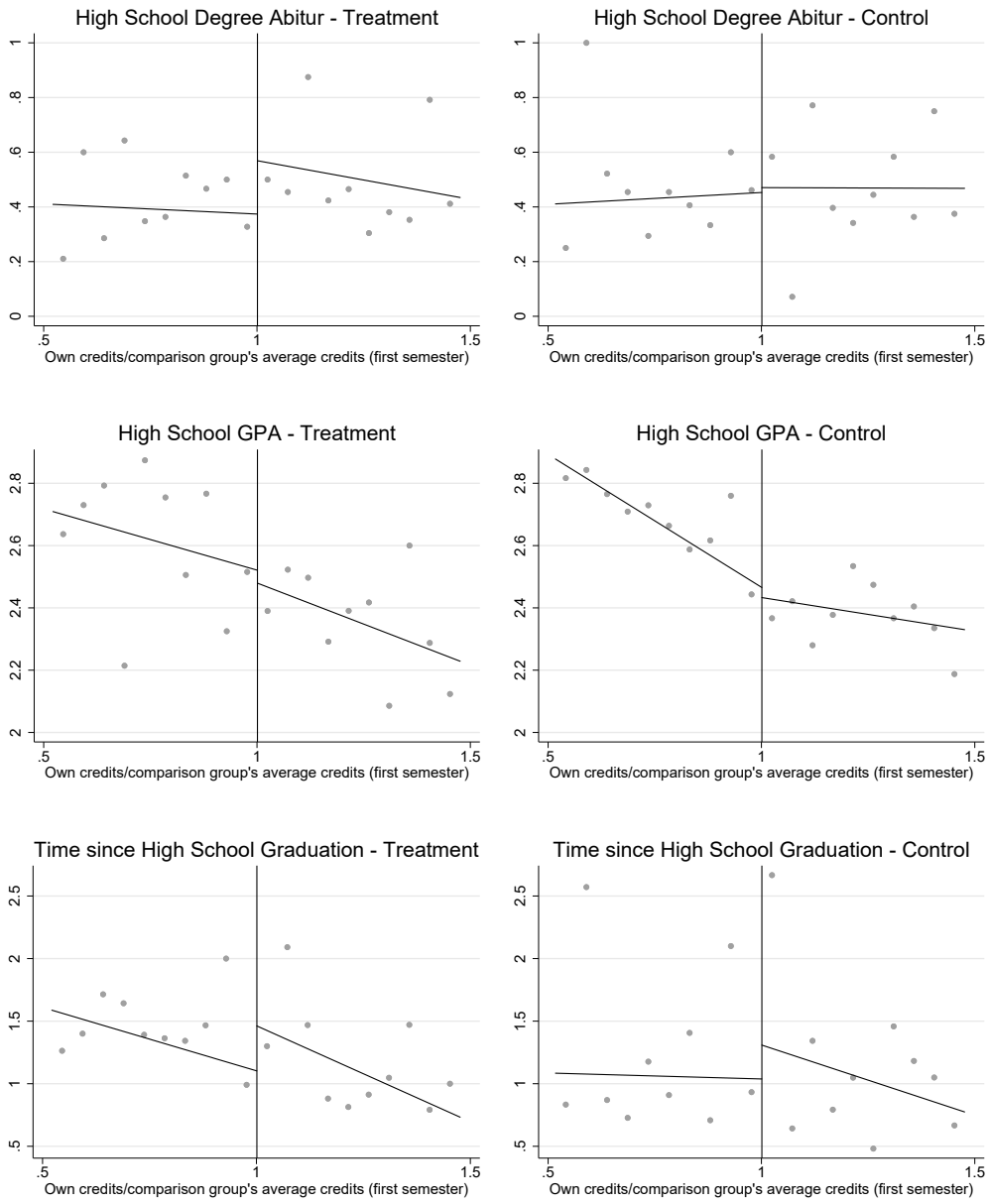
Note: Outcome variable: credits second semester; *study program FE:* study program FE, a cohort dummy, and the interaction of the cohort dummy with the study program FE; *running variable (r):* ratio of first semester credits as depicted in the feedback letter to the comparison group's average credits. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Figure B.1: Distribution of the running variable at average – pooled sample



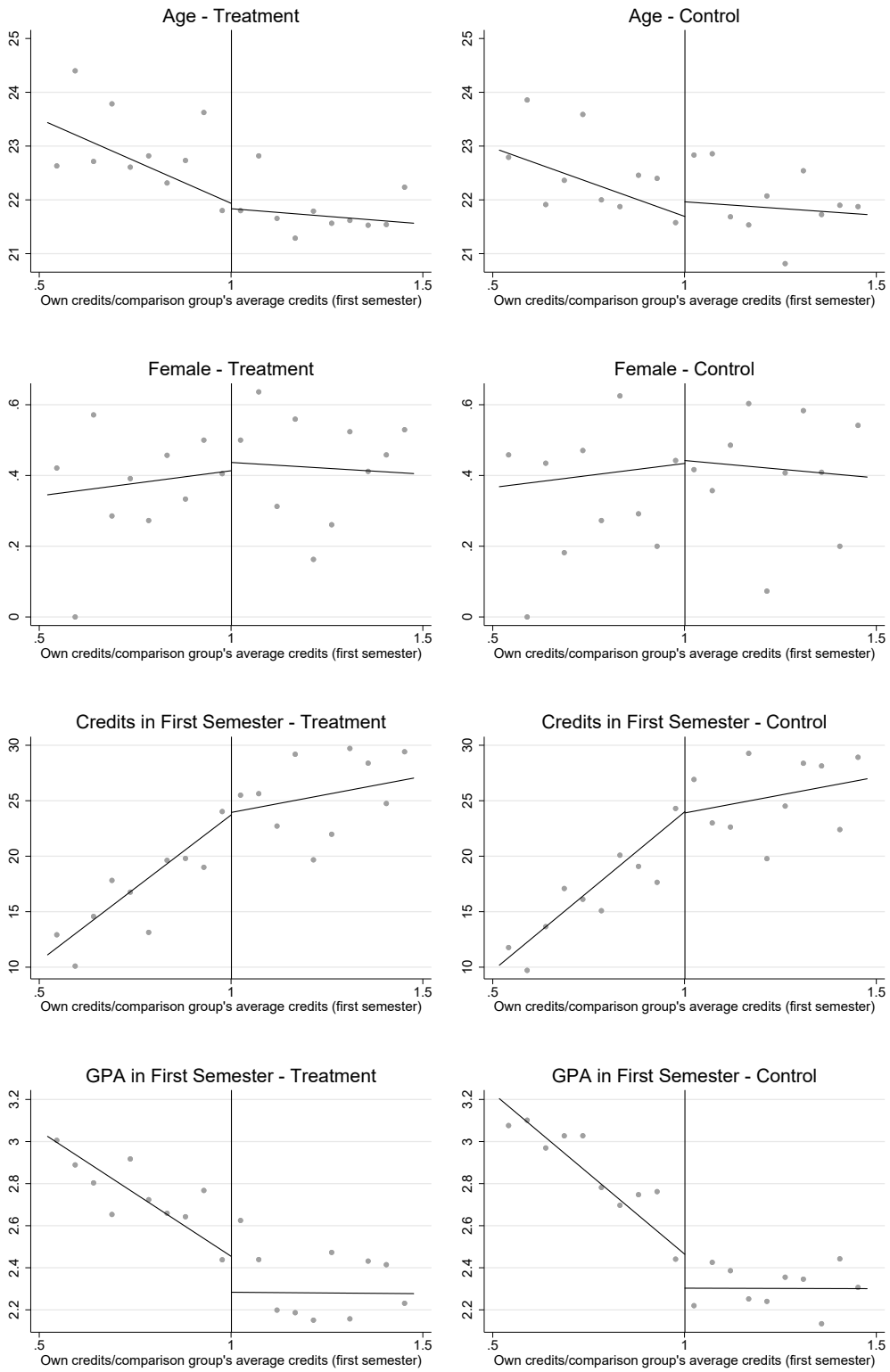
Note: The top panel shows the density of the running variable used for the RDD. Running variable is the ratio of first semester credits as depicted in the feedback letter to the comparison group's average credits. Observations with values lower or equal to 1 did not place above average and observations with values above 1 placed above average. The two bottom panels show the distribution of the distance to the comparison group's average in credit points. Observations with negative values or zero did not place above average and observations with positive values placed above average.

Figure B.2: RD plot at average for covariates – first order polynomial, pooled sample



Note: Binned scatterplots using first order polynomials. Running variable is the ratio of first semester credits as depicted in the feedback letter to the comparison group's average credits. Observations on the left side of the cutoff did not place above average. Observations on the right side placed above average.

Figure B.3: RD plot at average for covariates – first order polynomial, pooled sample (cont.)



Note: Binned scatterplots using first order polynomials. Running variable is the ratio of first semester credits as depicted in the feedback letter to the comparison group's average credits. Observations on the left side of the cutoff did not place above average. Observations on the right side placed above average.

B.2 RDD at top 20%

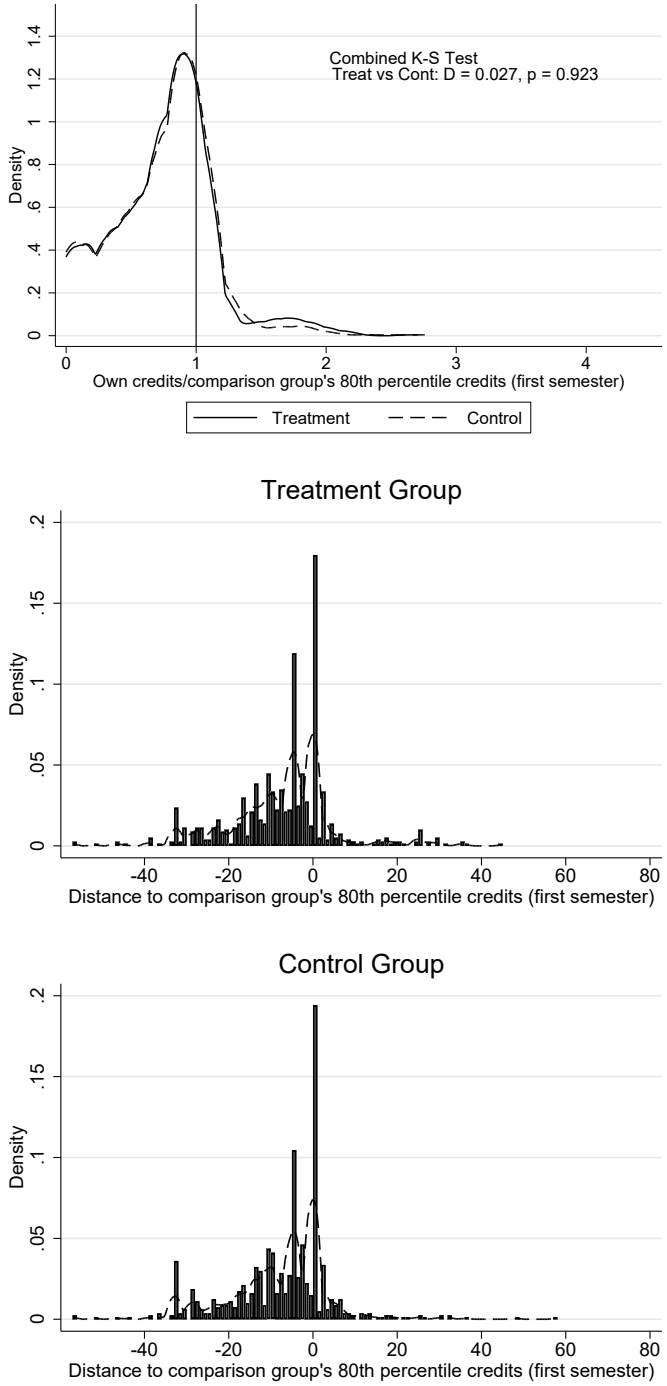
Table B.2: RD estimates at 80th percentile – different polynomials and discontinuity samples, pooled sample

a) Treatment Group				
	(1)	(2)	(3)	(4)
	$0 < r < 2$	$0.25 < r < 1.75$	$0.5 < r < 1.5$	$0.75 < r < 1.25$
1st Order Polynomial	2.134	2.632	0.891	2.969
	(1.569)	(1.720)	(2.038)	(3.129)
2nd Order Polynomial	3.889	2.964	5.941	0.269
	(2.429)	(3.275)	(3.866)	(5.011)
3rd Order Polynomial	3.681	2.556	0.582	7.251
	(3.788)	(3.982)	(4.407)	(7.762)
4th Order Polynomial	0.148	1.990	-8.383	3.371
	(3.931)	(4.686)	(8.173)	(10.507)
Study Program FE	Yes	Yes	Yes	Yes
N	730	667	552	396

b) Control Group				
	(1)	(2)	(3)	(4)
	$0 < r < 2$	$0.25 < r < 1.75$	$0.5 < r < 1.5$	$0.75 < r < 1.25$
1st Order Polynomial	-0.365	0.621	3.256	0.943
	(1.604)	(1.894)	(2.113)	(2.327)
2nd Order Polynomial	1.786	1.657	-1.757	-2.875
	(2.547)	(2.776)	(3.011)	(3.808)
3rd Order Polynomial	1.701	-0.136	-1.709	-21.222***
	(3.457)	(4.250)	(4.159)	(7.934)
4th Order Polynomial	-4.984	-9.215*	-18.092***	-9.113
	(4.480)	(4.760)	(5.671)	(9.029)
Study Program FE	Yes	Yes	Yes	Yes
N	726	669	558	402

Note: Outcome variable: credits second semester; *study program FE:* study program FE, a cohort dummy, and the interaction of the cohort dummy with the study program FE; *running variable (r):* ratio of first semester credits as depicted in the feedback letter to the 80th percentile of credits in the comparison group. Robust standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Figure B.4: Distribution of the running variable at 80th percentile – pooled sample



Note: The top panel shows the density of the running variable used for the RDD. Running variable is the ratio of first semester credits as depicted in the feedback letter to the 80th percentile of credits in the comparison group. Observations with values lower than 1 placed below the top 20% and observations with values equal to or above 1 placed in the top 20%. The two bottom panels show the distribution of the distance to the comparison group's 80th percentile in credit points. Observations with negative values placed below the top 20% and observations with zero or positive values placed in the top 20%.

C Data and methods appendix

The university provides student-level data on accumulated credit points (ACP), the cumulative grade point average (CGPA), individual exam performance, and 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. As described in Section 2.1, we provide students with absolute and relative feedback on their ACP and their CGPA by sending postal letters twice per semester (see Figure 1). The information in the first and in the second letter was mostly identical but for some students changes occured (e.g., during the first treatment semester for 133 (21) students in Experiment I (II) the university updated the information on ACP).[35] These changes appear if (i) exam results were not yet available at the time the first letter was composed, (ii) grades were changed after students inspected their exams, or (iii) due to administrative problems at the university.[36] As a result, a small number of students received different types of relative feedback in the two letters: 15 (17) students in Experiment I (II) no longer had an above-average performance in the second letter although they did so in the first letter, and 10 (12) no longer had a below-average performance. We estimate heterogeneous treatment effects based on the relative feedback types of the second letter, as it provides the most accurate information and because 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 exam or of several exams (sub-modules), all of which must be passed to complete the entire module. Module-level grades are based on the credit-weighted grades of the exams which make up a module. To compute the CGPA the university sums up the product of the grades and credit points of all modules and divides by the ACP.[37] Failing grades do not enter into the CGPA. It is important to note that the university only considers completed modules for the ACP and the CGPA.[38] We refer to the university's approach of accumulating performance measures as aggregation on the *module-level*.

Students can access their personal ACP and CGPA online on a website.[39] As mentioned before, we use the same variables to illustrate our feedback. Although information at the *exam-level* would have reflected their individual performance more accurately as it also includes partly completed multiple-exam-modules, we decided not to use it. The reason was that it would have led to conflicting numbers between the official information on the web and our letters, which could have caused questions and

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

[36] When the first letter of Experiment I was sent the university had not yet calculated the CGPA information for one of the smaller study programs (Business Engineering; N=61).

[37] For GPAs in the Business Administration program, the study regulations require the university to double weight each module scheduled after the first year.

[38] This procedure is in effect across all faculties for the CGPA, 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.

[39] Importantly, in absence of the treatment, the university does not provide any information on the students' relative performance.

potentially also complaints from students.

The variable ACP is defined as zero and the variable CGPA contains a missing value 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 full 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 missing values on the CGPA. In the feedback letters, the latter were replaced 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]". Regarding the ACP we printed the zero and no bar.

Randomization data. In both cohorts, randomization was carried out in the week before the second semester started using demographic information and the individual ACP and the CGPA.

We stratified on study program and ACP, and performed re-randomization (Morgan and Rubin 2012) based on CGPA, 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 ($ACP \leq 12$, $12 < ACP \leq 18$, $18 < ACP \leq 24$, $24 < ACP \leq 30$, $ACP > 30$). In Experiment II we defined ACP strata based on quantiles (Q); four ACP strata in the larger study programs BuA 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 ($ACP < Q_{0.5}$, $Q_{0.5} \leq ACP$). For the randomization in Experiment II, we filled missing values on the variables high school GPA (N=30) and CGPA (N=89) with a constant in order to avoid losing units in the randomization and to balance on the full sample.[40][41] Tables 1 and A.4 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 3 we calculate credits and GPA based on semester-exam-level data. We use the following outcome variables: credit points per semester net of credits granted for internships, accumulated credits net of credits granted for internships, dropout, GPA (excluding failing grades), and survey variables on students' well-being.[42] In contrast to the ACP and the CGPA, the credit points and GPA are now measured on the exam-level, i.e., if students only partly completed a multiple-exam-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 semester, but using the ACP and the AGP as outcome variables could also result in an overstated treatment coefficient.[43]

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

[41] After the randomization, the university was able to provide us with information on the high school GPA of 15 of the 30 missing observations. To use high school GPA as a covariate, we thus only had to impute 15 missing values (see below).

[42] 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.

[43] The upward bias occurs when a module consists of several exams which are taken in different semesters. To calculate the ACP and the CGPA 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

In the regressions we include stratification fixed effects (study program dummies, ACP strata dummies, and a cohort dummy in pooled estimations and its interaction with the study programs), balancing variables (age, sex, high school grades[44], and time since high school graduation), and further control variables (type of HS degree, exam-level first semester credits) as covariates. To keep the number of observations constant across specifications we did not include the CGPA 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 exam-level. The exam-level GPA still has missing values for students who attempted no exams or failed all exams they attempted (54 (66) in Experiment I (II) in the overall sample). 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 constant across estimations.

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). 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 satisfied students 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 studying. 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 cohorts and study programs and in the cases where multiple questions captured the same latent construct, we constructed our outcomes by averaging across the corresponding questions (see Table A.7 for the survey questions and how they were aggregated to obtain the variables used in the estimations). Furthermore, in Experiment II we also gathered pre- and post-treatment information on students' beliefs about their relative performance (see Table A.5).

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 credits. However, the actual performance difference between the two individuals in the treatment semester is only five credits.

[44] 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.