ifo WORKING PAPERS

396
2023

June 2023

Good or Bad News First? The Effect of Feedback Order on Motivation and Performance

Lavinia Kinne

ifo INSTITUTE

Leibniz Institute for Economic Research
at the University of Munich

Imprint:

ifo Working Papers
Publisher and distributor: ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49(0)89 9224 0, Telefax +49(0)89 985369, email ifo@ifo.de
www.ifo.de

An electronic version of the paper may be downloaded from the ifo website:
www.ifo.de

ifo Working Paper No. 396

Good or Bad News First?
The Effect of Feedback Order on Motivation and Performance*

Abstract

How to give feedback in learning environments is a widely discussed topic. I design a field experiment to understand whether the ordering of feedback elements matters for motivation and performance. In random order, university students get one positive and one negative feedback element on their performance in exam practice questions. Students who first receive positive feedback are more motivated to study for the exam compared to those receiving negative feedback first. This effect is driven by a drop in motivation after negative feedback when receiving it first, but not when receiving it second. Furthermore, students adjust their study content to the feedback topics. I find no significant effects of feedback ordering on exam performance overall, but students who first receive the positive feedback perform better if their negative-feedback topic is covered in the exam.

JEL Code: D83, I20, I23
Keywords: Education; feedback; motivation; performance

Lavinia Kinne
ifo Institute – Leibniz Institute for
Economic Research at the University of
Munich, University of Munich
Poschingerstr. 5
81679 Munich, Germany
kinne@ifo.de

* I gratefully acknowledge comments from Ludger Woessmann, Anik Ashraf, Raphael Brade, Luca Braghieri, Cara Ebert, Ingrid Haegele, Johannes Haushofer, Steven Hemelt, Fabian Kosse, Philipp Lergetporer, Jonas Loebbing, Rigissa Megalokonomou, Markus Nagler, Pietro Sancassani, Simeon Schudy, Leonhard Vollmer, Katharina Wedel, Johannes Wimmer, Andrej Woerner, Basit Zafar, Larissa Zierow, and participants at the 2022 International Workshop on Applied Economics of Education in Catanzaro, the 2022 Workshop on Economics of Education at LIfBi Bamberg, the BGPE Meeting 2022, the 2022 TUM Workshop on Field Experiments in Economics and Business, the EALE Conference 2022, the 2023 annual meeting of the Royal and Scottish Economic Society, the 2023 SOLE conference, and the ifo Center for the Economics of Education in Munich. I thank Philipp Braun, Hanna Brosch, Melchor de la Cruz, Elias Farnleitner, Silas Kaatz, Bennet Niederhöfer, Isabella Pau, Hannah Rehwinkel, Katia Werkmeister, and Olivia Wirth for excellent research assistance. The project was funded through an Add-On Fellowship from the Joachim Herz Foundation. IRB approval was obtained from the Ethics Committee of LMU Munich (Project 2021-19) and the experiment was pre-registered in the AEA RCT Registry (AEARCTR-0008953).

1. Introduction

Feedback is a crucial part of any learning environment. For example, in education settings, students repeatedly receive feedback from their instructors that intends to both motivate them and influence their performance. Learning about the assessment of one's performance is an important channel for analyzing and potentially changing behavior (Ammons, 1956; Kluger and DeNisi, 1996; Villeval, 2022), and the way feedback is provided may have a large impact on this mechanism. However, there is little consensus on how to 'best' provide feedback. Various theories and practices have emerged: from the famous 'feedback sandwich' that embeds one corrective feedback into two positive feedback elements (e.g. Davies and Jacobs (1985)) to the mostly criticizing feedback in academic contexts (Allgood et al., 2019; Dupas et al., 2021; Handlan and Sheng, 2023).

A potentially meaningful feature of feedback that is easy to adapt is the ordering of feedback elements. Classic economic theories emphasize the information content of feedback which helps to reduce uncertainty, e.g. about one's own performance. Individuals adjust their performance beliefs which is mostly assumed to follow Bayes' rule: prior and new information are combined to attain an assessment of one's performance. If that was the only aspect of feedback, the ordering should not matter as long as the information content of the differently ordered feedback elements is the same. Instead, theories from psychology support potential effects of feedback ordering due to emotional reactions to feedback (Choi et al., 2018) as well as attributing the feedback to the self rather than the performance on the task (Kluger and DeNisi, 1996).

In this paper, I conduct a randomized field experiment to study whether the ordering of positive and negative feedback elements matters for the motivation and performance of university students. Each treated student receives one positive and one negative feedback part on their performance in exam practice questions. The feedback refers to subtopics of the practice questions such that the positive (negative) feedback element informs students about their individually best (worst) subtopic, independently of their ranking compared to other students. Besides the respective topic, the wording of both feedback elements is identical for all students. I randomly vary the ordering of the two feedback elements across the treatment groups to test whether the order affects students' motivation to study for the exam and their performance in subsequent practice questions and the exam. I focus on motivation and performance as outcomes due to their importance for later labor market outcomes such as drop out from education, unemployment, and wages.[1]

[1] A large literature shows that higher educational performance is associated with higher wages and lower unemployment (Zax and Rees, 2002; Hanushek et al., 2015). Similarly, evidence from education science shows how intrinsic motivation for a specific activity is negatively related to drop out from (higher) education (Gillet et al., 2012; Cabus and De Witte, 2016; Rump et al., 2017). In economics, seminal work from Bénabou and Tirole (2002) contextualizes how self-confidence and motivation interact.

I find that students who first receive positive feedback are more motivated to study for the exam compared to those who received negative feedback first. To understand how this ordering effect emerges over the course of the experiment, I compare both treatment groups to the pre-treatment average motivation of a control group that does not receive any feedback on the practice questions. The difference in post-feedback motivation between the *positive-negative* (POSNEG) ordering and the *negative-positive* (NEGPOS) ordering is driven by a drop in motivation for the NEGPOS group after the negative feedback part that persists after they have also received the positive feedback element. This drop in motivation is not observed for participants in the POSNEG group who receive the negative feedback as a second element, which points to a shielding effect of the positive feedback when receiving it first. For either treatment group, positive feedback does not affect motivation compared to the initial motivation of the control group.

The effect of feedback ordering on motivation in the two treatment groups cannot be explained by their belief updating, i.e. the information updating process highlighted in economics. There is no significant effect on post-treatment beliefs about performance in the practice questions. Compared to the control group, performance beliefs follow closely the patterns from 'motivated beliefs' where positive feedback is over-weighted compared to negative feedback, although not differentially by feedback order. Instead, overall feelings about the feedback, elicited one day after the exam, are much better for students from the POSNEG group compared to the NEGPOS group. This suggests that the effects on motivation might be driven by emotional responses to the feedback received as suggested by theories from psychology (see e.g. Jacobs et al. (1973)).

Furthermore, I find that students strongly react to the feedback topics and adjust their study content accordingly. For both treatment groups, a clear pattern emerges compared to the pre-treatment allocation of study topics of the control group: students persistently remove course topics from their study list as soon as they receive the positive feedback on them. Similarly, students add chapters to their study list when receiving negative feedback on them. This happens more strongly for the negative-feedback topic than for the positive-feedback topic and students are slightly more cautious about removing the positive-feedback topic from their study list if they have received negative feedback first. These findings imply that students understand the content of the feedback and consider it relevant. The pattern can still be observed when asking students about their actual study content one day after the exam, although slightly weaker in magnitude. I find no effect on study hours on the remaining days before the exam.

Finally, there is no significant overall effect of feedback ordering on performance in the final exam. This might reflect hurdles in translating the effects of feedback ordering on motivation to students' exam performance. When having a look at the content of

Through motivation, self-confidence can improve individuals' perseverance and thus compensate for imperfect willpower.

the exam, I see differences in students' performance depending on which course topics were part of the exam. More specifically, I find a positive effect of the POSNEG ordering on exam grades for students who encounter their negative-feedback topic in the exam. For these students, it might be especially important to have avoided a drop in study motivation from negative feedback when dealing with this course topic in the exam.

The results of feedback ordering on motivation suggest that the information content of feedback might not be the only mechanism in place. In this setting, the informativeness of the feedback is the same for the two treatment groups since students in both groups receive feedback on a topic they performed best in as well as worst. The resulting treatment effects on motivation and post-exam feelings about feedback imply a more emotional or impulsive reaction that leads to a different perception of positive and negative feedback elements depending on their ordering. This is supported by the psychological literature on how feedback can cause emotional reactions (see e.g. Erickson et al. (2021)) that in turn can shape (cognitive) behavior (see e.g. Zadra and Clore (2011), Baumeister et al. (2007), and Tyng et al. (2017)). The findings hence add a new angle to the interpretation of feedback in economics, suggesting that individuals react to more than the information content of feedback.

The way feedback is provided in this intervention has a series of advantages that might nicely complement other commonly used types of feedback. Highlighting individual strengths and weaknesses can be crucial for personal development, especially at young ages. It might help students to develop an assessment of their capacities that is independent of others, which in turn can make it easier to adapt to new settings with different peers. More specifically, informing students about the course topics they are currently performing best and worst in when studying for an exam, could be helpful in terms of study efficiency. If we consider studying for an exam to be an optimization problem of how to best allocate study time and effort under a constrained time budget, advice on where to invest more time might be as useful as information on where to take that time away. Similarly, the within-person feedback used in this study can reduce inequality in the direction of feedback received by students: receiving relative performance feedback usually implies that students in the lower part of the distribution hardly ever receive positive feedback. Most likely though, these students also have strengths: they might either be low-performers just with respect to a specific peer group or have relative strengths in certain areas versus others. Hence, information on personal strengths and weaknesses can help especially lower achievers to make better decisions for future life choices. Furthermore, this within-person feedback can be an alternative solution when no clear reference group can be identified.

This paper mainly contributes to three strands of the literature. First, it adds to the literature on feedback interventions in education. A series of papers finds positive effects of absolute and relative performance feedback on performance at university (see

e.g. Bandiera et al. (2015) in the UK, Brade et al. (2022) in Germany, Kajitani et al. (2020) in Japan, and Dobrescu et al. (2021) in Australia), in secondary education (Azmat and Iriberri, 2010; Fischer and Wagner, 2018; Goulas and Megalokonomou, 2021), and in primary education (Muis et al., 2015; Hermes et al., 2021).[2] The contribution of my paper to this literature is twofold. First, I look at the *ordering* of feedback elements as a specific feature of how to provide feedback. Second, I provide feedback in a novel way compared to the literature by telling students how well they performed on specific topics within an overall performance. By moving away from peer-related relative performance feedback which is highly dependent on the reference group, I can furthermore study the effects of positive (negative) feedback on low (high) achievers while still providing truthful and informative feedback.

The second related literature about feedback is the one from behavioral economics. Eil and Rao (2011) show how individuals differ in the way they incorporate feedback on ego-relevant dimensions such as intelligence and beauty depending on whether it aligns with their prior assessment. Möbius et al. (2022) find similar evidence of such 'motivated beliefs' for undergraduate students, again especially in ego-relevant dimensions. Both studies show that individuals tend to over-weight positive feedback relative to negative feedback. In an emerging part of this literature, Zimmermann (2020) has highlighted the dynamic aspect of motivated beliefs. His findings confirm that individuals react more to positive feedback which becomes even more evident after some time passes. Following up on this approach, Coffman et al. (2021) document large and persistent gender gaps in beliefs and choices and show that women react more strongly to negative feedback compared to men. I add to this literature in two ways: first, my experiment looks at both the effects of feedback elements and their ordering. This is combined with a dynamic perspective on individuals' reactions to the separate feedback elements. Second, I provide evidence from a field experiment in a real-world setting as opposed to the laboratory context that has been used in most of the previous literature. This is especially important for education as one of the crucial periods not only for human capital accumulation leading to far-reaching later life decisions but also for the development of one's personality.

Lastly, the paper builds on previous work from psychology. Feedback has been a topic of interest in this literature for a very long time, reaching back to first important theoretical contributions from Thorndike (1913, 1927).[3] The first empirical studies on

[2]The exception to these largely positive effects are the studies by Azmat et al. (2019) and Bursztyn and Jensen (2015). Azmat et al. (2019) find short-run negative effects of providing feedback on college students' performance in Spain for those who initially underestimate their performance. Bursztyn and Jensen (2015) show that publicly revealing top performances makes the best high school students reduce their effort to avoid being exposed to their peer group in an L.A. low-income setting.

[3]Further important theoretical contributions about the impact of feedback on individuals came from Ammons (1956), Ilgen et al. (1979), and Kluger and DeNisi (1996). For example, Kluger and DeNisi (1996) highlight how the effectiveness of feedback decreases when it is given or perceived as closer to the self than the task.

feedback ordering have been conducted by in the 1970s and 1980s (Jacobs et al., 1973; Schaible and Jacobs, 1975; Davies and Jacobs, 1985), but these studies were conducted on very small and/or selected samples or only included a subset of potential feedback orderings such that the resulting evidence is only partially conclusive. More recently, empirical studies have focused on the 'feedback sandwich', with mixed results. For example, Slowiak and Lakowske (2017) and Henley and DiGennaro Reed (2015) show that there was no differential impact of different sandwich orderings on performance, but participants rather preferred the non-sandwich version with corrective-positive-positive feedback. Choi et al. (2018) provide evidence on feedback ordering in a work-related lab experiment and show how emotional responses can play a role in reactions to feedback. Whereas most of these approaches looking at the feedback sandwich cannot isolate the effect of the presence of more than one positive or negative feedback element, my study uses a clean design only using one positive and one negative element, in randomly varying order. Furthermore, it brings the analysis of feedback ordering to a field setting in education where feedback is especially relevant since students are constantly accompanied by different types of feedback givers throughout their education.

The remainder of the paper is structured as follows: section 2 details the experimental design whereas section 3 gives an overview of the sample and descriptive statistics. Section 4 shows the main experimental results, section 5 concludes.

2. Experimental Design

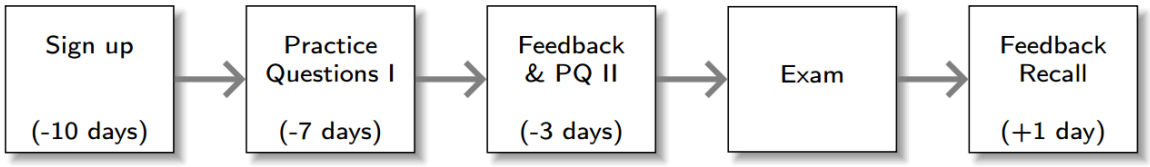
The field experiment was implemented in the setting of university courses that undergraduate Business and Economics students have to take at the University of Munich. I set up an additional exam preparation that students could sign up for. Students from two courses participated in February 2022 which was the exam period of the Winter Semester 2021/2022.[4] All information and questionnaires were presented in German according to the course language.[5]

The flow chart in figure 1 shows the overall sequence of the study design. The following subsections explain each of the parts in detail.

[4]Both courses were related to basic methods and concepts that are taught in the first semesters of a typical undergraduate Business and Economics program. Exams were concentrated in a period of approximately two weeks and are usually spread out as evenly as possible for those students who are taking the exams within the regular schedule of the respective major. Nonetheless, students would have multiple exams during a week, sometimes on subsequent days, which would impede the adjustment of their study schedule and hence attenuate my results.

[5]As the experimenter, I was neither involved in teaching nor in grading the courses. Students were informed multiple times that I was providing this service as an external person and that they could not infer anything about the exam content from my practice questions.

Figure 1: Overview of the study design



2.1. Sign up

The opportunity to participate in the study was announced both during (online) lectures as well as via email three to four weeks before the exam. Students were informed that they had the chance to practice their knowledge in exam-type questions and that they would receive personalized feedback on their performance. After the deadline for signing up (10 days before the exam), students received an email with the outline of the exam preparation. This email included the dates for the additional exam preparation questions (*seven* and *three* days before the exam), for the feedback I would provide (*three* days before the exam), and for a short post-exam questionnaire (*one* day after the exam). It was specified that only those who participated in all three questionnaires had the chance of winning a €50 voucher to be chosen from the platforms Amazon, Avocadostore, and Netflix. Around 20% of participants received such a voucher in the end. Finally, students were informed that they would be contacted again once they had received the grades for the exam and that each student who would then report the grade together with some proof (e.g. a screenshot of their student portal) would receive an additional €10 voucher to be chosen from the same platforms.[6]

2.2. First set of practice questions

Seven days before the exam, students received an email with a link to the first questionnaire including the first exam preparation questions. Prior to the practice questions, students answered a series of questions on some background characteristics as well as personality traits. The former contained questions on age, gender, semester, field of study, their high school GPA and last high-school math grade as a proxy for ability, their parents' highest education and employment status as a proxy for socio-economic status, and whether students and their parents were born in Germany as a proxy for culture. The surveyed personality traits include patience, risk-aversion, the big five personality traits, self-efficacy, and feedback aversion. Furthermore, I elicited

[6]Grade reporting was voluntary since the data protection guidelines of the university do not permit the use of personalized administrative data.

students' general motivation for their studies. Finally, participants stated how many hours per week they had spent so far studying for this course.

I elicited such a rich set of background characteristics for two main reasons. First, some of the characteristics were ex ante of interest as dimensions of heterogeneity based on results from other studies as well as theoretical considerations. This included gender, ability, socio-economic background, and personality traits. Second, since I could not predict how many students would sign up for the experiment, I wanted to make sure that the design itself would help to increase statistical power. To this end, I asked for the characteristics that should theoretically predict the outcomes well such that I would then be more likely to detect a change in outcomes caused by the treatment.

Furthermore, all later outcomes were elicited during this first questionnaire. This again was intended to help with statistical power in the later analyses since initial levels of the outcomes are most likely very good predictors of the final outcomes and can hence should improve the precision of the estimates.

One main outcome is motivation to study for the exam. It was elicited using a survey question with the following wording: 'How motivated are you to study for the exam of [name of course]?'. Students could choose one value from a five-point scale: 1 'not motivated at all', 2 'rather not motivated', 3 'neutral', 4 'rather motivated', and 5 'very motivated'. The same wording and scale was used every time this outcome was elicited.

In addition, students were asked about their study plan. This first elicited how many hours they planned to dedicate to studying for this course on each of the remaining days before the exam. Furthermore, students were presented with the full list of course chapters as outlined on the course syllabus. From these, they had to choose exactly three topics they wanted to focus on in their remaining study time. Due to the design of the feedback, these could later be connected with the feedback topics (see detailed explanation in section 2.3).

Lastly, students reported which performance they expected for the exam as well as the upcoming practice questions. Each assessment of students' performance beliefs contained two elements: absolute and relative expected performance. The absolute performance belief was elicited for exam grades on the German scale (from 1.0 to 5.0) and for the total points obtained in the first set of practice questions (out of 20). For the expected relative performance, students had to provide probabilities adding up to 100 for being placed in each quartile of the later performance distribution. For each expected absolute performance, I also elicited a self-evaluation, i.e. an assessment of how good or poor students would consider this performance on a five-point scale, as in Exley and Kessler (2022): 1 'very poor', 2 'poor', 3 'neutral', 4 'good', and 5 'very good'.

After this first block of the questionnaire, I elicited pre-treatment performance via the practice questions that students had to answer to receive feedback. Participants were

presented with five open questions about exam-relevant content. These were adapted to the pandemic situation with online exams and allowed for the same conditions as the later exam (e.g. open book and formula sheets). Students had 30 minutes to answer the practice questions, a visible timer ensured that students would stick to this time frame. All questions were presented on the same screen and contained a brief description of the task itself and an open field where students could type in their answer. Once students had submitted their answers or time had run out, they were forwarded to a last screen that elicited beliefs about their performance in the practice questions they had just answered (again in points out of 20). This measure will be used as a pre-treatment performance belief since it is a much more informed assessment than the one provided prior to working on the questions. For motivation and the study plan, pre-treatment outcomes will always refer to the answers provided before answering the practice questions.

The practice questions were then graded by a group of research assistants whom I provided with sample solutions and instructions for grading. A total score (out of 20) as well as an overall ranking of students was calculated. Furthermore, each of the practice questions referred to a specific chapter of the course outline, such that students also received a sub-score for each topic. According to these sub-scores, a within-student ranking of topics was created, resulting in an assessment of which topic students had answered best as well as worst. These topics would then be used for the positive and negative feedback element.[7]

2.3. Feedback and second set of practice questions

The core part of the experiment was the feedback treatment that followed four days after the practice questions, i.e. three days before the exam. Students again received a (personalized) link to visit a page with an individualized structure (see table 1). Students were randomized into three groups: a control group that did not receive any feedback at this point and the two treatment groups with positive and negative feedback in varying order. Randomization was performed on the individual level (by course) after receiving the answers to the first set of practice questions. Also, randomization was stratified by gender since ex-ante this was the main dimension of interest in terms of heterogeneity.

Participants from the control group did not receive any feedback on the practice questions performed four days before. These students were immediately directed to

[7]In case of equality of scores for the best and/or the worst topic, one of them was chosen randomly, depending on whether students put in any text at all. This affected 65 students, i.e. 28.9%, for the best topic and 118 students for the worst topic (52.4%), which amounts to a total of 166 students affected (73.8%). Only 67 students had two or more best or worst topics (29.8%), 19 students had at least three best or worst topics (8.4%). For students who left all questions blank, it happened that the best and the worst dimension coincided. In this case, randomization of best and worst topic for feedback was repeated until the topics were different. The latter only applied to two students of whom one did not participate in the feedback round. Excluding the other student does not change the results. The procedures were applied to students from both treatment groups and the control group.

questions about performance beliefs (including a self-evaluation) and motivation before moving on to the second set of practice questions. Since the main focus of this study was the effect of feedback ordering on the respective outcome, the control group was undersampled and comprised of 20% of participating students. Its main purpose was to observe time trends such as study progress that were independent of feedback as well as to replicate results from the literature showing the effect of feedback compared to no feedback.

Table 1: Overview of the feedback questionnaire

			motivation II		**motivation III**	
group	share	feedback I	*beliefs II* *study plan II*	feedback II	*beliefs III* *study plan III*	practice quest. II (**performance II**)
T1	0.4	positive	x	negative	x	total points
T2	0.4	negative	x	positive	x	total points
C	0.2	no feedback		no feedback	x	total points

Notes: Group shares and ordering of elements in the feedback questionnaire.

The remaining participants were equally distributed between the treatment groups *positive-negative (POSNEG)* and *negative-positive (NEGPOS)*. These students first saw their personal feedback before moving on to the second set of practice questions. Students were informed that their performance on the first set of practice questions had been corrected and evaluated. Then, they were presented with the first feedback element which was either positive or negative depending on their treatment group. According to the individual within-person ranking described in section 2.2, the positive (negative) feedback element would inform students about the topic they had performed best (worst) in. The feedback was communicated in the following way: 'According to all topics evaluated on your performance in the first set of practice questions, you did best/worst on XXX', where XXX stood for their personal best/worst out of the course topics. For the positive part, an additional sentence was added: 'This is your personal strength, great!'. Similarly, for negative feedback: 'This is your personal weakness, what a pity!'. Furthermore, a green (red) thumbs up (down) was added to make the positive and negative nature of feedback more salient. Examples of how positive and negative feedback elements looked like can be seen in figures B.1 and B.2 in the appendix (translated from German). The positive and negative feedback parts were split onto separate screens and would be seen at a distance of a few minutes.

Between and after the two pieces of feedback, participants from the two treatment groups were asked about all relevant outcomes considered in the later analyses. This included their current motivation to study for the exam assessed by the same survey question as in the first questionnaire, their performance beliefs as elicited four days before, and their study plan (hours and topics). Additionally, students were asked about their feelings regarding the single feedback elements. More specifically, they were asked after

each feedback element: 'How do you feel about the feedback you just received?' and had the answering options 'very bad' (1), 'bad', 'neutral', 'good', and 'very good' (5).

After receiving feedback and responding to questions on all outcomes, all students answered an additional set of practice questions. The provision of such a second set of practice questions had the goal of being able to measure any potential treatment effects on immediate performance. This is an interesting outcome in itself and also stands in contrast to the exam a few days later: students didn't have time to prepare for these questions after receiving feedback, but they were also associated with relatively low stakes. The questions consisted of 10 multiple choice questions related to the exam content. Each question had a time limit of 60 seconds in order to be more similar to the later exam situation under time pressure.[8] Students could choose exactly one out of three options for each question and would receive one point for each correct answer. No penalties for wrong answers were applied. Each multiple choice question was presented on a separate screen that included a timer indicating the remaining answer time.

Lastly, all students were automatically informed about their score in the second set of multiple choice questions. After responding to the last question, participants were forwarded to a screen thanking them for their participation and informing them about their total number of correct answers out of 10.[9] Hence, students from the control group in the end also received some feedback but only on their total number of points in the second set of multiple choice practice questions. This implies that comparisons of any outcome measured after this point in time between the treatment groups and the control group measure the effect of additional, more detailed feedback for the treatment groups.

2.4. Post-exam questionnaire

One day after the exam, students received the link to a post-exam survey. It contained questions about their exam performance beliefs (including a self-evaluation), the study plan they had actually implemented (hours dedicated to studying for this exam on each of the six days before the exam and topics on which they had focused), and their perception of the difficulty of the practice questions provided in the experiment compared to the exam. It also asked about the perceived usefulness of the practice questions as an exam preparation, a recall of the feedback, questions about how they felt about the feedback in general, and how useful they thought the feedback was.

These outcomes were of interest by themselves, but also had the aim of uncovering potential mechanisms of treatment effects. Whether students remember the feedback

[8] In one of the two courses, there were only multiple choice questions in the later exam, even though not as time-constrained. The later exam of the other course consisted mostly of questions as presented in the first set of practice questions. This information could be inferred from previous exams held while teaching online during the pandemic and was hence known to me when designing the treatment.

[9] A large majority of participants still remembers their score in these multiple choice questions after the exam as will be discussed in section 4.5.

and its ordering might play an important role for how immediate effects could translate into medium-term effects related to the exam. Furthermore, students' feelings about the feedback in the longer run might contain an emotional component of feedback that would be expected from psychological theories. Similarly, if students perceived the practice questions or the feedback as not very useful, this might explain their reactions to the feedback. Finally, their actually implemented study plan gives a glimpse into their behavior between the feedback intervention and the exam. This might be an important explanatory factor for any effects on exam grades.

As soon as official grades were released, students were contacted again. They were asked to report their course grade together with some proof, e.g. a screenshot of their transcript. Each student who reported her grade received a €10 voucher after I had checked the grade proof.

Exam grades are an interesting outcome for a series of reasons. First, they are a high-stake outcome for the students. This is especially true in this setting where students are at a very early stage of their studies and hence receive one of the first external assessments of their suitability for their university studies. Considering that this intervention took place during the pandemic where students had only experienced online teaching, additional feedback might have been especially valuable to them and hence may have had particularly strong effects. Second, compared to the immediate performance in the multiple choice questions answered right after the feedback, the exam provides a medium-term learning outcome. Students had another three days after the feedback to study for the exam. In these remaining days, they might have adjusted their study plan on two margins: quantity and quality. Adjusting the quantity would imply increasing or decreasing the study hours while the quality or efficiency of study could be reflected in the content they focus on while studying.

3. Sample and Descriptive Statistics

This section gives an overview of the participants of the study, including their descriptive statistics and balancing checks for background characteristics and pre-treatment outcomes.

Table 2 shows the number of participants and the corresponding attrition rates between the different stages of the experiment, by course. The main sample for the analysis consists of 225 students who participated in the feedback intervention (see column 5).[10] Although attrition rates were rather low from one stage to another, the

[10]These 225 students consisted of 86 participants from course I and 139 students from course II, 145 minus the duplicate observations from 6 students who participated in the experiment for both courses. For these 6 students, only the observation from course I was used since it had all stages exactly one day before course II.

final estimation sample reduces to around 63% of those who initially registered for the experiment. The numbers about grade reporting in column 7 are relative to the main sample in column 5 since grade reporting did not depend on answering the post-exam questionnaire and was incentivized separately.

Table 2: Number of observations by stage of the experiment and course

	Course size	Sign-up	PQI	FB + PQII	Without Duplicates	Post-Exam w/o dupl.	Grades w/o dupl.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Course I							
Observations	644	132	101	86	86	81	66
% of previous stage		20.5	76.52	85.15	100	94.19	76.74
% of initial obs.		100.00	76.52	65.15	65.15	61.36	50
Course II							
Observations	1006	227	165	145	139	129	103
% of previous stage		22.56	72.69	87.88	95.86	88.96	74.1
% of initial obs.		100.00	72.69	63.88	61.23	56.83	45.37
Total							
Observations	1650	359	266	231	225	210	169
% of previous stage		21.76	74.09	86.84	97.40	93.33	75.11
% of initial obs.		100.00	74.09	64.34	62.67	58.49	47.07

Notes: Number of observations and relative shares of participants at all experimental stages, from sign-up to grade reporting, by course. The course size refers to the number of students who took the exam on the first available date. Column (5) excludes the observation for the second course of all students who participated in the experiment for both courses. Data source col. 1: official university statistics.

Table 3: Number of observations by stage of the experiment and treatment status

	PQI	FB + PQII	Without Duplicates	Post-Exam w/o dupl.	Grades w/o dupl.
	(1)	(2)	(3)	(4)	(5)
Control Group					
Observations	51	44	44	42	29
% of previous stage		86.27	100	95.45	65.91
% of initial obs.	100	86.27	85.27	82.35	56.86
POSNEG					
Observations	106	91	89	81	69
% of previous stage		85.85	97.80	91.01	77.53
% of initial obs.	100	85.85	83.96	76.41	65.09
NEGPOS					
Observations	109	96	92	87	71
% of previous stage		83.49	95.83	94.56	77.17
% of initial obs.	100	83.49	84.40	79.82	65.14

Notes: Number of observations and relative shares of participants from the first practice questions to grade reporting, by treatment status. Column (3) excludes the observation for the second course of all students who participated in the experiment for both courses.

Table 3 shows attrition rates by treatment status. The sample after the first questionnaire comprised of 266 students that were randomized into one control and two treatment groups as described above, by course and gender. The control group consisted of 20% of the sample, i.e. 51 students. Of those, 44 students actually received the

feedback in the second stage. Some students had participated in the experiment for both courses such that for them only the observation from course I is used. The remaining 42 non-duplicate observations are then part of the analysis sample.

The *positive-negative* (POSNEG) and *negative-positive* (NEGPOS) treatment groups comprised of 40% of the sample each, i.e. 106 students for POSNEG and 109 for NEGPOS. Of these students assigned to the treatment groups, 89 (92) finally participated in the treatment for the POSNEG (NEGPOS) group, excluding one of the observations for the duplicate students who participated in both courses. Participation in the second questionnaire (column 2) was not related to the later treatment status of students (not shown). Similarly, attrition in the post-exam questionnaire (column 4) is not related to the realized treatment status (not shown).

Incentivized grade reporting was offered to everyone from the main sample, i.e. those who participated in the second set of practice questions (column 3). Hence, all relative numbers in column 5 refer to this group. Although a slightly larger attrition rate can be observed for the control group, the difference to any of the treatment groups separately or jointly is not statistically significant (not shown). Nonetheless, I will focus on comparisons between the two treatment groups only for analyses looking at exam grades as outcomes.

Table 4 shows descriptive statistics for the main control variables to be used in some specifications of the regression analyses. The final estimation sample is almost gender-balanced and has a median and mean age of 20 years. The latter is in line with students being recruited from two courses that are meant to be taken in the first and in the second year of their undergraduate studies. Participants mostly come from Business Administration and Economics majors, a small part is enrolled in Business and Economics as a minor or other majors (mostly teaching degrees for economics). Most students are in the semester corresponding to the study plan for their major, i.e. at a very early stage of their undergraduate studies. 90% of the sample obtained their high school degree in Germany and hence provided a German high school GPA and their last math grade from high school. Grades are presented on the German scale (1, best, to 6, worst) and are in line with admission policies of German universities. 54% (63%) of students state that their mother or other female guardian (father or other male guardian) has a university degree.

This points at a rather more academic sample in the experiment compared to average Business and Economics students in Germany. According to the most recent NEPS data (a representative survey of university students in Germany, see NEPS-Netzwerk (2021)), 31% of first-semester Business and Economic students in Germany have at least one parent with an academic background, compared to 71% in my sample. This might also be due to the setting in the city of Munich where living costs are comparably high and hence the student population might be positively selected to

Table 4: Descriptive Statistics

	Obs. (1)	Mean (2)	Median (3)	Std. Dev. (4)	Min. (5)	Max. (6)
Female	225	0.56	1.00	0.50	0	1
Age	224	20.18	20.00	3.00	18	54
Business Administration	225	0.56	1.00	0.50	0	1
Economics	225	0.17	0.00	0.38	0	1
Bus. and Econ. Education	225	0.06	0.00	0.23	0	1
Minor Bus./Econ./Education	225	0.12	0.00	0.33	0	1
Other major	225	0.10	0.00	0.30	0	1
Semester	225	1.93	1.00	1.21	1	7
German high school degree	225	0.90	1.00	0.30	0	1
High school GPA	202	1.74	1.70	0.46	1	3
Last math grade	201	1.85	2.00	0.87	1	5
Mother university degree	224	0.54	1.00	0.50	0	1
Father university degree	225	0.63	1.00	0.48	0	1
Mother employed	224	0.85	1.00	0.36	0	1
Father employed	225	0.87	1.00	0.34	0	1
First-generation migrant	225	0.16	0.00	0.37	0	1
Second-generation migrant	225	0.20	0.00	0.40	0	1
Patience	225	3.56	3.67	0.69	2	5
Risk aversion	225	3.03	3.00	0.94	1	5
Conscientiousness	225	3.83	4.00	0.78	2	5
Neuroticism	225	3.04	3.00	1.06	1	5
Openness	225	3.46	3.50	1.07	1	5
Extraversion	225	3.39	3.50	1.12	1	5
Agreeableness	225	3.22	3.00	0.88	1	5
Feedback aversion	225	2.10	2.00	0.74	1	5
Self efficacy	225	3.89	4.00	0.69	1	5
Motivation university studies	225	3.86	4.00	0.96	1	5
Weekly hours invested in course	225	6.43	5.00	9.69	1	135

Notes: Descriptive statistics of control variables used in the later analyses. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course.

begin with. Contrarily, my sample also has a rather high level of migration background. In the NEPS data, 18% of Business and Economics students are born abroad or had a parent who was born abroad, while these students make up 36% of my sample.

Furthermore, students self-reported their patience, risk-aversion, big five personality traits, feedback aversion, and self-efficacy, on five-point scales. Students consider themselves rather patient and risk-averse, and they assign quite high scores for conscientiousness, openness, extraversion, and self-efficacy. Instead, they state to be less neurotic and agreeable. Furthermore, participating students attribute themselves a low level of feedback aversion.

Finally, students were asked how motivated they were to study for their university studies in general as well as how many hours they had spent per week studying for the specific course so far. Motivation was measured from 'not motivated at all' (1) to 'very motivated' (5) and students on average seem to be quite motivated. Also, students on average invested 6.5 hours per week to study for the respective course (with a median of 5 hours and one large outlier at 135).

A similar descriptive table for pre-treatment outcomes seven days before the exam can be found in table B.1. Self-reported motivation to study for the exam again was measured via a survey question that asked students how motivated they currently were to study for the exam. The scale ranged from 1 ('not motivated at all') to 5 ('very motivated'). Overall motivation was relatively high with a mean of 3.5 and a median value of 4.

Students' pre-treatment performance was measured using their score on the first set of practice questions. Out of 20 possible points, students on average scored 8.5, implying that they were still lacking quite some course knowledge one week before the exam. Interestingly, students had higher expectations regarding their points before answering the questions (12.6 points on average), but adjusted these expectations downwards after having responded to the practice questions (mean of 8.7 points). This immediate absolute performance belief is on average quite accurate. The self-evaluation of this performance was assessed by asking students whether they would consider this expected performance 'very poor' (1), 'poor', 'neutral', 'good', or 'very good' (5). Again, students considered their expected performance much better before answering the practice questions (mean of 2.9 before compared to a mean of 1.9 after).

Expectations about future exam grades were also elicited before treatment, i.e. seven days before the exam. The German scale for grades ranges from 1 (best) to 5 (fail) and only contains decimal values at .3 and .7 between integer values. Interestingly, no student expected to fail the class one week before the exam. On average, students expected a grade of 2.3 which would be classified as 'good'. Self-evaluation of exam grades follows more closely the self evaluation of the first set of practice questions before answering them and is even more optimistic with a mean of 3.2.

Finally, students were also asked to state their study plan during the first questionnaire. On the one hand, this included questions on their planned study hours on each of the remaining days before the exam. On the other hand, students were asked to pick three topics from the course syllabus that they were planning to focus on in their remaining study time. Students planned to steadily increase their study hours when moving closer to the exam with the highest value on the last day before the exam (4.2 hours on average). Furthermore, 27% (29%) of students had the topic they would perform best (worst) in during the later practice questions on their list of prioritized study topics in this pre-treatment questionnaire.

Table 5 presents a balancing check of all controls for the three groups students were randomized into. Columns 4 to 6 show differences across the treatment and control groups, including asterisks indicating a statistically significant difference. Only three differences turn out to be significant at the five percent level which could as well be expected by chance. Similarly, a joint F-test on all differences cannot reject the hypothesis that they are jointly zero (not shown). Furthermore, all variables seen in this table are controlled for in some specifications of the presented regressions.

A similar balancing check for pre-treatment outcomes can be found in table B.2. Pre-treatment motivation and performance did not significantly vary between the control and treatment groups. Some minor differences can be observed for performance beliefs about the first set of practice questions. Students in the POSNEG group overall seem to be slightly more optimistic before answering the questions which is less visible after all students have actually done the practice. Both treatment groups are mildly more optimistic about their future grade (on a German scale) compared to the control group. No differences can be observed regarding students pre-treatment study plan. In the later regressions, I will mostly control for the respective pre-treatment outcome which should take care of any remaining imbalance.

4. Results

This study aims to evaluate the effects of feedback ordering on the main outcomes motivation and performance as well as the potential mechanisms beliefs and study behavior. I will first show results on motivation (section 4.1), then move on to beliefs (section 4.2) and the study plan (section 4.3) before showing the results on performance (section 4.4). Section 4.5 explores some further mechanisms of the treatment effects, whereas section 4.6 summarizes heterogeneous treatment effects that are described more in detail in appendix A. Finally, section 4.7 looks at the effects of receiving any feedback.

Table 5: Balancing check for all control variables

	Control	POSNEG	NEGPOS	(2) vs (1)	(3) vs (1)	(2) vs (3)
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.614	0.562	0.543	-0.052	-0.070	0.018
Age	20.273	20.056	20.253	-0.217	-0.020	-0.197
Business Administration	0.568	0.573	0.533	0.005	-0.036	0.040
Economics	0.227	0.180	0.130	-0.047	-0.097	0.049
Bus. and Econ. Education	0.068	0.090	0.022	0.022	-0.046	0.068**
Minor Bus./Econ./Education	0.068	0.101	0.163	0.033	0.095	-0.062
Other major	0.068	0.056	0.152	-0.012	0.084	-0.096**
Semester	2.045	2.045	1.772	-0.001	-0.274	0.273
German high school degree	0.909	0.899	0.902	-0.010	-0.007	-0.003
High school GPA	1.740	1.725	1.742	-0.015	0.002	-0.017
Last math grade	2.017	1.746	1.873	-0.271*	-0.144	-0.127
Mother university degree	0.591	0.545	0.500	-0.045	-0.091	0.045
Father university degree	0.614	0.618	0.641	0.004	0.028	-0.023
Mother employed	0.909	0.830	0.848	-0.080	-0.061	-0.018
Father employed	0.841	0.865	0.891	0.024	0.050	-0.026
First-generation migrant	0.136	0.202	0.141	0.066	0.005	0.061
Second-generation migrant	0.182	0.169	0.228	-0.013	0.046	-0.060
Patience	3.621	3.543	3.540	-0.078	-0.081	0.003
Risk aversion	3.261	2.955	2.984	-0.306*	-0.278	-0.029
Conscientiousness	3.875	3.865	3.783	-0.010	-0.092	0.083
Neuroticism	3.375	2.871	3.049	-0.504***	-0.326*	-0.178
Openness	3.523	3.399	3.500	-0.124	-0.023	-0.101
Extraversion	3.489	3.483	3.250	-0.005	-0.239	0.233
Agreeableness	3.352	3.169	3.196	-0.184	-0.157	-0.027
Feedback aversion	2.030	2.094	2.149	0.063	0.118	-0.055
Self efficacy	3.947	3.910	3.841	-0.037	-0.106	0.070
Motivation university studies	3.977	3.798	3.870	-0.180	-0.108	-0.072
Weekly hours invested in course	5.500	7.635	5.710	2.135	0.210	1.925
Observations	44	89	92	133	136	181

Notes: Balancing check between control and treatment groups on all controls used in the analyses. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course. Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.1. Treatment effects on motivation

Students were asked multiple times how motivated they were to study for the exam of the specific course for which they were participating in the experiment. All students received this question before solving the first set of practice questions (seven days before the exam). Control group students were asked again once before solving the second set of practice questions (three days before the exam). Students from the two treatment groups received the question about topic-specific study motivation twice before solving the second set of practice questions (three days before the exam), after each of the separate feedback elements (see table 1).

This setup allows me to estimate various types of effects: the effect of feedback ordering by comparing post-feedback outcomes for the two treatment groups POSNEG and NEGPOS, the effect of positive versus negative feedback only with a between-

feedback comparison of the two treatment groups, and the effect of the positive and negative feedback respectively compared to no feedback by comparing treatment groups to the control group.[11]

I start by looking at the effects of feedback ordering on motivation across treatment groups. To do so, I use the following basic regression that only includes the two treatment groups who received feedback on the first set of practice questions:

$$O_i = \beta_0 + \beta_1 POSNEG_i + female \times course + best_i + worst_i + \beta_2 O_{i0} + \beta_3 performance_{i0} + \mathbf{X'}_i \boldsymbol{\beta}_4 + \varepsilon_i \quad (1)$$

O_i refers to the different outcomes as described above, in this case motivation. $POSNEG_i$ is a dummy that takes the value one for the treatment group with positive feedback first and negative feedback afterwards, and zero for the group receiving negative feedback first. ε_i is an individual error term, I use robust standard errors. Furthermore, equation 1 contains dummies for the course-gender cell within which I randomized as well as for the topics students received feedback on since these were not random in this setting. Also, most specifications will include pre-treatment performance controls from the practice questions as well as the respective pre-treatment outcome.[12]

In different specifications, I then sometimes include a larger set of controls to increase the precision of my estimates. $\mathbf{X}_i$ includes the background characteristics described in section 3 (age, gender, semester, field of study, high-school GPA, last math grade in high school, parents' education and occupation, migration status) as well as the personality traits elicited prior to the first set of practice questions (patience, risk-aversion, big five, self-efficacy, feedback aversion). Not all students reported high-school grades, depending on whether they had completed their high-school degree in Germany. For all those with missing values, I impute high-school performance with the mean of the sample and add an imputation dummy to all respective regressions.

Table 6 presents the results of this estimation with motivation as an outcome. Columns 1-3 show the results for motivation measured between the two feedback elements, columns 4-6 refer to the question on motivation after both feedback elements had been provided. All regressions only include students from the two treatment groups POSNEG and NEGPOS. As indicated by the coefficient for the POSNEG dummy, receiving feedback in the ordering *positive-negative* compared to *negative-positive* leads to higher post-feedback motivation to study for the exam. This effect emerges after the

[11] All three types of analyses (for all presented outcomes) were pre-specified in the pre-analysis plan that is part of the AEA RCT Registry entry AEARCTR-0008953.

[12] Results look very similar when simply controlling for the female and the course dummies individually, without adding the interaction term (not shown). Similarly, results do not change much if I include the points in the best and worst topic as a pre-treatment indicator of performance rather than overall performance in the first set of practice questions (not shown).

first feedback element has been presented (columns 1-3) which can be interpreted as the pure effect of positive versus negative feedback as single elements where the information content of the feedback is different.

Table 6: Treatment effects on motivation between and after feedback elements

	After first feedback			After both feedback elements		
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	0.343**	0.411***	0.434***	0.220	0.292**	0.349**
	(0.151)	(0.128)	(0.136)	(0.154)	(0.134)	(0.136)
Female × Course Indicator	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓
Pre-treatment motivation		✓	✓		✓	✓
Pre-treatment performance		✓	✓		✓	✓
Controls			✓			✓
Observations	181	181	179	181	181	179
R^2	0.144	0.366	0.512	0.104	0.347	0.498

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the exam between feedback elements (col.s 1-3) and after both feedback elements (col.s 4-6). OLS regressions. Dependent variable: (standardized) motivation to study for the exam. Col.s 1 and 4 only include a female × course indicator for each randomization cell and feedback topic dummies. Col.s 2 add 5 pre-treatment motivation and performance. Finally, col.s 3 and 6 add all controls from table 4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

After students have received the second feedback element as well, this information asymmetry is no longer present since students from both groups were informed about the best and worst topic of their performance in the practice questions. Hence, the estimates in columns 4-6 show the treatment effect of the feedback ordering on study motivation. The effect is positive and significant, which is especially evident when including pre-treatment motivation and performance. The latter are strong predictors of post-treatment motivation as can be seen from the values of the R-squared in columns 2 and 5 and hence they increase the statistical power of the estimation.

The measure for motivation is standardized here such that the treatment effect can be interpreted in standard deviations. This implies that first receiving positive feedback leads to almost 0.3 standard deviations higher post-feedback motivation to study for the exam compared to first receiving negative feedback in my preferred specification in column 5. Including a large set of controls in columns 3 and 6 increases the coefficient size, but also loses two observations due to missing values of control variables.[13]

[13]Tables B.4 and B.5 show a full set of specifications for both outcomes, including a raw comparison

Overall, coefficient sizes vary to a certain degree across columns 3-6 of table 6 (and in the full table B.5) which might raise concerns about the validity of the results. The changes might be related to the relatively low sample size: even if carried out correctly, randomization in small samples may still produce some correlation between the treatment and the control variables. This could explain why coefficient estimates are affected by the inclusion of controls. Including many controls should then in principle reduce omitted variable bias when estimating the treatment effects. In this rather small sample though, the inclusion of many controls may lead to overfitting of the specification. If that were an issue here, we would see the standard errors of the coefficient on the POSNEG dummy increase which does not seem to be the case in column 6 of table 6.

To rule out that the inclusion or exclusion of certain controls drives the observed treatment effects, I perform both LASSO and double LASSO procedures that choose covariates to be included in the model based on their relevance for predicting the outcome only (LASSO) or both the outcome and the treatment variable (double LASSO). Table B.3 shows the treatment coefficient of the resulting estimations with selected covariates. For both methods, I run two variants: one forces the procedure to include the main indicators from column 5 of table 6 (female × course indicator, feedback topic dummies, pre-treatment motivation and performance) into the estimation, the other one allows for free choice among all indicators, dummies, and controls. The coefficient size of the preferred restricted versions in columns 1 and 3 in table B.3 is very close to column 5 in table 6 which confirms choosing this as a preferred specification. When allowing the selection procedure to choose freely from all controls, coefficients on the treatment dummy become slightly smaller, standard errors change very little though.

To understand what is driving the treatment effects on motivation, I next compare the evolution of motivation in the treatment groups to the initial motivation of the control group. Borrowing from Coffman et al. (2021), I estimate the following equation:

$$\begin{aligned} O_{it} = \beta_0 &+ \beta_1 POSNEG\_initial_{it} + \beta_2 POSNEG\_fb1_{it} + \beta_3 POSNEG\_fb2_{it} \\ &+ \beta_4 NEGPOS\_initial_{it} + \beta_5 NEGPOS\_fb1_{it} + \beta_6 NEGPOS\_fb2_{it} \\ &+ \beta_7 CONTROL\_after_{it} + female \times course + best_i + worst_i + \mathbf{X'}_i \boldsymbol{\beta}_8 + \varepsilon_{it} \end{aligned} \tag{2}$$

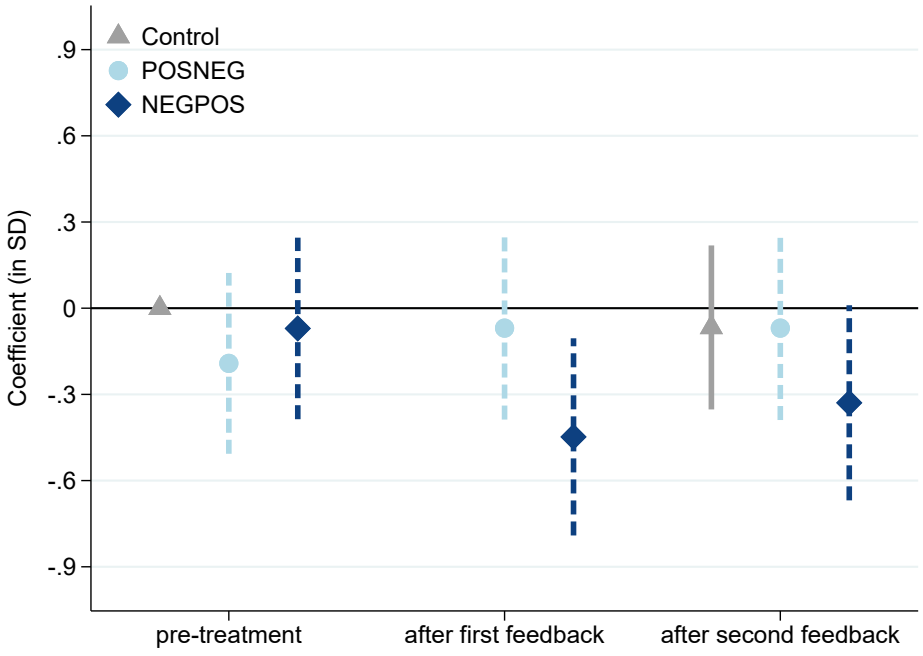
All outcomes are standardized with respect to the initial average outcome of the control group which serves as a reference category, i.e. the outcomes are demeaned and divided by the standard deviation of the pre-treatment control group outcome. O_{it} refers to the respective outcome (here motivation) at three points in time: pre-treatment (initial), after the first feedback (fb1) and after the second feedback (fb2). The coefficients

without any controls. It then adds all controls separately.

β_1 to β_7 refer to dummies that indicate whether an observation belongs to one of the three groups and when it was measured. This way, the resulting coefficients report deviations from the initial control mean of the respective outcome. The control variables are the same as in equation 1. Standard errors are clustered at the individual level since there are multiple observations for each individual now.

Figure 2 shows the resulting coefficients for the two treatment groups as well as the control group. First, it can be seen that there is no significant difference between the pre-treatment motivation of the three groups of students ('*initial*'). Furthermore, there is no movement over time of the control group's motivation (gray plots). Instead, for the NEGPOS treatment group there is a large drop in motivation following the first, i.e. negative, feedback ('*fb1*'). This does not fully recover to the initial levels of the control group after the second, i.e. positive, feedback, and remains statistically distinguishable from these initial levels. More interestingly though, this drop in motivation cannot be observed for the POSNEG group when they receive the negative feedback part ('*fb2*'). This suggests that first receiving positive feedback shields respondents from experiencing a decrease in motivation after negative feedback. Maybe surprisingly, motivation does not increase (much) after receiving positive feedback for the POSNEG group compared to the control group's initial motivation. This might hint towards a ceiling effect: students were initially very motivated to study for the exam and might hence not react much to positive feedback on this dimension.

Figure 2: Evolution of motivation to study for the exam



Notes: Plot of coefficients corresponding to equation 2 with motivation as an outcome, including the 95% confidence intervals. The corresponding regression table can be found in appendix table B.6.

The results for motivation have some important implications. Positive and negative feedback individually have very different effects on students' motivation in this setting. Positive feedback doesn't affect motivation while (early) negative feedback reduces motivation. Often though, feedback givers do not have the choice between either positive or negative feedback, but they rather want to provide suggestions for improvement that necessarily imply some negative feedback. My findings suggest that there is no need to avoid giving negative feedback when both negative and positive feedback are provided. Simply paying attention to the ordering of the feedback elements might be a solution to avoid demotivating students with negative feedback.

4.2. Treatment effects on beliefs

To understand whether the observed effects of feedback ordering on motivation can be explained by students' belief updating, I next turn to the results on performance beliefs. Performance feedback is often associated with updating beliefs about one's performance, which in turn might have an impact on motivation. Beliefs about performance were elicited on three measures: performance in the first set of practice questions, in the second set of practice questions, and in the exam. All belief elicitation blocks contained three elements: students had to first state which exact points/grade they thought they had achieved/would achieve. Secondly, they had to evaluate this expected performance as either poor or good as described above. Lastly, students had to provide probabilities adding up to 100 about the likelihood of being ranked in each quarter of the performance distribution of all participants.

Performance beliefs were surveyed pre-treatment (seven days before the exam) and both between the two feedback elements as well as after the second feedback (three days before the exam). Since feedback was given on the first set of practice questions, between-feedback belief elicitation focused on the performance in the first set of practice questions and in the exam. Instead, beliefs about performance in the second set of practice questions were only elicited once, right before performing the questions. This leads to three belief outcomes to study: retrospective beliefs about the performance in the practice questions that students receive feedback on (right after performing the questions as well as after each of the feedbacks), immediate forward-looking beliefs about the second set of practice questions, and medium-term forward looking beliefs about the exam grade. In the following, I will present results on absolute performance beliefs only.

Table 7 shows the results for all explicit beliefs about performance regarding the points achieved in both of the practice sets provided during the experiment as well as the exam grade. It uses equation 1 for my preferred specification from col. 5 of table 6 which includes a female × course indicator, feedback topic dummies, and pre-treatment expectations and performance, but no further controls. Both forward-looking beliefs

Table 7: Treatment effects on beliefs

	Practice set I		Exam		Practice set II
	after fb1	after fb2	after fb1	after fb2	after fb2
	(1)	(2)	(3)	(4)	(5)
POSNEG	0.501***	-0.025	0.136*	0.058	0.069
	(0.081)	(0.090)	(0.077)	(0.075)	(0.145)
Female × Course Indicator	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓
Pre-treatment expectations	✓	✓	✓	✓	
Pre-treatment performance	✓	✓	✓	✓	✓
Observations	181	181	181	181	181
R^2	0.770	0.678	0.760	0.755	0.115

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on beliefs. OLS regressions. Dependent variable: (standardized) beliefs about performance in practice set 1 (col.s 1 and 2), exam grades (col.s 3 and 4) and practice set 2 (col. 5). All columns include a female × course indicator for each randomization cell, feedback topic dummies, and pre-treatment beliefs (where applicable) and performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

about future exam performance as well as retrospective beliefs about past performance in the first set of practice questions are higher in the POSNEG group compared to the NEGPOS group after the first feedback element, although more convincingly for beliefs about practice set 1. The effect does not persist after the second feedback element for both practice set 1 (on which feedback was given) as well as exam grades. No effect can be seen for the second set of practice questions where beliefs were elicited only once right before answering the questions which was right after the two feedback elements. This implies that belief updating cannot explain the treatment effects on motivation since there is no treatment effect of feedback ordering on performance beliefs.

Figure B.3 shows the evolution of performance beliefs about the first set of practice questions compared to the initial mean of the control group. Interestingly and in contrast to the reactions to motivation, when updating beliefs students follow patterns known from 'motivated beliefs': they react strongly to the positive feedback, but much less to the negative feedback. This is especially true for students receiving the positive feedback as a first element whereas students who first receive negative feedback react more cautiously to positive feedback, even though the difference between the reactions of the two treatment groups to positive feedback is not statistically significant (not shown).[14]

[14]To test formally whether the reactions to positive feedback are different across the two treatment groups, I can compare the coefficient on the dummy $POSNEG\_fb1_{it}$, i.e. the reaction of the POSNEG group to the positive feedback, to the isolated reaction of the NEGPOS group to the positive feedback which is given by the difference between $NEGPOS\_fb2_{i,t}$ and $NEGPOS\_fb1_{i,t}$.

The results on performance beliefs provide some important insights to the mechanisms behind the treatment effects on motivation. Belief updating has been extensively studied in economic theory such that there is a clear idea on how students would react to the feedback elements I provide. The fact that students seem to react as expected helps to disentangle mechanisms of the treatment effect on motivation that can be explained by economic theory from other potential mechanisms that we currently don't have good explanations for. Such other mechanisms could be found in other disciplines such as psychology where theories suggest emotional reactions to feedback. These will be analyzed in further detail in section 4.5 using data from the post-exam questionnaire.

4.3. Treatment effects on study behavior

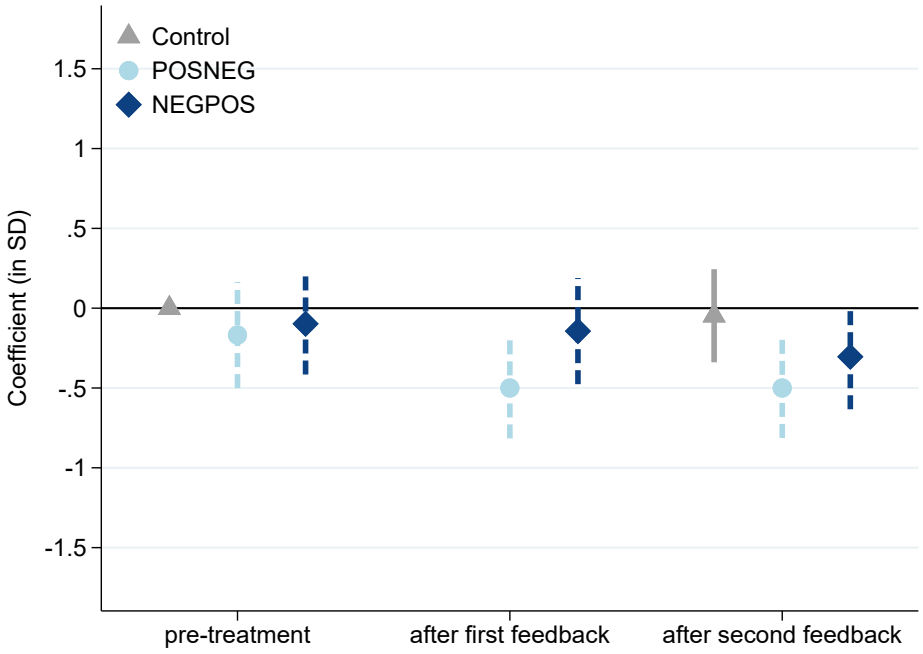
To understand potential post-treatment changes in the study behavior of students, participants were also asked about their study plan. This included questions about the hours they were planning to study for the exam on each of the remaining days before the exam. These were also surveyed after the exam to see the actually implemented study behavior of students. Furthermore, students were asked multiple times to name three topics from the course outline that they were planning to focus on in the remaining study time. These could be chosen from the list of chapters included in the respective course. Participants from the treatment groups received these questions three times: pre-treatment (seven days before the exam), after the first feedback and after the second feedback (three days before the exam). Control students only responded twice to these question blocks as for motivation and beliefs, once seven days before the exam and once three days before the exam.

Table B.8 in the appendix shows the treatment effect on the aggregated study hours students planned to spend on this course in the last two days before the exam. The coefficients are not statistically distinguishable from zero and if anything negative for those receiving positive feedback first.

Instead, figures 3 and 4 illustrate how students reacted to the topics they have been given feedback on. For each student, positive and negative feedback was given individually since it depended on which of the topics they had performed in best and worst during the first set of practice questions. The measure is also available for control group students since their practice questions were corrected as well. Hence, they also have a best and worst topic, but didn't receive feedback on them. This allows me to construct an indicator for each student at each time they responded to this question that is one if students have the respective best/worst topic on their list of top three study topics and zero if they don't. Then, I can repeat the estimation from equation 2 for this indicator at three (two) points in time for the treatment (control) groups.

Figure 3 shows the evolution of this indicator for the topics individuals performed

Figure 3: Evolution of whether positive feedback topic is on study list



Notes: Plot of coefficients corresponding to equation 2 with an indicator of whether a person has the topic on their study list they have or will receive positive feedback on as an outcome, including the 95% confidence intervals. The corresponding regression table can be found in appendix table B.9.

in best and eventually received positive feedback on if they were in one of the treatment groups. At the initial stage, i.e. pre-treatment, there is no significant difference between the control and treatment groups on how likely individuals were to have this individual-specific best topic on their list of study priorities. After receiving positive feedback on the respective topic, the POSNEG group shows a significant drop in the share of individuals who include this topic on their study list compared to the initial control mean. Similarly, this happens for the NEGPOS group after they have received positive feedback, even if slightly weaker in magnitude and statistical significance. This suggests that students are more cautious to remove a topic from the study list if they received negative feedback before.[15] No change over time can be seen for the control group, confirming that the results for the treatment groups do not include a general trend.

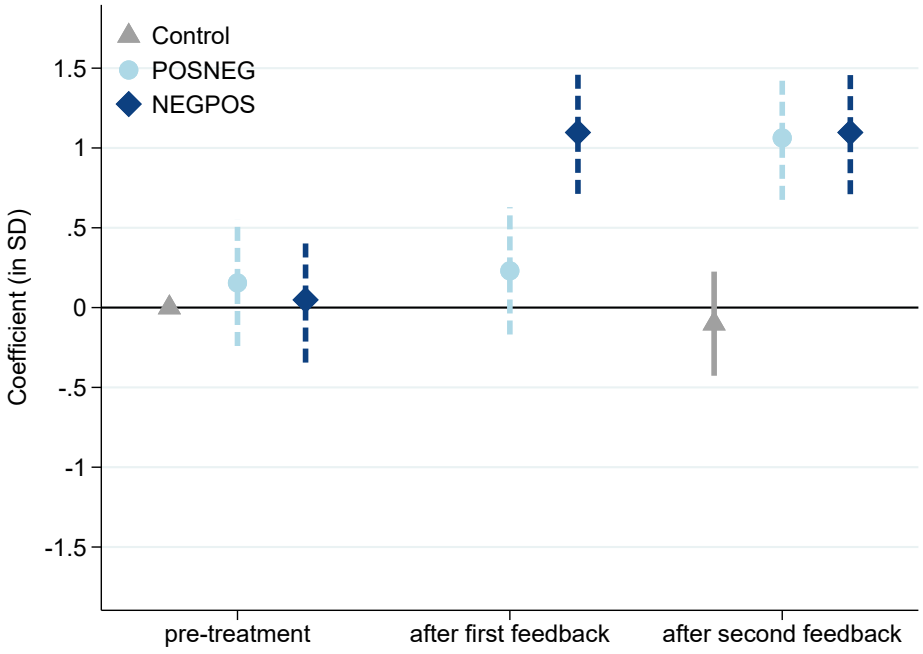
Interestingly, this development is still visible when looking at students' responses to the question which topics they had actually focused a few days later (see figure B.4 in the appendix). Even one day after the exam, individuals report less often to have this personally best topic as part of their study list in the POSNEG group, but not significantly so in the NEGPOS group. The coefficients for both groups are smaller in magnitude and significantly different from the coefficients after the second feedback

[15]In fact, the difference between $POSNEG\_fb1_{i,t}$, i.e. the reaction of the POSNEG group to the positive feedback, and the isolated reaction of the NEGPOS group to the positive feedback, $NEGPOS\_fb2_{i,t} - NEGPOS\_fb1_{i,t}$, is significantly different from zero.

element, at the 5% level for the POSNEG group and at the 10% level for the NEGPOS group. This suggests that students face some constraints in following their updated study plan, but it might as well mean that students remember their positive feedback topic less accurately. Again, no general trend can be seen for the control group.

Similarly, figure 4 shows how the inclusion of worst topics onto individuals' study lists evolves over the course of the experiment. Again, students react to the feedback provided during the experiment: in this case, they add topics to their study list they received negative feedback on. This seems to be equally strong for both treatment groups, the coefficients for $NEGPOS\_fb1_{i,t}$ and the difference of the coefficients for $POSNEG\_fb2_{i,t}$ and $POSNEG\_fb1_{i,t}$ are not significantly different from each other. The control group again does not show any changes which is plausible given that they never received the information on their best and worst topic.

Figure 4: Evolution of whether negative-feedback topic is on study list



Notes: Plot of coefficients corresponding to equation 2 with an indicator of whether a person has the topic on their study list they have or will receive negative feedback on as an outcome, including the 95% confidence intervals. The corresponding regression table can be found in appendix table B.11.

Equivalently to the positive feedback topics, the results for negative-feedback topics persist to the post-exam questionnaire about actual study topics as well (see figure B.5). Again, effect sizes become smaller and are significantly different to the coefficients from before the exam. But they are still significantly different from the initial control mean and there is no development for the control group.

The results for students' study behavior again help to get a better understanding of the mechanisms during the experiment. The fact that students do not adjust their

study time suggests that there is limited flexibility in students' learning schedule, but might also imply that they consider other margins of adjustment more relevant in this context. In fact, students seem to adapt their study efficiency rather than the quantity by changing the topics according to the feedback they received. This has important implications for highly time-constrained settings in which students might still benefit from feedback although they cannot increase their study time. Since students react in a similar fashion in both treatment groups, this finding can be generalized to receiving any feedback rather than being unique to a specific feedback ordering.

4.4. Treatment effects on performance

The treatment effect on performance can be analyzed using two measures: immediate performance in the second set of practice questions (points out of 10) and medium-term high-stake performance in the exam (German-scale grade). Both measures presented here are standardized across treatment groups to have a mean of zero and a standard deviation of one. This way, coefficients can be interpreted as changes in standard deviations. Furthermore, exam grades are inverted such that higher grades correspond to better performances. This is necessary because in the German grading scale smaller numbers correspond to better grades.

Table 8: Treatment effects on performance

	Practice set II			Exam		
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	-0.074	-0.095	-0.008	0.164	0.099	0.034
	(0.145)	(0.128)	(0.137)	(0.154)	(0.130)	(0.124)
Female × Course Indicator	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓
Pre-treatment performance		✓	✓		✓	✓
Controls			✓			✓
Observations	181	181	179	140	140	139
R^2	0.168	0.327	0.505	0.318	0.526	0.693

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on performance in the second set of practice questions and the exam. OLS regressions. Dependent variable: (standardized) performance points (practice set 2, col.s 1-3) and standardized and inverted exam grades (col-s 4-6). Columns 1 and 4 show estimates without controls but include a female × course indicator for each randomization cell and feedback topic dummies. Col.s 2 and 5 add pre-treatment performance. Finally, col.s 3 and 6 add all controls from table 4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table 8 shows the treatment effects on both performance measures according to an

estimation of equation 1. Columns 1-3 focus on immediate performance in the second set of multiple choice practice questions. Coefficients are statistically not distinguishable from zero and if anything negative. When turning to the grades in the exam (col.s 4-6), a similar picture emerges: coefficients are now slightly positive, but still insignificant. From simply looking at the confidence intervals, I can exclude effect sizes larger than 0.358 at the 95% confidence level (0.316 at the 90% level, 0.441 at the 99% level). These results imply that the changes in motivation as well as the adjustment of study topics do not necessarily translate into changes in immediate or medium-term, high-stake performance.

A reason for this could be the relatively short time frame. Students do not have the time to prepare further for the second set of practice questions since they answer them right after having received the feedback. Fischer and Wagner (2018) show in a field experiment that students react negatively to feedback given right before an exam and positively to similar feedback given a few days earlier. My results are in line with the direction of the coefficients, although they are not statistically different from zero.

Another potential reason for not finding treatment effects is the lack of statistical power in detecting relatively small effect sizes. To be able to detect an effect size of 0.099 from my preferred specification in column 5 of table 8 with a confidence level of 90% and statistical power of 80%, I would need a total sample size of around 2,500 students. This could be reduced to about 800 students when considering the large explanatory power that the control variables of the regression have.[16]

Additionally, the sample size of the students who voluntarily reported their grade is smaller than the full estimation sample for the other treatment effects. This is caused by grade reporting being voluntary rather than me having access to administrative data for all students. Grade reporting was incentivized with a €10 voucher and around 75% of students sent their grade, including a proof from their student portal (see table 2).

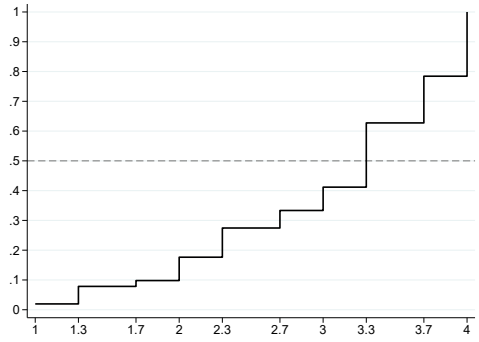
This raises concerns about sample selection into grade reporting. To get an idea of such potential selection, I compare the grades reported by students in the experiment with the aggregated statistics of all students that can be found in the publicly released grade distribution.[17]

Figure 5 shows the experimental and the official cumulative distribution functions for all students who passed course I. Grades on the x-axis are on a German scale, i.e. from 1.0, best, to 4.0, worst. The students who reported their grades in my experimental sample seem to be slightly positively selected in all parts of the distribution. The median grade is 3.3 in both distributions, but the share of students with the lowest passing grade

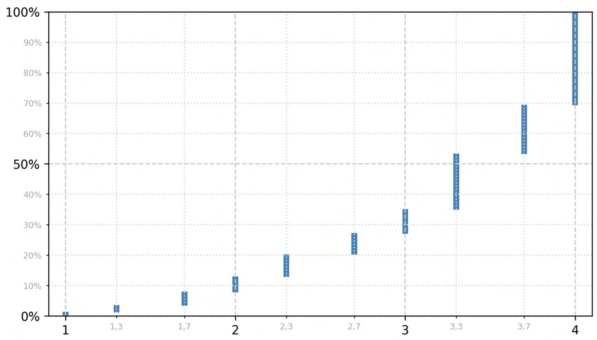
[16]The underlying calculations have been performed with the STATA command *power* as well as with Optimal Design Software, see Raudenbush, S. W., et al. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software]. Available from www.wtgrantfoundation.org (last accessed on October 28 2022).

[17]Official university statistics only report the graphs in panels B of figures 5 and 6, without the underlying data, and, separately, the respective passing rates. I will hence consider selection separately by course using the available plots.

Figure 5: Selection into grade reporting (course I)



(A) Cumulative distribution function of grades in experiment



(B) Official cumulative distribution function (all students)

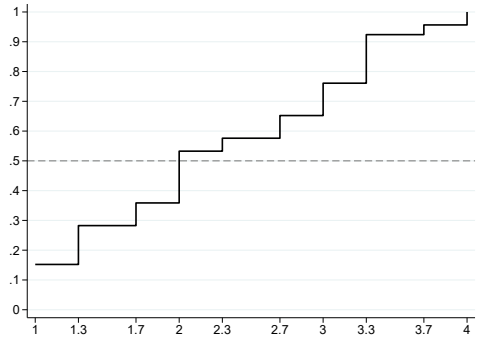
Notes: Cumulative distribution functions of exam grades in course I for the experimental sample and all students. Data source panel B: official university statistics.

(4.0) is higher in the full student population (around 30%) than in the experimental sample (around 20%). The share of students who failed the course differs markedly between the two groups: only 22.7% of students in my sample provided proof of them failing the course whereas official statistics report that 50.8% failed the course. Since the difference is so stark on this dimension but not in the group of students who passed the exam, this might imply that students who failed the course disproportionately often did not report their grade in the experiment.

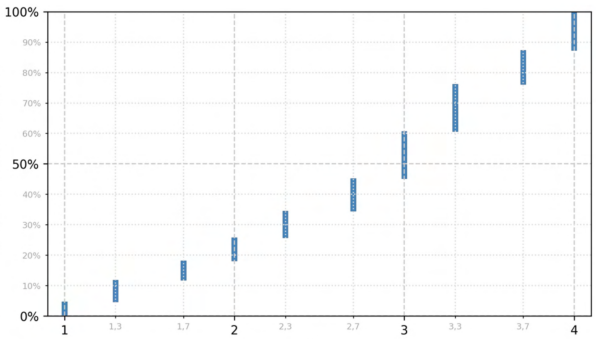
To shed more light on this argument, table B.20 shows a balancing check for students in course I, distinguishing by whether students reported their grade or not after having participated in the treatment. The results are rather surprising since it seems that students with a lower feedback aversion tended not to report their grades. Similarly, the students who failed to report their grade more commonly have a mother who is working and has a university degree. Also, they are more likely to have a second-generation migration background and grade reporting does not depend on initial motivation. Less surprisingly, those students who performed worse on the first set of practice questions are less likely to report their final exam grade. Given the amount of comparisons presented here and the often marginal level of statistical significance, these differences could also be observed by chance.

Figure 6 shows the corresponding distributions for course 2. Here, the positive selection into grade reporting is even more evident: median grades differ by one full grade point (2.0 in my sample vs. 3.0 in the full course) and the share of students who reported the lowest passing grade is around 4% in my sample vs more than 10% in the entire course. The difference between the two distributions is especially visible in the upper half where in my sample 40% of the students perform better or equal than a 2.0 whereas this share is only 20% in the full course distribution. When looking at passing rates, again, students in my sample are more likely to have passed the exam (around

Figure 6: Selection into grade reporting (course II)



(A) Cumulative distribution function of grades in experiment



(B) Official cumulative distribution function (all students)

Notes: Cumulative distribution functions of exam grades in course II for the experimental sample and all students. Data source panel B: official university statistics.

90%) than in the full course distribution (around 58%).

Table B.21 compares all characteristics used in the regressions between those who did or did not report their grade after receiving treatment in course 2. Students who do not report their grade are more risk-averse and are more likely to have a migration background, this time in the first generation. Furthermore, they have different majors, report being less conscientious, and show higher levels of extraversion. Also, they seem to report higher, i.e. worse, high school GPAs which is in favor of the explanation posed above that the lowest performers, i.e. those most likely to fail the exam might avoid reporting their grades in the framework of the experiment, even though they are paid for it. Again, grade reporting does not depend on initial motivation but is more likely for students with higher performance in the first set of practice questions.

Generally, the evidence from both courses is consistent with a notion that the money I offered to students for reporting their grades (€10 vouchers for Amazon or Avocadostore) did not fully compensate for other reasons why students might want to avoid reporting their grades. In particular, students with higher risk aversion might have been more cautious about potential consequences or might feel shame associated with very bad outcomes. Similarly, students with relatively highly educated mothers, as in course I, might be less willing to communicate a very bad grade. This could e.g. be driven by the fact that the monetary reward is less relevant for them or that they come from a higher feeling of failure when receiving a worse grade if higher education in the household is associated with more pressure about exam grades. Interestingly, no differences by gender can be observed.

Additionally, the reaction of students regarding their performance in the exam might depend on which (feedback) topics were included in the final exam. Which topics are ultimately part of the exam is plausibly exogenous to students, absent any prior indication from the teaching staff about excluded/included topics. In table 9, I hence

show an exploratory analysis using the course topics present in the final exam.[18] The results from columns 1-3 of table 9 show that there is a positive treatment effect of the POSNEG order compared to NEGPOS for those students who faced their worst practice topic in the exam. This applies to almost half of the student, i.e. 67/140, and is quite evenly distributed among the treatment groups (not shown). In general, students performed better if their negative-feedback topic was not in the exam. Furthermore, not encountering the worst topic in the exam reduces the benefit of the POSNEG order. No clear pattern can be observed for the positive feedback topics (col.s 4-6).

Table 9: Heterogeneities by exam-feedback topic correspondence for exam grades

	Outcome: exam performance (stand.)					
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG=1	0.734***	0.633***	0.396**	0.123	0.052	-0.002
	(0.219)	(0.172)	(0.165)	(0.212)	(0.163)	(0.170)
Worst fb topic not in exam=1	1.337***	1.390***	1.026***			
	(0.494)	(0.392)	(0.352)			
POSNEG=1 × Worst fb topic not in exam=1	-1.131***	-1.065***	-0.711***			
	(0.290)	(0.235)	(0.259)			
Best fb topic not in exam=1				0.809	0.411*	0.472
				(0.644)	(0.232)	(0.292)
POSNEG=1 × Best fb topic not in exam=1				0.033	0.068	0.033
				(0.292)	(0.245)	(0.245)
Female × Course Indicator	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓
Pre-treatment performance		✓	✓		✓	✓
Controls			✓			✓
Observations	140	140	139	140	140	139
R^2	0.407	0.611	0.721	0.327	0.529	0.696

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on performance in the exam. OLS regressions. Dependent variable: (standardized) exam grades. Columns 1 and 4 only include a female × course indicator for each randomization cell and feedback topic dummies. Col.s 2 and 5 add pre-treatment performance. Finally, col.s 3 and 6 add all controls from table 4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

These results imply that the effect of feedback ordering on exam grades might depend on which topics students actually face during the exam. Students from both treatment groups were equally likely to add the negative-feedback topic to their study list after hearing they had performed worst in it. If this topic then does not show up in the exam, students might feel differently about the adjustment of their study content. This might in turn affect their exam performance if focusing on this topic does not pay off

[18] I was provided with the full text of the final exams for both participating courses and identified which course topics were part of the exam. From that, I can construct a simple indicator of whether the individually best and worst topics from the practice questions were included in the exam or not.

for the exam. The latter might be especially true for students who received the negative feedback as a second element and hence were more motivated to study for the exam than those from the NEGPOS group. If these POSNEG students were especially eager to follow their adjusted study plan, they might be relatively more frustrated when not finding the worst feedback topic in the exam.

Students who instead encounter the worst feedback topic in the exam, might also perform differently based on the feedback ordering they were assigned to. Participants from the POSNEG group might remember that their motivation did not suffer from the negative feedback associated with that topic since they received positive feedback before. This might reduce the negative impact of encountering this topic in the exam when heving been in the POSNEG group. Hence, although the effects on motivation do not translate into performance effects on average, they might be a mediator of treatment effects on performance when students are faced with their negative-feedback topic in the exam.

4.5. Feelings and thoughts about feedback

This section looks at students' feelings and thoughts about the feedback I provided, which could uncover important additional mechanisms for the observed treatment effects. The outcome measures come from two parts of the questionnaires. Feelings about the separate feedback elements were elicited right after each feedback part three days before the exam. All other outcomes presented in this section come from the post-exam questionnaire that was sent to students one day after the exam.[19]

To see whether students fully grasped the feedback they received, I present two important checks. The first one refers to the treatment groups only and compares how they stated they felt about each of the separate feedback elements. The scale ranges from 'very bad' (1) to 'rather bad' (2), 'neutral' (3), 'rather good' (4), and 'very good' (5). Table 10 shows that feelings after positive feedback (col. 1 row 1 and col. 2 row 2) are generally better than after negative feedback (col. 1 row 2 and col. 2 row 1) for both groups. Furthermore, the differences between the two treatment groups are significant for both feedback elements, i.e. participants generally feel very differently after positive and negative feedback.

[19] Around 86% of students who answered this questionnaire responded on the day after the exam, another 5% within two days after the exam. Only around 2% of students answered the post-exam questionnaire 4 or 5 days after exam.

Table 11: Descriptive Statistics from the Post-Exam Questionnaire

	Obs. (1)	Mean (2)	Median (3)	Std. Dev. (4)	Min. (5)	Max. (6)
Difficulty Exam	210	2.15	2.00	0.90	1	5
Difficulty PQ I	210	3.20	3.00	1.05	1	5
Difficulty PQ II	210	3.87	4.00	0.94	1	5
Usefulness PQ I	210	3.17	3.00	1.24	1	5
Usefulness PQ II	210	2.73	3.00	1.08	1	5
Correct recall FB1 Yes/No	210	0.84	1.00	0.36	0	1
Correct recall FB1 Elements	138	0.88	1.00	0.32	0	1
Correct recall FB1 Ordering	122	0.91	1.00	0.29	0	1
Usefulness Feedback PQ I	141	2.39	2.00	1.13	1	5
Feelings Feedback PQI	141	2.86	3.00	0.80	1	5
Correct recall FB2 Yes/No	210	0.89	1.00	0.32	0	1
Correct recall FB2 Points	186	0.95	1.00	0.22	0	1
Usefulness Feedback PQ II	186	2.64	2.50	1.17	1	5

Notes: Descriptive statistics from the post-exam questionnaire.

Table 10: Feelings after Feedback for Treatment Groups

	POSNEG (1)	NEGPOS (2)	(2) vs (1) (3)
Feelings after FB I	3.562	2.250	-1.312***
Feelings after FB II	2.404	3.957	1.552***
Observations	89	92	181

Notes: Feelings about each of the separate feedback elements for the two treatment groups.

Table 11 shows some descriptive statistics from the post-exam questionnaire. Students were asked about the difficulty of the exam from 'very difficult' (1) to 'very easy' (5) and about the relative difficulty of the practice questions I and II with respect to the exam ('much more difficult' (1) to 'much easier' (5)). On average, students perceived the exam as rather difficult and the first (second) set of practice questions as similarly difficult (slightly easier) than the exam. Furthermore, on a scale from 'not useful at all' (1) to 'very useful' (5), participants on average rated the first and second set of practice questions as neutral in terms of their usefulness to prepare for the exam. Interestingly, this differs by whether the best or worst feedback topic was in the exam. Students whose best feedback topic was part of the exam perceived the first set of practice questions to be much more useful (not shown). Similarly, students whose worst topic was part of the exam perceived the first set of practice questions as significantly less useful. No differences can be observed for the second set of practice questions (not shown).

Furthermore, students were asked to recall the feedback they received on both sets of practice questions. All students received feedback in the form of a final score for the second set of practice questions, and 89% of students remember or saw this final score. Of these students, 95% correctly recall their final score. Furthermore, these students on average perceive the feedback as 'neutral' in terms of its usefulness.

Finally, students were also asked to recall feedback on the first set of practice questions which only the two treatment groups received. 84% of participants correctly recall having or not having received this feedback. Of those who correctly remember receiving feedback, 88% correctly recall having received both positive and negative feedback. From these remaining 122 participants, 91% remember the ordering of feedback. The usefulness of the feedback on the first set of practice questions was perceived as slightly lower than that on the second set.

Overall feelings about the feedback received on the first set of practice questions were elicited with the question 'How did you overall feel about the feedback you received on the first set of practice questions?'. The scale ranged from 'very bad' (1) to 'very good' (5) and students on average, state to have felt 'neutral'.

Additionally, I look at whether there is an effect of feedback ordering on any of the post-exam outcomes. Table B.22 shows how all outcomes from the post-exam questionnaire differ by treatment status. Most outcomes don't seem to be affected by students' treatment status. The exception is the answer to the question regarding overall feelings about the feedback on the first set of practice questions. These overall feelings reported one day after the exam are significantly better in the POSNEG group compared to the NEGPOS group. This is confirmed by a regression of the same specifications used in tables B.4 and B.5 with post-exam feelings about feedback as an outcome. Table 12 shows the results of these regressions now using standardized post-exam feelings about feedback. For students in the POSNEG group, the measure of overall feelings is almost 0.9 standard deviations higher than for the NEGPOS group in my preferred specification in column 5. This corresponds to 0.7 points better feelings on the five-point scale. Although these effects are estimated on the slightly smaller sample of students that responded to the post-exam questionnaire, they can most likely be generalized to the main estimation sample since the results from section 3 suggest that there was no differential attrition by treatment status. The results on feelings about feedback point to emotions being a potential mechanism for the treatment effects on motivation which is in line with the hypotheses from psychology mentioned above.

4.6. Heterogeneous treatment effects

In the pre-analysis plan, I specified a series of dimensions of heterogeneity that ex ante were of interest. Due to the relatively small final sample size, the interpretability of

Table 12: Treatment effects on overall feelings about feedback

	Outcome: feelings about feedback (stand.)					
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	0.791***	0.802***	0.829***	0.836***	0.875***	0.859***
	(0.157)	(0.157)	(0.156)	(0.161)	(0.160)	(0.190)
Female × Course Indicator		✓	✓	✓	✓	✓
Feedback topic dummies			✓	✓	✓	✓
Pre-treatment motivation				✓	✓	✓
Pre-treatment performance					✓	✓
Controls						✓
Observations	138	138	138	138	138	136
R^2	0.157	0.183	0.240	0.241	0.260	0.436

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on post-exam overall feelings about the feedback on the first set of practice questions. OLS regressions. Dependent variable: (standardized) feelings about feedback. Col. 1 shows the regression without any controls, col. 2 only includes a female × course indicator for each randomization cell. Col.s 3-5 gradually add dummies for the topics individuals received positive and negative feedback on, and pre-treatment motivation and performance. Finally, col. 6 adds all controls from table 4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

these heterogeneous treatment effects is rather limited, such that I report the results in more detail in appendix A. In the following paragraphs, I will summarize some suggestive evidence from these analyses.

A student's gender was the main dimension of interest because of prior evidence highlighting how women react differently to feedback than men (Buser, 2016; Coffman et al., 2021; Goulas and Megalokonomou, 2021). In my experiment, I do not find clear evidence of any gender differences in reacting to feedback ordering for any of the outcomes (section A.1). Coefficients on the interaction between treatment and the female dummy are mostly positive, with the exception of exam performance where the POSNEG feedback ordering might be harmful for women.

The initial performance level was the second heterogeneity of interest since prior evidence suggests that low- and high-achievers can have different reactions to feedback (Bandiera et al., 2015; Goulas and Megalokonomou, 2021; Hermes et al., 2021). Indeed, I also find that the treatment effect on motivation is smaller the higher an individual's pre-treatment performance was, indicating that relatively lower-achievers are even more motivated by the feedback sequence highlighting their strength first (section A.2).

A further characteristic that might influence the perception of feedback ordering is a student's socio-economic background (SES). I define SES with a dummy indicating

whether at least one parent has university education, hereby assuming that there might be systematic differences in feedback culture between families with and without academic parents. Maybe surprisingly, students from *non-academic* families overall seem to react less to the feedback ordering, e.g. the treatment effect on motivation is fully concentrated among those with at least one academic parent (section A.3).

Lastly, individuals' personality traits do not seem to systematically matter for their reaction to the feedback sequence. Some of the coefficients of the interaction terms with the treatment dummy are significantly different from zero, but no stable pattern can be observed (section A.4).

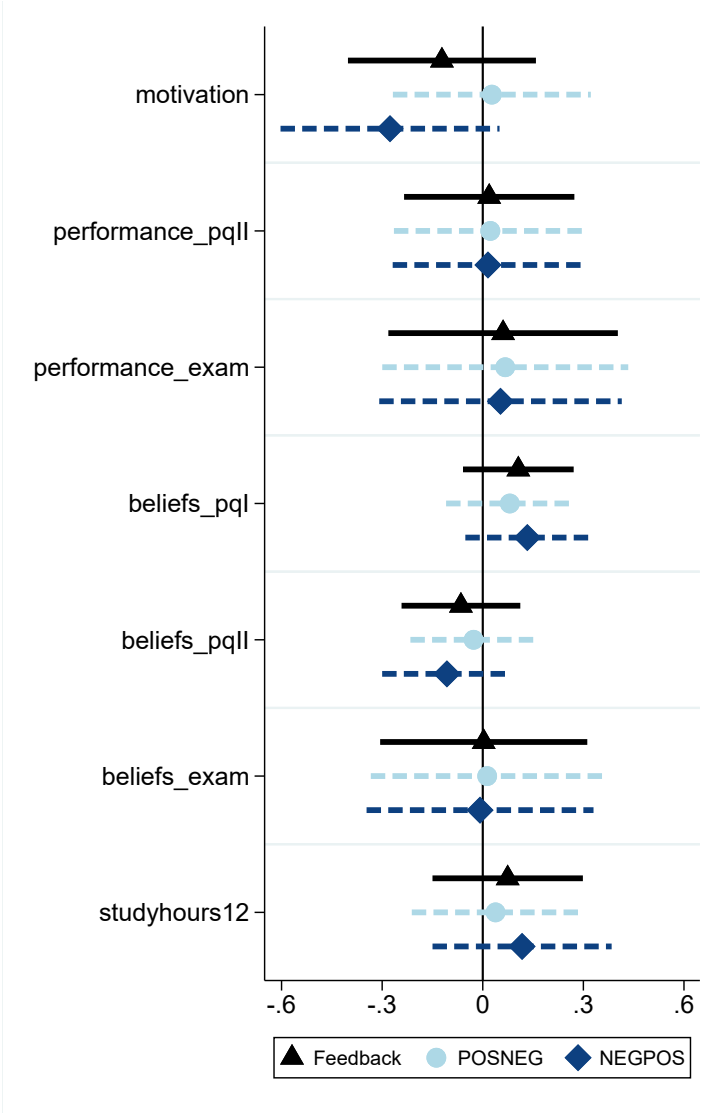
4.7. Feedback vs no feedback

In the following, I want to present some evidence on the effect of feedback compared to no feedback. An important caveat to this analysis is the small sample size of the control group that did not receive feedback on the first set of practice questions (44 individuals). The shares of the groups within the final sample were pre-specified and set to 40% of the sample for each of the treatment groups and 20% for the control group. These proportions were set this way to ensure that there would be a sufficient number of observations for the main comparison of the treatment groups with different feedback ordering, at the expense of potential comparisons with a control group not receiving any feedback.

Furthermore, participants in the control group as well as those in the treatment groups received feedback on the second set of practice questions in form of a total of points out of 10. This implies that any observed effect would actually be the effect of the additional feedback for the first set of practice questions compared to just the one on the second set of practice questions. Nevertheless, this comparison is useful in getting an understanding of whether in general the feedback was useful to students and whether the POSNEG (NEGPOS) ordering actually increased (decreased) motivation overall.

Figure 7 shows the treatment effects of receiving feedback compared to the control group, both on average across the two treatment groups as well as separately. These values come from the regressions of all outcomes either on a dummy for being in one of the treatment groups (black plots) or on dummies for each of the treatment groups separately (blue plots). All regressions contain the full set of controls to increase precision of the estimates. No average effect of receiving feedback vs not receiving feedback on the first set of practice questions is significantly different from zero. Furthermore, none of the coefficients on the separate treatment dummies is significantly different from zero, with the exception of the coefficient on the NEGPOS dummy for motivation (significant at the 10% level). This is in line with figure 2 and shows that in this setting, only receiving feedback in the NEGPOS ordering is statistically different from receiving no feedback on the first set of practice questions.

Figure 7: Effects of feedback vs no feedback



Notes: Plot of coefficients corresponding to equation 1 with dummies for being treated in general (any of the sequences, *Feedback*) or in one of the treatment groups (*POSNEG* and *NEGPOS*) compared to the control group, including the 95% confidence intervals. Dependent variables as described in the legend, all outcomes are standardized across all groups. OLS regressions. All regressions contain a female × course indicator for each of the randomization cells, dummies for the topics individuals received positive and negative feedback on, pre-treatment outcome and performance, and all controls from table 4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of all treatment and control groups.

5. Conclusion

This field experiment varies the ordering of a negative and a positive feedback element on university students' performance in exam practice questions. I find that first giving positive feedback increases post-feedback motivation to study for the exam. Furthermore, students react to the feedback by adjusting their study plan to the content of the feedback elements. I do not find effects of feedback ordering on performance beliefs and average exam performance.

These results hint at feedback operating beyond its pure information channel. Students in both treatment groups get positive and negative feedback on their performance and hence receive the same level of information regarding their pre-treatment level of knowledge. The observed effects therefore suggest an emotional or impulsive reaction, especially to the feedback students receive first. This can be confirmed by a treatment effect of feedback ordering on students' feelings about feedback.

The present study adds to the literature by looking at dynamic reactions to feedback in a field setting. The exam context of university students as a real-life setting with high stakes can provide important insights beyond a lab setting. This is an important step towards understanding how to give feedback in the education context where learning individuals are especially dependent on their instructors' assessment of their performance.

The results from this experiment can also help to shed light on how to motivate students to pursue their studies. In fact, motivation has shown to be related to self-confidence (Bénabou and Tirole, 2002) and negatively associated with dropout from education (Gillet et al., 2012; Cabus and De Witte, 2016; Rump et al., 2017). My findings show that it is not necessary to cut down on the corrective feedback that aims to help students improve their performance. Rather, to avoid a drop in study motivation, it is important to highlight their personal strengths first rather than afterwards.

References

Allgood, S., Badgett, L., Bayer, A., Bertrand, M., Black, S. E., Bloom, N., and Cook, L. D. (2019). AEA professional climate survey: Final report. *American Economic Assocation*.

Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. *The Journal of General Psychology*, 54(2):279–299.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (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):3714–3736.

Azmat, G. and Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, 94(7-8):435–452.

Bandiera, O., Larcinese, V., and Rasul, I. (2015). Blissful ignorance? A natural experiment on the effect of feedback on students' performance. *Labour Economics*, 34:13–25.

Baumeister, R. F., Vohs, K. D., Nathan DeWall, C., and Zhang, L. (2007). How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, 11(2):167–203.

Bénabou, R. and Tirole, J. (2002). Self-confidence and personal motivation. *The Quarterly Journal of Economics*, 117(3):871–915.

Brade, R., Himmler, O., and Jäckle, R. (2022). Relative performance feedback and the effects of being above average-field experiment and replication. *Economics of Education Review*, 89:102268.

Bursztyn, L. and Jensen, R. (2015). How does peer pressure affect educational investments? *The Quarterly Journal of Economics*, 130(3):1329–1367.

Buser, T. (2016). The impact of losing in a competition on the willingness to seek further challenges. *Management Science*, 62(12):3439–3449.

Cabus, S. J. and De Witte, K. (2016). Why do students leave education early? Theory and evidence on high school dropout rates. *Journal of Forecasting*, 35(8):690–702.

Choi, E., Johnson, D. A., Moon, K., and Oah, S. (2018). Effects of positive and negative feedback sequence on work performance and emotional responses. *Journal of Organizational Behavior Management*, 38(2-3):97–115.

Coffman, K. B., Araya, P. U., and Zafar, B. (2021). A (dynamic) investigation of stereotypes, belief-updating, and behavior. *NBER Working Paper*.

Davies, D. and Jacobs, A. (1985). 'Sandwiching' complex interpersonal feedback. *Small Group Behavior*, 16(3):387–396.

Dobrescu, L., Faravelli, M., Megalokonomou, R., and Motta, A. (2021). Relative performance feedback in education: Evidence from a randomised controlled trial. *The Economic Journal*, 131(640):3145–3181.

Dupas, P., Modestino, A. S., Niederle, M., Wolfers, J., et al. (2021). Gender and the dynamics of economics seminars. *NBER Working Paper*.

Eil, D. and Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics*, 3(2):114–38.

Erickson, D., Holderness Jr, D. K., Olsen, K. J., and Thornock, T. A. (2021). Feedback with feeling? How emotional language in feedback affects individual performance. *Accounting, Organizations and Society*, page 101329.

Exley, C. L. and Kessler, J. B. (2022). The gender gap in self-promotion. *The Quarterly Journal of Economics*, 137(3):1345–1381.

Fischer, M. and Wagner, V. (2018). Effects of timing and reference frame of feedback: Evidence from a field experiment. *IZA Discussion Paper*.

Gillet, N., Berjot, S., Vallerand, R. J., and Amoura, S. (2012). The role of autonomy support and motivation in the prediction of interest and dropout intentions in sport and education settings. *Basic and Applied Social Psychology*, 34(3):278–286.

Goulas, S. and Megalokonomou, R. (2021). Knowing who you actually are: The effect of feedback on short-and longer-term outcomes. *Journal of Economic Behavior & Organization*, 183:589–615.

Handlan, A. and Sheng, H. (2023). Gender and tone in recorded economics presentations: Audio analysis with machine learning. *Available at SSRN 4316513*.

Hanushek, E. A., Schwerdt, G., Wiederhold, S., and Woessmann, L. (2015). Returns to skills around the world: Evidence from PIAAC. *European Economic Review*, 73:103–130.

Henley, A. J. and DiGennaro Reed, F. D. (2015). Should you order the feedback sandwich? Efficacy of feedback sequence and timing. *Journal of Organizational Behavior Management*, 35(3-4):321–335.

Hermes, H., Huschens, M., Rothlauf, F., and Schunk, D. (2021). Motivating low-achievers - relative performance feedback in primary schools. *Journal of Economic Behavior & Organization*, 187:45–59.

Ilgen, D. R., Fisher, C. D., and Taylor, M. S. (1979). Consequences of individual feedback on behavior in organizations. *Journal of Applied Psychology*, 64(4):349.

Jacobs, M., Jacobs, A., Gatz, M., and Schaible, T. (1973). Credibility and desirability of positive and negative structured feedback in groups. *Journal of Consulting and Clinical Psychology*, 40(2):244.

Kajitani, S., Morimoto, K., and Suzuki, S. (2020). Information feedback in relative grading: Evidence from a field experiment. *PloS one*, 15(4):e0231548.

Kluger, A. N. and DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing self-confidence: Theory and experimental evidence. *Management Science, forthcoming*.

Muis, K. R., Ranellucci, J., Trevors, G., and Duffy, M. C. (2015). The effects of technology-mediated immediate feedback on kindergarten students' attitudes, emotions, engagement and learning outcomes during literacy skills development. *Learning and Instruction*, 38:1–13.

NEPS-Netzwerk (2021). Nationales Bildungspanel, Scientific Use File der Startkohorte Studierende. Technical report, Leibniz-Institut für Bildungsverläufe (LIfBi), Bamberg.

Rump, M., Esdar, W., and Wild, E. (2017). Individual differences in the effects of academic motivation on higher education students' intention to drop out. *European Journal of Higher Education*, 7(4):341–355.

Schaible, T. D. and Jacobs, A. (1975). Feedback iii: Sequence effects: Enhancement of feedback acceptance and group attractiveness by manipulation of the sequence and valence of feedback. *Small Group Behavior*, 6(2):151–173.

Slowiak, J. M. and Lakowske, A. M. (2017). The influence of feedback statement sequence and goals on task performance. *Behavior Analysis: Research and Practice*, 17(4):357.

Thorndike, E. L. (1913). Educational psychology, vol 1: The original nature of man. *Teachers College*.

Thorndike, E. L. (1927). The law of effect. *The American journal of psychology*, 39(1/4):212–222.

Tyng, C. M., Amin, H. U., Saad, M. N., and Malik, A. S. (2017). The influences of emotion on learning and memory. *Frontiers in Psychology*, 8:1454.

Villeval, M. C. (2022). The cognitive and motivational effects of performance feedback. In *Encyclopedia of Labor Studies*. Edward Elgar Publishers.

Zadra, J. R. and Clore, G. L. (2011). Emotion and perception: The role of affective information. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(6):676–685.

Zax, J. S. and Rees, D. I. (2002). IQ, academic performance, environment, and earnings. *Review of Economics and Statistics*, 84(4):600–616.

Zimmermann, F. (2020). The dynamics of motivated beliefs. *American Economic Review*, 110(2):337–61.

A. Appendix: Report of Heterogeneous Treatment Effects as Specified in the Pre-Analysis Plan

In this section, I will present more detailed results on all pre-specified dimensions of heterogeneity that from a theoretical perspective might have been of interest. Due to the relatively small sample size, the interpretability of these results is limited, but for completeness, I will present all respective estimations.

A.1. Differences by gender

The main dimension along which I expected to find heterogeneous effects, was the gender of participants. In fact, prior evidence on feedback suggests that women react more strongly to feedback, especially to negative elements (Buser, 2016; Coffman et al., 2021; Goulas and Megalokonomou, 2021). Interestingly in my case, no differences by gender can be observed for most of the outcomes, including motivation, as can be seen in table A.1. The only outcome that shows a gender component is exam grades in column 3. Maybe surprisingly, it seems as if only male students benefit from the positive feedback first whereas female students might overall actually be negatively affected.[20] As can be seen in column 7, this does not operate through differential effects on study hours for the last two days before the exam.[21]

Furthermore, figures B.6 and B.7 in appendix B show that this can hardly be explained by differences in reactions to feedback topics. Male students do seem to react slightly more optimistically to the positive feedback topic, i.e. are even more likely to take it away from their study plan than females. This is especially the case after the second (positive) feedback element for the NEGPOS group where women almost don't react at all to the positive feedback topic but men do. In fact, when running a regression according to equation 2 adding female interaction terms for all variables, all interaction terms with dummies from the NEGPOS group are positive but not significant, indicating that women are less likely to remove the positive feedback topic from their study plan in this treatment group (not shown). This is unfortunately also the case for the dummies indicating the initial share of females and males who had that topic on their study list in the NEGPOS group which poses a caveat to the above interpretation, especially since the coefficient on this interaction term is statistically significantly different from zero. For negative-feedback topics, eyeballing the graphs would suggest a similar pattern where women seem to be more reactive to the negative feedback independently of when they received it. Econometrically though, none of the respective dummy×female interaction

[20] The coefficient on this interaction term is not statistically distinguishable from zero anymore after correcting for multiple hypothesis testing (not shown).

[21] All regressions in this section follow the preferred specification of my main analyses without the full set of controls. Results in general look very similar when including all controls from table 4 (not shown).

Table A.1: Differences by gender for all outcomes

	Motivation	Performance		Beliefs			Study
		PQ II	Exam	PQ I	PQ II	Exam	hours
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
POSNEG=1	0.236	-0.303*	0.359*	0.028	-0.095	-0.064	0.025
	(0.188)	(0.181)	(0.194)	(0.134)	(0.259)	(0.095)	(0.152)
Female=1	-0.322	-0.578**	-0.105	-0.363*	-0.534**	-0.396***	0.193
	(0.278)	(0.247)	(0.273)	(0.214)	(0.249)	(0.140)	(0.221)
POSNEG=1 × Female=1	0.102	0.376	-0.456*	-0.096	0.296	0.220	-0.122
	(0.250)	(0.255)	(0.253)	(0.181)	(0.322)	(0.147)	(0.217)
Female × Course Indicator	✓	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓	✓
Pre-treatment motivation	✓						
Pre-treatment performance	✓	✓	✓	✓	✓	✓	✓
Pre-treatment expectations				✓		✓	
Pre-treatment study hours							✓
Observations	181	181	140	181	181	181	181
R^2	0.347	0.335	0.538	0.678	0.120	0.758	0.575

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the exam, performance in the second set of practice questions and the exam, beliefs about past and future performance, and study hours, interacted with being female. OLS regressions. Dependent variable: respective (standardized) outcome. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, pre-treatment performance. Finally, all columns (besides column 5) include pre-treatment outcomes. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

terms is statistically significantly different from zero (not shown).

A.2. Differences by initial performance level

Another dimension of interest is students' initial performance level because of the prior evidence on differences in reactions to feedback between low- and high-achievers (Bandiera et al., 2015; Goulas and Megalokonomou, 2021; Hermes et al., 2021). Indeed, one goal of this paper was to be able to look at differential effects on low- and high-achievers. This is usually not possible when feedback is given relative to others, i.e. where low- (high-) achievers naturally always receive negative (positive) feedback only. In my setting, due to the nature of the feedback as a within-person performance comparison, I can instead look at the effects of early positive feedback on low-achievers. Table A.2 shows treatment effects on individuals depending on their performance in the first set of practice questions which took place before treatment. The only outcomes for which this

comparison seems to matter after both feedback elements were received, are motivation and beliefs about immediate performance. The negative coefficient on the interaction term between receiving positive feedback first and the performance points suggests that for better-performing students the (beneficial) effect of positive feedback first fades with increasing performance.[22]

Table A.2: Differences by initial performance for all outcomes

	Motivation	Performance		Beliefs			Study
		PQ II	Exam	PQ I	PQ II	Exam	hours
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
POSNEG=1	0.287**	-0.094	0.123	-0.025	0.066	0.058	-0.042
	(0.132)	(0.128)	(0.133)	(0.090)	(0.143)	(0.075)	(0.102)
Points PQ I (stand.)	-0.064	0.445***	0.690***	0.131	0.306***	0.048	-0.090
	(0.122)	(0.090)	(0.116)	(0.090)	(0.109)	(0.072)	(0.095)
POSNEG=1 × Points PQ I (stand.)	-0.295**	0.087	-0.153	0.022	-0.353**	-0.015	0.077
	(0.136)	(0.127)	(0.129)	(0.094)	(0.168)	(0.079)	(0.117)
Female × Course Indicator	✓	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓	✓
Pre-treatment motivation	✓						
Pre-treatment expectations				✓		✓	
Pre-treatment study hours							✓
Observations	181	181	140	181	181	181	181
R^2	0.367	0.329	0.531	0.678	0.144	0.755	0.575

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

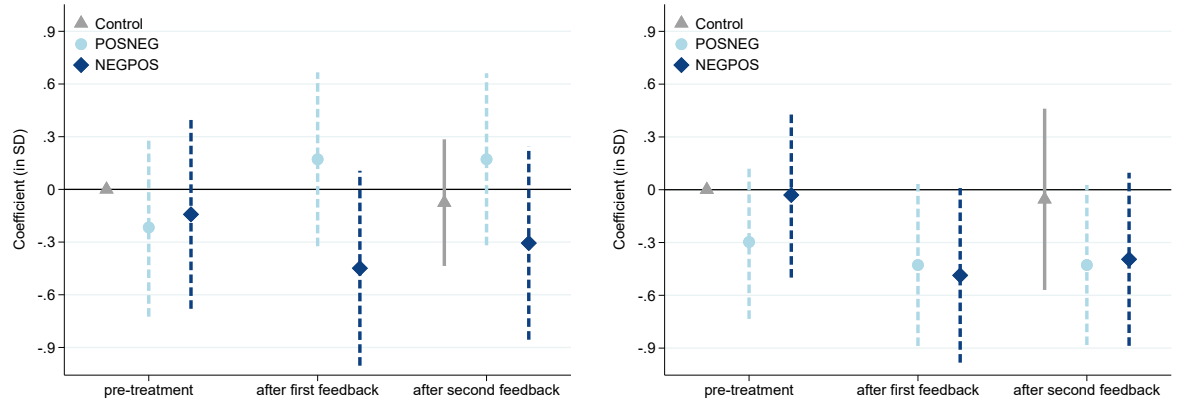
Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the exam, performance in the second set of practice questions and the exam, beliefs about past and future performance, and study hours, interacted with initial performance in the first set of practice questions. OLS regressions. Dependent variable: respective (standardized) outcome. All columns contain a female × course indicator for each of the randomization cells, dummies for the topics individuals received positive and negative feedback on. Finally, all columns (besides column 5) include pre-treatment outcomes. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Again, this can be explored more in detail when looking at the evolution of motivation compared to the control group receiving no feedback on the first set of practice questions. Figure A.1 shows that the pattern observed in figure 2 is only clearly visible for those with below-median performance on the first set of practice questions. Instead, for high-achievers there is a drop in motivation after negative feedback for individuals from the POSNEG group as well. This could have two reasons: either it just reflects a general trend of falling motivation for high-achievers when moving closer to the exam or high-achievers are demotivated by positive and negative feedback equally. The fact that we don't see a clear drop for the control group rather speaks in favor of the second explanation, even though there is no significant difference in feedback aversion as

[22] Again, these coefficients do not seem to be significantly different from zero anymore when correcting for multiple hypothesis testing (not shown).

measured pre-treatment between these two groups (not shown). From a fully saturated regression, I can also infer that the coefficients on POSNEG_fb1 and POSNEG_fb2 are statistically significantly different for the two performance groups. These results again hint towards a ceiling effect for motivation that might in this case be especially pronounced for high-performers. In fact, initial average levels of motivation to study for the exam were significantly higher for those with later above-median performance in the practice questions (not shown).

Figure A.1: Evolution of motivation for pre-treatment performance below-/above-median



(A) Below-median pre-treatment performance

(B) Above-median pre-treatment performance

Notes: Plot of coefficients corresponding to equation 2 with motivation as an outcome, including the 95% confidence intervals, by pre-treatment performance above or below the sample median. The corresponding regression tables can be found in appendix tables B.17 and B.18.

A.3. Differences by socio-economic background

A further dimension of heterogeneity that could theoretically be of interest, is the socio-economic background (SES) of students. Students' socio-economic status could matter for their reaction to the ordering of positive and negative feedback elements if the (order of) feedback they are used to receiving is connected to the parents' academic background. For example, one might assume that parents without academic background might be more likely to emphasize positive feedback regarding their children's performance in the academic context given that it exceeds their own highest level of education and hence might also be more likely to mention this feedback element first.

In this paper, I measure SES as having at least one parent with a university degree (high SES) compared to none (low SES). Table A.3 shows regressions including an interaction term for the treatment variable and an individual's SES measured in this way. The general picture suggests that, for individuals with no academic parent, receiving positive feedback first is less relevant than for those with academic parents. This is especially visible for beliefs where the coefficient on the interaction term is statistically

significant for the first set of practice questions and the exam.[23]

These findings point towards some structurally different ways of how students with (non-)academic parents react to feedback and in particular the ordering of positive and negative elements. One explanation for this could be that these students are particularly highly selected, i.e. only the very well-performing students from this 'low-SES' background are in my sample. Alternatively, there could be a difference in how feedback dynamics operate in their families as described above. With my data, I cannot make any statement about the second channel, but I can have a closer look at these students' background characteristics.

Table A.3: Differences by socio-economic background for all outcomes

	Motivation	Performance		Beliefs			Study
		PQ II	Exam	PQ I	PQ II	Exam	hours
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
POSNEG=1	0.380**	-0.012	0.009	0.094	0.145	0.189**	0.078
	(0.158)	(0.154)	(0.161)	(0.108)	(0.185)	(0.084)	(0.107)
No academic parent=1	0.196	0.017	-0.017	0.130	-0.005	0.238**	0.195
	(0.201)	(0.171)	(0.205)	(0.138)	(0.195)	(0.117)	(0.174)
POSNEG=1 × No academic parent=1	-0.295	-0.294	0.296	-0.411**	-0.270	-0.447**	-0.414*
	(0.282)	(0.288)	(0.281)	(0.195)	(0.307)	(0.183)	(0.241)
Female × Course Indicator	✓	✓	✓	✓	✓	✓	✓
Feedback topic dummies	✓	✓	✓	✓	✓	✓	✓
Pre-treatment motivation	✓						
Pre-treatment performance	✓	✓	✓	✓	✓	✓	✓
Pre-treatment expectations				✓		✓	
Pre-treatment study hours							✓
Observations	181	181	140	181	181	181	181
R^2	0.352	0.334	0.533	0.687	0.122	0.765	0.582

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the exam, performance in the second set of practice questions and the exam, beliefs about past and future performance, and study hours, interacted with socio-economic background as measured by having none or at least one parent with a university degree. OLS regressions. Dependent variable: respective (standardized) outcome. All columns contain a female × course indicator for each of the randomization cells, dummies for the topics individuals received positive and negative feedback on, pre-treatment performance. Finally, all columns (besides column 5) include pre-treatment outcomes as well. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

In fact, students with different socio-economic backgrounds seem to be systematically different on some characteristics as can be seen from table B.19. Students with 'low SES' as measured in this setting tend to be older and in higher semesters, they choose different majors and have worse high-school outcomes (on a German scale). Most interestingly, they markedly differ with respect to their personality traits, i.e. individuals

[23] Again, none of the significance levels of the presented coefficients survives the correction for multiple hypothesis testing.

with non-academic parents report being less conscientious, i.e. more lazy/less thorough, and more neurotic, i.e. more stressed and anxious. Finally, they tend to be less motivated to study for university in general, but have invested more hours thus far in studying for this course. These pieces of evidence suggest that rather the second mechanism described above is what can explain differences in these students' reactions to feedback ordering since the 'low-SES' students from this study are not particularly highly selected.

A.4. Differences by personality traits

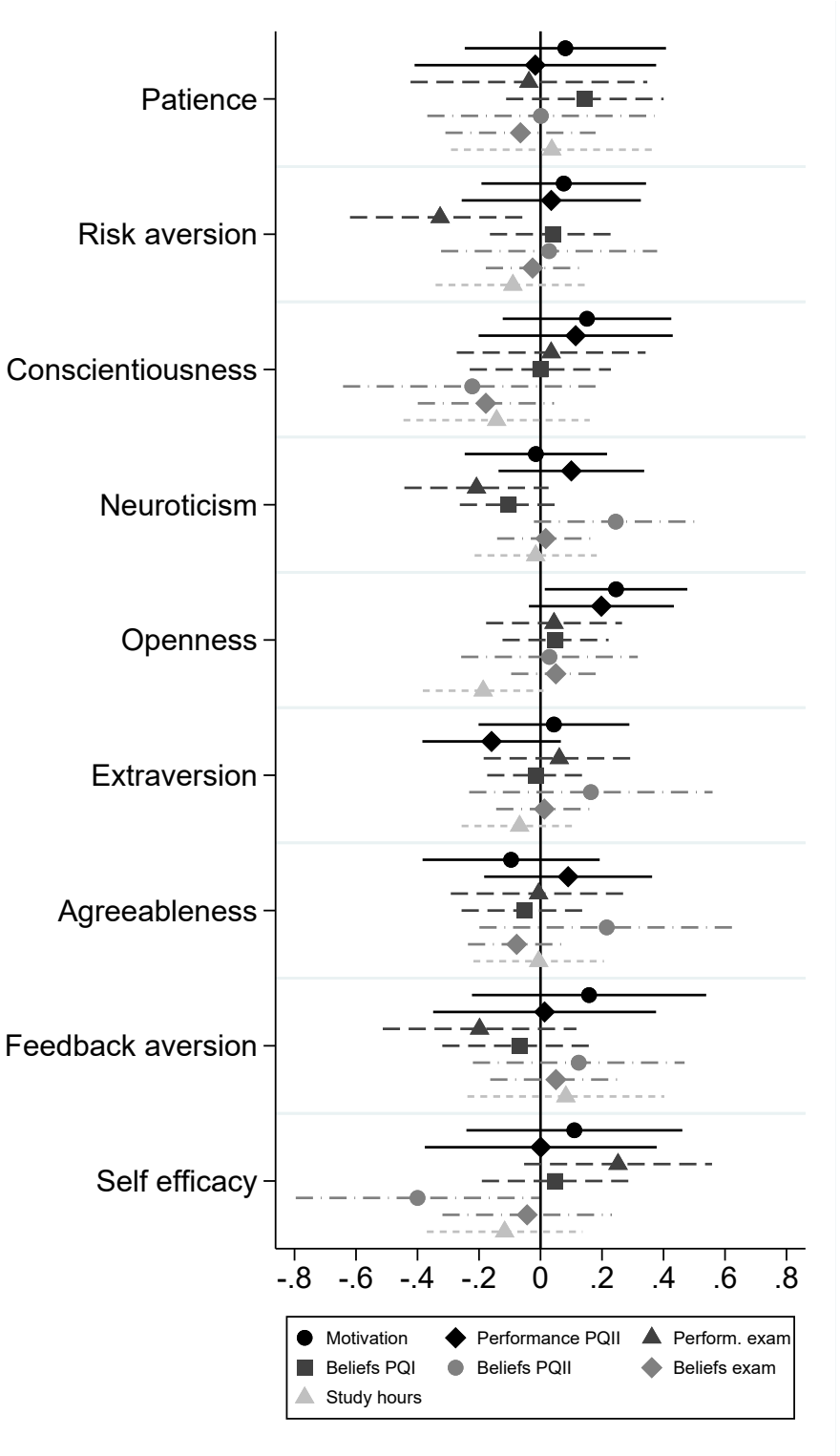
Similar heterogeneity analyses can be run with patience, risk-aversion, the big five personality traits, feedback aversion, and self-efficacy. Figure A.2 shows a plot for the coefficients on the interaction term of the treatment dummy and the respective trait from regressions of all presented outcomes on all mentioned traits.

Overall, there is no clear pattern of any trait influencing treatment effects into a certain direction, but there are some exceptions that might be worth looking into.[24] For example, risk aversion seems to play a role for the treatment effects on exam grades: the more risk-averse students are, the less important it is for them to receive the positive feedback element first (while more risk-averse students in general have higher grades, not shown). Similarly, more neurotic students benefit less from first receiving positive feedback first for their exam performance. On the other hand, neurotic students benefit more from the *positive-negative* feedback order for beliefs about their immediate performance in the second set of practice questions which suggests that neuroticism operates very differently for intuitive reactions compared to medium-term processes. This seems to hold true for openness as well: more open students benefit more from the *positive-negative* ordering both for motivation as well as for their immediate performance, but they also increase their study hours on the last two days before the exam less compared to less open students.

Lastly, self-efficacy plays a role for the treatment effect of the feedback sequence on beliefs about immediate performance after the feedback: more self-efficacious students benefit less from first receiving positive feedback in having higher post-feedback beliefs about their performance in the second set of practice questions (overall, self-efficacy is positively related to immediate performance beliefs, not shown). Maybe surprisingly, no clear patterns can be observed for patience, conscientiousness, extraversion, agreeableness, and feedback aversion.

[24] None of the presented statistically significant coefficients remains significantly different from zero according to conventional levels once I correct for multiple hypothesis testing for the regressions of each trait (not shown). The following descriptions hence are more suggestive evidence.

Figure A.2: Plot of all coefficients on personality trait interactions for all outcomes

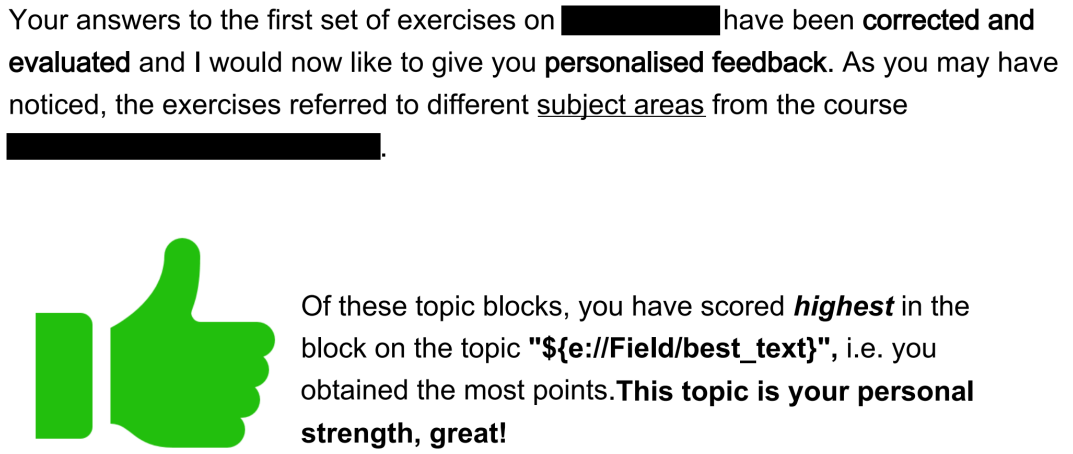


Notes: Coefficient estimates on the interaction term between a treatment dummy and the respective personality trait, including the 95% confidence intervals. Dependent variables as described in the legend, all outcomes are standardized within the treatment groups. OLS regressions. All regressions contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and pre-treatment outcomes and performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

B. Appendix: Figures and Tables

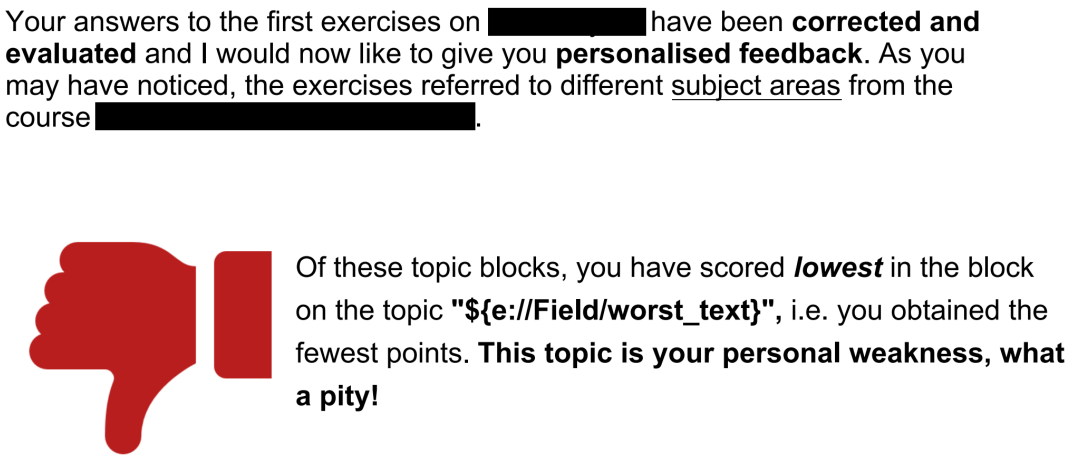
B.1. Figures

Figure B.1: Example of positive feedback (translated from German)



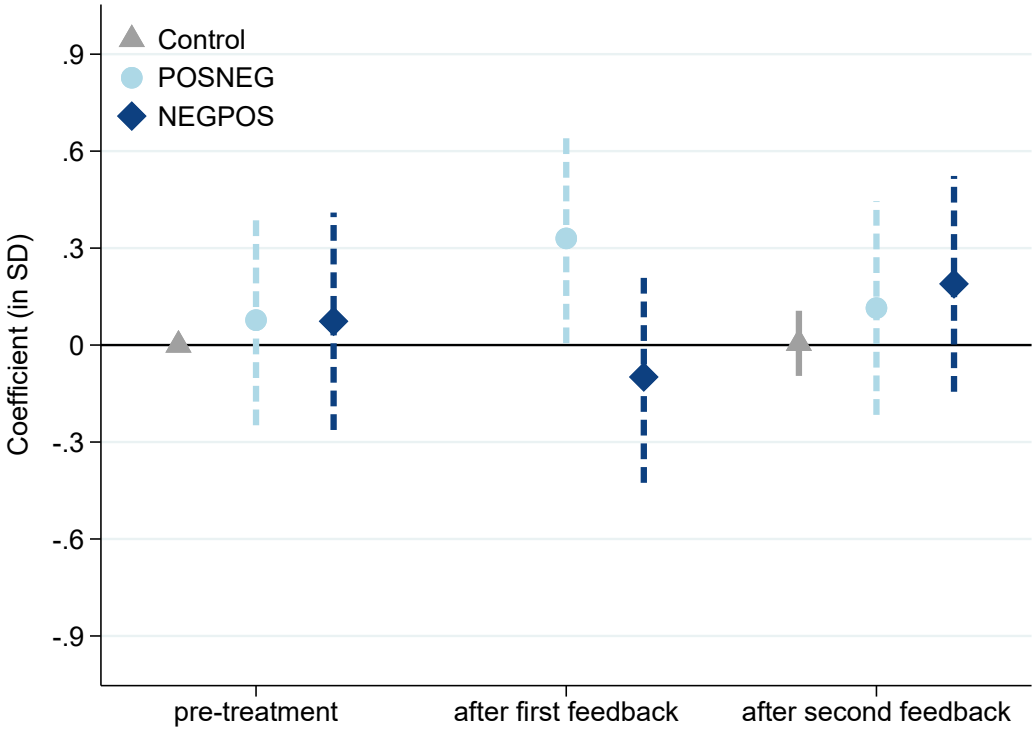
Notes: Screenshot from the feedback questionnaire. Details on the specific course and date have been blackened.

Figure B.2: Example of negative feedback (translated from German)



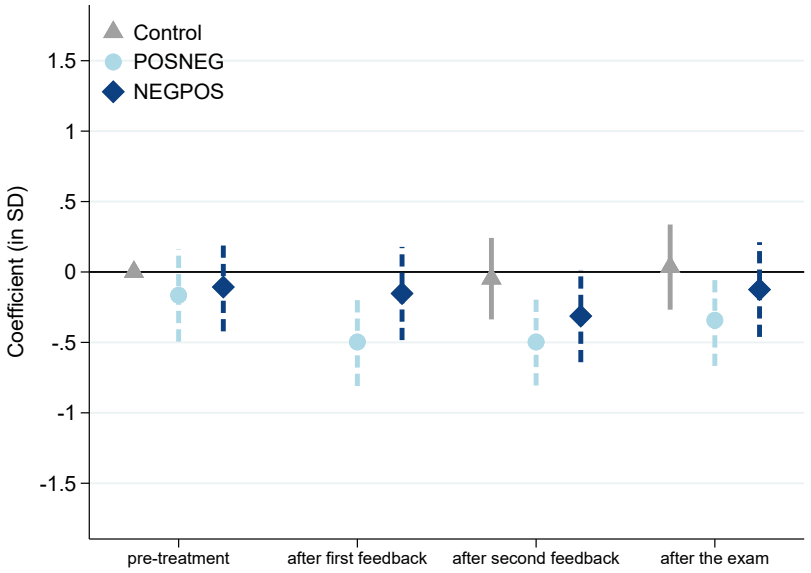
Notes: Screenshot from the feedback questionnaire. Details on the specific course and date have been blackened.

Figure B.3: Evolution of performance beliefs about PQI



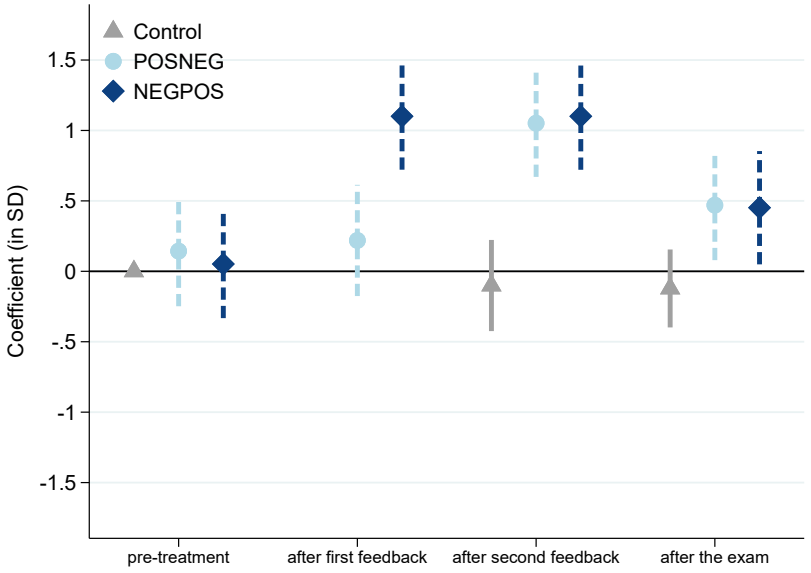
Notes: Plot of coefficients corresponding to equation 2 with standardized beliefs about performance in the first set of practice questions, including the 95% confidence intervals. The corresponding regression table can be found in appendix table B.7.

Figure B.4: Evolution of whether **positive** feedback topic is on study list, incl. post exam



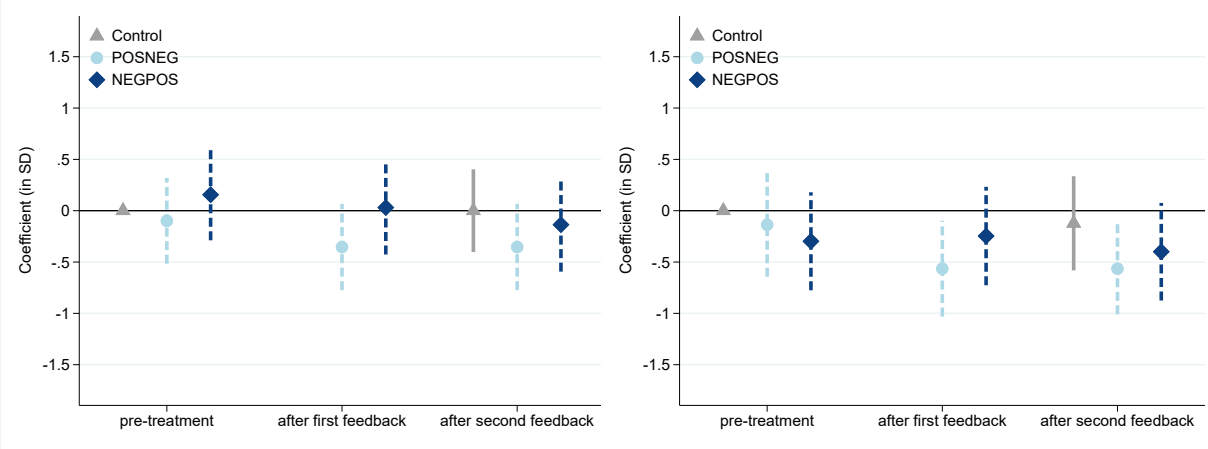
Notes: Plot of coefficients corresponding to equation 2 with an indicator of whether a person has the topic on their study list they have or will receive positive feedback on as an outcome, including the 95% confidence intervals. This version is enriched by the students' assessment after the exam. The corresponding regression table can be found in appendix table B.10.

Figure B.5: Evolution of whether **negative** feedback topic is on study list, incl. post exam



Notes: Plot of coefficients corresponding to equation 2 with an indicator of whether a person has the topic on their study list they have or will receive negative feedback on as an outcome, including the 95% confidence intervals. This version is enriched by the students' assessment after the exam. The corresponding regression table can be found in appendix table B.12.

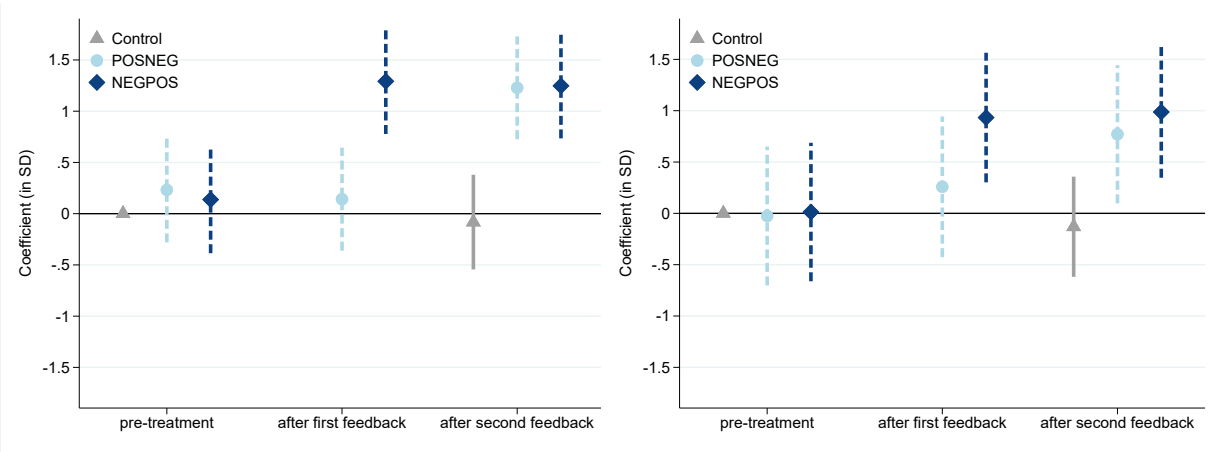
Figure B.6: Evolution of whether **positive** feedback topic is on study list, by gender



(A) Females (B) Males

Notes: Plot of coefficients corresponding to equation 2 with an indicator of whether a person has the topic on their study list they have or will receive positive feedback on as an outcome, including the 95% confidence intervals, by gender. The corresponding regression table can be found in appendix tables B.13 and B.14.

Figure B.7: Evolution of whether **negative** feedback topic is on study list, by gender



(A) Females (B) Males

Notes: Plot of coefficients corresponding to equation 2 with an indicator of whether a person has the topic on their study list they have or will receive negative feedback on as an outcome, including the 95% confidence intervals, by gender. The corresponding regression table can be found in appendix tables B.15 and B.16.

Table B.1: Descriptive Statistics

	Obs. (1)	Mean (2)	Median (3)	Std. Dev. (4)	Min. (5)	Max. (6)
Motivation study exam	225	3.54	4.00	1.06	1	5
Points PQ I	225	8.46	8.50	3.82	0	18
Before answering						
Expected points PQ I	225	12.60	13.00	3.58	1	20
Self evaluation exp. pts PQI	225	2.87	3.00	1.05	1	5
Expected prob. Q1 (PQI)	225	17.72	10.00	21.47	0	100
Expected prob. Q2 (PQI)	225	30.46	30.00	20.01	0	100
Expected prob. Q3 (PQI)	225	32.88	30.00	20.04	0	90
Expected prob. Q4 (PQI)	225	18.94	10.00	23.21	0	100
After answering						
Expected points PQ I	225	8.72	8.00	4.27	0	18
Self evaluation exp. pts PQI	225	1.89	2.00	0.92	1	5
Expected prob. Q1 (PQI)	225	37.64	30.00	33.48	0	100
Expected prob. Q2 (PQI)	225	33.57	35.00	23.11	0	100
Expected prob. Q3 (PQI)	225	20.94	15.00	21.10	0	80
Expected prob. Q4 (PQI)	225	7.84	0.00	16.86	0	100
Expected grade (German scale)	225	2.31	2.30	0.69	1	4
Self evaluation exp. grade	225	3.22	3.00	1.06	1	5
Expected prob. Q1 (grade)	225	12.92	5.00	18.56	0	100
Expected prob. Q2 (grade)	225	25.35	25.00	18.43	0	80
Expected prob. Q3 (grade)	225	36.27	35.00	20.31	0	100
Expected prob. Q4 (grade)	225	25.46	10.00	28.42	0	100
Planned study hours (-6)	225	2.44	2.00	1.92	0	12
Planned study hours (-5)	225	2.48	2.00	1.89	0	10
Planned study hours (-4)	225	2.13	2.00	1.83	0	8
Planned study hours (-3)	225	2.04	2.00	1.80	0	12
Planned study hours (-2)	225	3.14	3.00	2.45	0	12
Planned study hours (-1)	225	4.21	4.00	2.47	0	12
Positive fb topic on study list	225	0.27	0.00	0.44	0	1
Negative fb topic on study list	225	0.29	0.00	0.46	0	1

Notes: Descriptive statistics of all pre-treatment outcomes. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course.

Table B.2: Balancing check for pre-treatment outcomes

	Control	POSNEG	NEGPOS	(2) vs (1)	(3) vs (1)	(2) vs (3)
	(1)	(2)	(3)	(4)	(5)	(6)
Motivation study exam	3.591	3.461	3.598	-0.130	0.007	-0.137
Points PQ I	7.852	8.584	8.636	0.732	0.784	-0.052
Before answering						
Expected points PQ I (pre)	12.136	12.910	12.511	0.774	0.375	0.399
Self evaluation exp. pts PQI (pre)	2.773	2.955	2.826	0.182	0.053	0.129
Expected prob. Q1 (PQI) (pre)	24.182	14.079	18.152	-10.103***	-6.030	-4.074
Expected prob. Q2 (PQI) (pre)	29.295	29.157	32.272	-0.138	2.976	-3.114
Expected prob. Q3 (PQI) (pre)	31.045	37.303	29.489	6.258*	-1.556	7.814***
Expected prob. Q4 (PQI) (pre)	15.477	19.461	20.087	3.983	4.610	-0.626
After answering						
Expected points PQ I (post)	7.932	9.090	8.728	1.158	0.796	0.362
Self evaluation exp. pts PQI (post)	1.886	1.933	1.859	0.046	-0.028	0.074
Expected prob. Q1 (PQI) (post)	40.227	33.764	40.163	-6.463	-0.064	-6.399
Expected prob. Q2 (PQI) (post)	33.523	37.056	30.228	3.533	-3.294	6.828**
Expected prob. Q3 (PQI) (post)	18.955	23.506	19.402	4.551	0.448	4.103
Expected prob. Q4 (PQI) (post)	7.295	5.674	10.207	-1.621	2.911	-4.532*
Expected grade (German scale)	2.486	2.275	2.254	-0.211*	-0.232*	0.021
Self evaluation exp. grade	3.159	3.157	3.304	-0.002	0.145	-0.147
Expected prob. Q1 (grade)	17.364	10.315	13.315	-7.049**	-4.048	-3.001
Expected prob. Q2 (grade)	26.750	24.910	25.109	-1.840	-1.641	-0.199
Expected prob. Q3 (grade)	35.227	40.067	33.098	4.840	-2.129	6.970**
Expected prob. Q4 (grade)	20.659	24.708	28.478	4.049	7.819	-3.770
Planned study hours (-6)	2.455	2.669	2.223	0.214	-0.232	0.446
Planned study hours (-5)	2.557	2.562	2.353	0.005	-0.204	0.209
Planned study hours (-4)	1.739	2.185	2.261	0.447	0.522	-0.075
Planned study hours (-3)	1.841	2.247	1.937	0.406	0.096	0.310
Planned study hours (-2)	2.591	3.163	3.370	0.572	0.779*	-0.207
Planned study hours (-1)	3.932	4.208	4.337	0.276	0.405	-0.129
Positive fb topic on study list	0.341	0.213	0.283	-0.127	-0.058	-0.069
Negative fb topic on study list	0.273	0.326	0.272	0.053	-0.001	0.054
Observations	44	89	92	133	136	181

Notes: Balancing check between control and treatment groups on all pre-treatment outcomes. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course.

Table B.3: Treatment effects on motivation with LASSO regressions

	LASSO		Double LASSO	
	(restricted)	(unrestricted)	(restricted)	(unrestricted)
	(1)	(2)	(3)	(4)
POSNEG	0.298**	0.243**	0.290**	0.236*
	(0.122)	(0.123)	(0.129)	(0.121)
Variables selected				
...for outcome	female × course	pre-treatment motivation	female × course	pre-treatment motivation
	feedback topic dummies	major 'Bus. and Econ. Education'	feedback topic dummies	second semester dummy
	pre-treatment motivation	patience	pre-treatment motivation	conscientiousness
	pre-treatment performance	conscientiousness	pre-treatment performance	neuroticism
	major 'Bus. and Econ. Education'	neuroticism	conscientiousness	general study motivation
	father university degree	general study motivation		
	patience	feedback-aversion		
	risk-aversion			
	conscientiousness			
	neuroticism			
...for treatment (double LASSO only)			female × course	
			feedback topic dummies	
			pre-treatment motivation	
			pre-treatment performance	
N	179	179	179	179

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the exam. LASSO (*lasso* in STATA, col.s 1-2) and double LASSO (*dsregress* in STATA, col.s 3-4) regressions with selection of covariates and default parameters. Dependent variable: (standardized) motivation to study for the exam. Columns 1 and 3 select choice of controls on those described in table 4 whereas col.s 2 and 4 also allow to select among the female × course indicator, feedback topic dummies, and pre-treatment motivation and performance. Sample comprises of treatment groups only, the control group is excluded here.

Table B.4: Treatment effects on motivation between feedback elements

	Outcome: motivation to study (stand.)					
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	0.302**	0.305**	0.343**	0.399***	0.411***	0.434***
	(0.147)	(0.147)	(0.151)	(0.130)	(0.128)	(0.136)
Female × Course Indicator		✓	✓	✓	✓	✓
Feedback topic dummies			✓	✓	✓	✓
Pre-treatment motivation				✓	✓	✓
Pre-treatment performance					✓	✓
Controls						✓
Observations	181	181	181	181	181	179
R^2	0.023	0.045	0.144	0.343	0.366	0.512

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the exam between feedback elements. OLS regressions. Dependent variable: (standardized) motivation to study for the exam. Col. 1 shows the regression without any controls, col. 1 only includes a female × course indicator for each randomization cell. Col.s 2-5 gradually add dummies for the topics individuals received positive and negative feedback on, and pre-treatment motivation and performance. Finally, col. 6 adds all controls from table 4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table B.5: Treatment effects on motivation after both feedback elements

	Outcome: motivation to study (stand.)					
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	0.201	0.202	0.220	0.278**	0.292**	0.349**
	(0.148)	(0.148)	(0.154)	(0.135)	(0.134)	(0.136)
Female × Course Indicator		✓	✓	✓	✓	✓
Feedback topic dummies			✓	✓	✓	✓
Pre-treatment motivation				✓	✓	✓
Pre-treatment performance					✓	✓
Controls						✓
Observations	181	181	181	181	181	179
R^2	0.010	0.029	0.104	0.315	0.347	0.498

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on motivation to study for the exam after both feedback elements. OLS regressions. Dependent variable: (standardized) motivation to study for the exam. Col. 1 shows the regression without any controls, col. 1 only includes a female × course indicator for each randomization cell. Col.s 2-5 gradually add dummies for the topics individuals received positive and negative feedback on, and pre-treatment motivation and performance. Finally, col. 6 adds all controls from table 4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table B.6: Evolution of motivation to study for the exam

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.192	-0.180	-0.273*
	(0.160)	(0.159)	(0.164)
After fb1 POSNEG	-0.069	-0.057	-0.146
	(0.162)	(0.160)	(0.168)
After fb2 POSNEG	-0.069	-0.057	-0.171
	(0.162)	(0.161)	(0.168)
Initial NEGPOS	-0.070	-0.058	-0.066
	(0.160)	(0.160)	(0.166)
After fb1 NEGPOS	-0.448**	-0.436**	-0.437**
	(0.174)	(0.174)	(0.180)
After fb2 NEGPOS	-0.329*	-0.317*	-0.317*
	(0.172)	(0.172)	(0.177)
After-treatment control	-0.067	-0.067	-0.074
	(0.145)	(0.144)	(0.138)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	625	625	557
R^2	0.408	0.405	0.406

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure 2 using equation 2. Dependent variable: (standardized) motivation to study for the exam. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.7: Evolution of performance beliefs about PQI

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	0.077	0.112	0.004
	(0.165)	(0.166)	(0.173)
After fb1 POSNEG	0.330**	0.364**	0.268
	(0.165)	(0.164)	(0.176)
After fb2 POSNEG	0.115	0.149	0.046
	(0.168)	(0.167)	(0.179)
Initial NEGPOS	0.074	0.089	-0.013
	(0.171)	(0.174)	(0.180)
After fb1 NEGPOS	-0.099	-0.083	-0.173
	(0.166)	(0.170)	(0.175)
After fb2 NEGPOS	0.189	0.205	0.081
	(0.169)	(0.172)	(0.178)
After-treatment control	0.005	0.005	-0.006
	(0.051)	(0.051)	(0.056)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	625	625	557
R^2	0.430	0.408	0.446

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure B.3 using equation 2. Dependent variable: (standardized) performance beliefs about practice questions 1. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Col. 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, col. 2 excludes high-school performance as a control, and col. 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.8: Treatment effects on study hours days 1-2 before the exam

	Outcome: study hours (stand.)					
	(1)	(2)	(3)	(4)	(5)	(6)
POSNEG	-0.122	-0.120	-0.113	-0.045	-0.043	-0.073
	(0.148)	(0.143)	(0.144)	(0.102)	(0.103)	(0.120)
Female × Course Indicator		✓	✓	✓	✓	✓
Feedback topic dummies			✓	✓	✓	✓
Pre-treatment study plan				✓	✓	✓
Pre-treatment performance					✓	✓
Controls						✓
Observations	181	181	181	181	181	179
R^2	0.004	0.089	0.124	0.572	0.574	0.661

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Treatment effects of *positive-negative* feedback compared to *negative-positive* feedback (base group) on total study hours one and two days before the exam. OLS regressions. Dependent variable: (standardized) total hours students planned to study for the exam on the last two days before the exam. Col. 1 shows the regression without any controls, col. 2 only includes a female × course indicator for each randomization cell. Col.s 3-5 gradually add dummies for the topics individuals received positive and negative feedback on, and pre-treatment planned study hours and performance. Finally, col. 6 adds all controls from table 4, including an imputation dummy for individuals who did not report their high-school performance. Robust standard errors are used throughout. Sample comprises of treatment groups only, the control group is excluded here.

Table B.9: Evolution of positive feedback topics

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.168	-0.165	-0.165
	(0.169)	(0.167)	(0.186)
After fb1 POSNEG	-0.500***	-0.496***	-0.544***
	(0.160)	(0.159)	(0.178)
After fb2 POSNEG	-0.500***	-0.496***	-0.544***
	(0.159)	(0.157)	(0.176)
Initial NEGPOS	-0.098	-0.098	-0.198
	(0.161)	(0.161)	(0.175)
After fb1 NEGPOS	-0.144	-0.144	-0.224
	(0.169)	(0.168)	(0.183)
After fb2 NEGPOS	-0.304*	-0.304*	-0.376**
	(0.167)	(0.166)	(0.182)
After-treatment control	-0.047	-0.047	-0.104
	(0.148)	(0.148)	(0.153)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	625	625	557
R^2	0.329	0.328	0.343

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure 3 using equation 2. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive positive feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.10: Evolution of positive feedback topics, incl. post exam

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.165	-0.166	-0.163
	(0.167)	(0.165)	(0.184)
After fb1 POSNEG	-0.497***	-0.497***	-0.542***
	(0.159)	(0.157)	(0.175)
After fb2 POSNEG	-0.497***	-0.497***	-0.542***
	(0.157)	(0.155)	(0.174)
Post-exam POSNEG	-0.343**	-0.344**	-0.357*
	(0.165)	(0.163)	(0.183)
Initial NEGPOS	-0.107	-0.110	-0.207
	(0.160)	(0.159)	(0.173)
After fb1 NEGPOS	-0.153	-0.156	-0.233
	(0.168)	(0.167)	(0.182)
After fb2 NEGPOS	-0.313*	-0.316*	-0.385**
	(0.166)	(0.166)	(0.181)
Post-exam NEGPOS	-0.125	-0.127	-0.186
	(0.171)	(0.170)	(0.185)
After-treatment control	-0.047	-0.047	-0.104
	(0.147)	(0.147)	(0.152)
Post-exam control	0.035	0.035	0.027
	(0.153)	(0.153)	(0.167)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	833	833	743
R^2	0.329	0.328	0.345

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure B.4 using the extended version of equation 2. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive positive feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.11: Evolution of negative-feedback topics

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	0.155	0.139	0.202
	(0.201)	(0.200)	(0.201)
After fb1 POSNEG	0.230	0.215	0.318
	(0.202)	(0.202)	(0.204)
After fb2 POSNEG	1.063***	1.047***	1.154***
	(0.197)	(0.197)	(0.195)
Initial NEGPOS	0.048	0.038	0.081
	(0.199)	(0.200)	(0.198)
After fb1 NEGPOS	1.097***	1.087***	1.137***
	(0.195)	(0.195)	(0.193)
After fb2 NEGPOS	1.097***	1.087***	1.137***
	(0.196)	(0.196)	(0.194)
After-treatment control	-0.101	-0.101	0.000
	(0.166)	(0.165)	(0.164)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	625	625	557
R^2	0.335	0.333	0.339

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure 4 using equation 2. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive negative feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.12: Evolution of negative-feedback topics, incl. post exam

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	0.144	0.131	0.200
	(0.199)	(0.200)	(0.201)
After fb1 POSNEG	0.219	0.207	0.316
	(0.200)	(0.202)	(0.204)
After fb2 POSNEG	1.052***	1.039***	1.152***
	(0.193)	(0.195)	(0.194)
Post-exam POSNEG	0.470**	0.457**	0.510**
	(0.198)	(0.199)	(0.199)
Initial NEGPOS	0.052	0.047	0.105
	(0.195)	(0.196)	(0.193)
After fb1 NEGPOS	1.101***	1.096***	1.161***
	(0.192)	(0.193)	(0.191)
After fb2 NEGPOS	1.101***	1.096***	1.161***
	(0.193)	(0.193)	(0.191)
Post-exam NEGPOS	0.452**	0.446**	0.484**
	(0.204)	(0.205)	(0.204)
After-treatment control	-0.101	-0.101	-0.000
	(0.164)	(0.164)	(0.163)
Post-exam control	-0.122	-0.123	-0.091
	(0.140)	(0.140)	(0.143)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	833	833	743
R^2	0.301	0.300	0.306

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to figure B.5 using the extended version of equation 2. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive negative feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.13: Evolution of positive feedback topics, females only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.099	-0.110	-0.124
	(0.211)	(0.212)	(0.233)
After fb1 POSNEG	-0.354*	-0.365*	-0.437*
	(0.213)	(0.211)	(0.235)
After fb2 POSNEG	-0.354*	-0.365*	-0.437*
	(0.213)	(0.211)	(0.235)
Initial NEGPOS	0.156	0.147	-0.004
	(0.224)	(0.223)	(0.242)
After fb1 NEGPOS	0.031	0.022	-0.093
	(0.231)	(0.230)	(0.254)
After fb2 NEGPOS	-0.136	-0.145	-0.270
	(0.231)	(0.230)	(0.251)
After-treatment control	0.000	0.000	-0.087
	(0.203)	(0.202)	(0.210)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	351	351	309
R^2	0.436	0.433	0.446

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (A) of figure B.6 using equation 2 for females only. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive positive feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.14: Evolution of positive feedback topics, males only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.137	-0.097	-0.171
	(0.255)	(0.250)	(0.274)
After fb1 POSNEG	-0.565**	-0.524**	-0.622**
	(0.235)	(0.226)	(0.253)
After fb2 POSNEG	-0.565**	-0.524**	-0.622**
	(0.224)	(0.218)	(0.241)
Initial NEGPOS	-0.298	-0.276	-0.348
	(0.240)	(0.237)	(0.255)
After fb1 NEGPOS	-0.247	-0.225	-0.288
	(0.242)	(0.244)	(0.254)
After fb2 NEGPOS	-0.399*	-0.378	-0.407
	(0.239)	(0.242)	(0.252)
After-treatment control	-0.123	-0.123	-0.130
	(0.231)	(0.229)	(0.247)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	274	274	248
R^2	0.378	0.367	0.385

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (B) of figure B.6 using equation 2 for males only. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive positive feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.15: Evolution of negative-feedback topics, females only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	0.232	0.199	0.465*
	(0.259)	(0.253)	(0.253)
After fb1 POSNEG	0.141	0.109	0.410
	(0.254)	(0.249)	(0.250)
After fb2 POSNEG	1.229***	1.196***	1.520***
	(0.253)	(0.249)	(0.246)
Initial NEGPOS	0.138	0.116	0.368
	(0.264)	(0.262)	(0.254)
After fb1 NEGPOS	1.292***	1.271***	1.501***
	(0.260)	(0.256)	(0.254)
After fb2 NEGPOS	1.247***	1.226***	1.454***
	(0.259)	(0.255)	(0.253)
After-treatment control	-0.082	-0.082	0.092
	(0.233)	(0.232)	(0.223)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	351	351	309
R^2	0.408	0.403	0.417

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (A) of figure B.7 using equation 2 for females only. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive negative feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.16: Evolution of negative-feedback topics, males only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.026	-0.018	-0.131
	(0.341)	(0.337)	(0.344)
After fb1 POSNEG	0.259	0.266	0.169
	(0.345)	(0.342)	(0.341)
After fb2 POSNEG	0.771**	0.779**	0.709**
	(0.339)	(0.336)	(0.337)
Initial NEGPOS	0.013	0.025	-0.042
	(0.340)	(0.328)	(0.359)
After fb1 NEGPOS	0.933***	0.945***	0.909***
	(0.318)	(0.310)	(0.332)
After fb2 NEGPOS	0.987***	1.000***	0.972***
	(0.323)	(0.313)	(0.340)
After-treatment control	-0.131	-0.131	-0.139
	(0.246)	(0.244)	(0.263)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	274	274	248
R^2	0.361	0.359	0.376

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (B) of figure B.7 using equation 2 for males only. Dependent variable: indicator of whether a person has the topic on their study list they have or will receive negative feedback on. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.17: Evolution of motivation, below-median performers only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.217	-0.233	-0.286
	(0.256)	(0.256)	(0.281)
After fb1 POSNEG	0.171	0.155	0.112
	(0.250)	(0.248)	(0.264)
After fb2 POSNEG	0.171	0.155	0.059
	(0.247)	(0.247)	(0.263)
Initial NEGPOS	-0.142	-0.162	-0.259
	(0.271)	(0.267)	(0.280)
After fb1 NEGPOS	-0.449	-0.468*	-0.543*
	(0.280)	(0.279)	(0.286)
After fb2 NEGPOS	-0.306	-0.325	-0.390
	(0.278)	(0.277)	(0.284)
After-treatment control	-0.075	-0.075	-0.089
	(0.182)	(0.181)	(0.153)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	325	325	290
R^2	0.448	0.445	0.407

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (A) of figure A.1 using equation 2 only for those with below-median performance in the first set of practice questions. Dependent variable: (standardized) motivation to study for the exam. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.18: Evolution of motivation, above-median performers only

	High-school performance		
	Imputed	None	Only reported
	(1)	(2)	(3)
Initial POSNEG	-0.297	-0.176	-0.384*
	(0.220)	(0.220)	(0.226)
After fb1 POSNEG	-0.428*	-0.307	-0.506**
	(0.232)	(0.231)	(0.232)
After fb2 POSNEG	-0.428*	-0.307	-0.506**
	(0.229)	(0.230)	(0.229)
Initial NEGPOS	-0.030	0.085	0.164
	(0.237)	(0.231)	(0.233)
After fb1 NEGPOS	-0.487*	-0.371	-0.313
	(0.250)	(0.246)	(0.241)
After fb2 NEGPOS	-0.395	-0.280	-0.234
	(0.248)	(0.245)	(0.239)
After-treatment control	-0.055	-0.055	-0.055
	(0.260)	(0.258)	(0.262)
Female × Course Indicator	✓	✓	✓
Feedback topic dummies	✓	✓	✓
Controls	✓	✓	✓
Observations	300	300	267
R^2	0.539	0.520	0.607

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Regression table corresponding to panel (B) of figure A.1 using equation 2 only for those with above-median performance in the first set of practice questions. Dependent variable: (standardized) motivation to study for the exam. All columns contain a female × course indicator for each randomization cell, dummies for the topics individuals received positive and negative feedback on, and all controls from table 4. Column 1 additionally includes an imputation dummy for individuals who did not report their high-school performance and imputes the missing value with the sample mean, column 2 excludes high-school performance as a control, and column 3 only uses the subset of individuals who completed high school in Germany and reported their final high-school performance. Standard errors are clustered at the individual level. Sample comprises of treatment groups as well as the control group.

Table B.19: Balancing check for all control variables, by SES

	No academic parent (1)	At least one academic parent (2)	(1) vs (2) (3)
Female	0.662	0.525	0.137*
Age	21.369	19.692	1.677***
Business Administration	0.554	0.556	-0.002
Economics	0.092	0.200	-0.108*
Bus. and Econ. Education	0.077	0.050	0.027
Minor Bus./Econ./Education	0.092	0.131	-0.039
Other major	0.185	0.063	0.122***
Semester	2.200	1.825	0.375**
German high school degree	0.908	0.900	0.008
High school GPA	1.825	1.698	0.128*
Last math grade	2.034	1.779	0.255*
Mother university degree	0.000	0.750	-0.750***
Father university degree	0.000	0.881	-0.881***
Mother employed	0.859	0.850	0.009
Father employed	0.877	0.869	0.008
First-generation migrant	0.108	0.188	-0.080
Second-generation migrant	0.246	0.175	0.071
Patience	3.456	3.598	-0.142
Risk aversion	3.038	3.022	0.017
Conscientiousness	3.669	3.900	-0.231**
Neuroticism	3.331	2.925	0.406***
Openness	3.523	3.441	0.082
Extraversion	3.385	3.391	-0.006
Agreeableness	3.262	3.197	0.065
Feedback aversion	2.174	2.075	0.099
Self efficacy	3.944	3.867	0.077
Motivation university studies	3.677	3.938	-0.261*
Weekly hours invested in course	8.877	5.436	3.441**
Observations	65	160	225

Notes: Balancing check between high- and low-SES individuals on all controls used in the analyses. Sample comprises of all participants who received feedback as a treatment in the second questionnaire of the experiment, without those who participated in the experiment for more than one course.

Table B.20: Balancing check for all control variables for course 1, by grade reporting

	Grade reported (1)	No grade reported (2)	(1) vs (2) (3)
Female	0.682	0.650	-0.032
Age	21.121	20.350	-0.771
Business Administration	0.576	0.650	0.074
Economics	0.000	0.000	0.000
Bus. and Econ. Education	0.076	0.050	-0.026
Minor Bus./Econ./Education	0.152	0.200	0.048
Other major	0.197	0.100	-0.097
Semester	2.788	3.100	0.312
German high school degree	0.955	0.850	-0.105
High school GPA	1.710	1.700	-0.010
Last math grade	1.819	2.156	0.337
Mother university degree	0.446	0.750	0.304**
Father university degree	0.561	0.650	0.089
Mother employed	0.862	1.000	0.138*
Father employed	0.909	0.850	-0.059
First-generation migrant	0.121	0.150	0.029
Second-generation migrant	0.106	0.300	0.194**
Patience	3.530	3.700	0.170
Risk aversion	3.311	2.900	-0.411
Conscientiousness	3.902	4.150	0.248
Neuroticism	3.212	3.275	0.063
Openness	3.500	3.575	0.075
Extraversion	3.303	3.575	0.272
Agreeableness	3.250	3.225	-0.025
Feedback aversion	2.278	1.900	-0.378*
Self efficacy	3.939	4.033	0.094
Motivation university studies	3.864	3.700	-0.164
Weekly hours invested in course	5.189	5.150	-0.039
Initial motivation	3.591	3.700	0.109
Points PQI	8.136	6.875	-1.261*
Observations	66	20	86

Notes: Balancing check between individuals who did or did not report their grades on all controls used in the analyses. Sample comprises of all participants who reported their grades and came from course I.

Table B.21: Balancing check for all control variables for course 2, by grade reporting

	Grade reported (1)	No grade reported (2)	(1) vs (2) (3)
Female	0.515	0.444	-0.070
Age	19.777	19.486	-0.291
Business Administration	0.476	0.694	0.219**
Economics	0.311	0.167	-0.144*
Bus. and Econ. Education	0.068	0.000	-0.068
Minor Bus./Econ./Education	0.087	0.111	0.024
Other major	0.058	0.028	-0.030
Semester	1.350	1.389	0.039
German high school degree	0.903	0.833	-0.070
High school GPA	1.704	1.903	0.199*
Last math grade	1.801	1.920	0.119
Mother university degree	0.524	0.611	0.087
Father university degree	0.650	0.667	0.016
Mother employed	0.806	0.889	0.083
Father employed	0.845	0.889	0.044
First-generation migrant	0.146	0.306	0.160**
Second-generation migrant	0.262	0.111	-0.151*
Patience	3.602	3.398	-0.204
Risk aversion	3.005	2.639	-0.366**
Conscientiousness	3.830	3.542	-0.288*
Neuroticism	2.966	2.819	-0.147
Openness	3.388	3.556	0.167
Extraversion	3.306	3.681	0.375*
Agreeableness	3.252	3.042	-0.211
Feedback aversion	2.068	2.000	-0.068
Self efficacy	3.819	3.917	0.098
Motivation university studies	3.864	3.944	0.080
Weekly hours invested in course	7.318	6.875	-0.443
Initial motivation	3.515	3.444	-0.070
Points PQI	9.505	6.958	-2.547***
Observations	103	36	139

Notes: Balancing check between individuals who did or did not report their grades on all controls used in the analyses. Sample comprises of all participants who reported their grades and came from course II.

Table B.22: Descriptive Statistics from the Post-Exam Questionnaire

	Control	POSNEG	NEGPOS	(2) vs (1)	(3) vs (1)	(2) vs (3)
	(1)	(2)	(3)	(4)	(5)	(6)
Difficulty Exam	2.262	2.062	2.172	-0.200	-0.089	-0.111
Difficulty PQ I	2.976	3.309	3.195	0.332*	0.219	0.113
Difficulty PQ II	3.643	3.975	3.885	0.332*	0.242	0.090
Usefulness PQ I	3.190	3.099	3.230	-0.092	0.039	-0.131
Usefulness PQ II	2.762	2.704	2.736	-0.058	-0.026	-0.032
Correct recall FB1 Yes/No	0.929	0.852	0.793	-0.077	-0.135*	0.059
Correct recall FB1 Elements		0.855	0.913	0.000	0.000	-0.058
Correct recall FB1 Ordering		0.881	0.937	0.000	0.000	-0.055
Usefulness Feedback PQ I	3.333	2.420	2.319	-0.913	-1.014	0.101
Feelings Feedback PQI	3.000	3.174	2.536	0.174	-0.464	0.638***
Correct recall FB2 Yes/No	0.929	0.889	0.862	-0.040	-0.067	0.027
Correct recall FB2 Points	1.000	0.944	0.933	-0.056	-0.067	0.011
Usefulness Feedback PQ II	2.897	2.556	2.587	-0.342	-0.311	-0.031
Observations	42	81	87	123	129	168

Notes: Descriptive statistics from the post-exam questionnaire (non-standardized), by treatment status.