Economics of Education Review 92 (2023) 102331

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

The effect of goal-setting prompts in a blended learning environment—evidence from a field experiment

Erwin Amann [a], Sylvi Rzepka [b,*]

[a] University Duisburg–Essen, Germany

[b] Center for Economic Policy Analysis (CEPA), University of Potsdam, RWI - Leibniz Institute for Economic Research, Germany

ARTICLE INFO

JEL classification:
I21
I23
C93
D91

Keywords:
Natural field experiment
Blended learning
Behavioral economics
Goalsetting

ABSTRACT

Previous literature has shown that task-based goal-setting and distributed learning is beneficial to university-level course performance. We investigate the effects of making these insights salient to students by sending out goal-setting prompts in a blended learning environment with bi-weekly quizzes. The randomized field experiment in a large mandatory economics course shows promising results: the treated students outperform the control group. They are 18.8% (0.20 SD) more likely to pass the exam and earn 6.7% (0.19 SD) more points on the exam. While we cannot causally disentangle the effects of goal-setting from the prompt sent, we observe that treated students use the online learning platform earlier in the semester and attempt more online exercises compared to the control group. The heterogeneity analysis suggests that higher treatment effects are associated with low performance at the beginning of the course.

1. Introduction

Higher education can be tough. The OECD average dropout rate amounted to 33% in 2017 (OECD, 2019). University drop-out is often driven by poor performance (Stinebrickner & Stinebrickner, 2014) and associated with behavioral bias such as time-inconsistency and self-control issues (Lavecchia et al., 2016). Online higher education appears even more demanding and especially challenging for some student populations, most prominently, male students and low achievers (Figlio et al., 2013). Yet, with the recent pandemic accelerating technological change in the educational sphere, some form of online education is likely to remain in higher education. It is important to make it work to the benefit of students who struggle with higher and online education.

In this paper, we test whether a blended learning environment can make the benefits of task-based goal setting and distributed learning salient to students and improve their course performance. For this, we send specific information on this learning approach as well as reminders to a randomly selected group of students and investigate how their study effort and study timing on the online learning platform differs from the control group.

We build on two strands of literature. First, the behavioral economics of education.[1] has seen a surge in experimental studies that have yielded mixed results. Studies on task-based goal-setting (Clark et al., 2020) and reminders (O'Connell & Lang, 2018) have proven to be effective in improving course-level performance, especially among male students likely to be more prone to low self-control. Yet, there are a few studies which tackle study time or test goal-setting in a broader setting that record null results. For instance, Dobronyi et al. (2019) and Oreopoulos et al. (2019) link goal-setting to academic and personal life in general. Himmler et al. (2019) send non-course specific reminders for staying on track, and Clark et al. (2020) and van Lent (2019) ask for grade-based goals for a course. Based on these results, we expect that prompting students to set task-based goals in a blended learning environment of a single course should improve student performance, especially for male students.

Second, we use insights from the psychological literature on effective study strategies and make these insights salient to students using the blended learning environment. This provides guidance to students at a scale that is not fathomable in traditional classroom settings. First, while "cramming" may be effective for academic performance, it has proven less effective for long-term retention of material than spaced learning (Carpenter et al., 2012; Kerdijk et al., 2015; Küpper-Tetzel, 2014). Second, self-testing is deemed a very effective learning strategy by psychologists, but students may need guidance to implement it (Tullis et al., 2013). In our blended learning environment, we combine prompts to task-based goals with bi-weekly quizzes. Hence, we

* Corresponding author.
E-mail addresses: erwin.amann@uni-due.de (E. Amann), sylvi.rzepka@gmail.com (S. Rzepka).

[1] For a general overview on the behavioral economics of education, see Damgaard and Nielsen (2018) and Lavecchia et al. (2016)

[2] For a review on the effectiveness of educational technology see Escueta et al. (2020).

https://doi.org/10.1016/j.econedurev.2022.102331
Received 10 February 2021; Received in revised form 19 June 2022; Accepted 27 October 2022
Available online 9 December 2022

use educational technology to communicate these insights to students in an action-driven way.[2] We expect the combination of prompts and bi-weekly quizzes to shift student learning over a longer period of time and funnel student efforts into active learning and testing practices, i.e. attempting more exercises on the online learning platform.

We ran a randomized controlled trial in the field, specifically a large introductory economics class with a diverse student body. For several years, the class had already multiple blended learning elements and incentives for starting to study early in the semester. Most importantly, students have access to JACK, a computer-assisted online learning platform that provides a range of online exercises, including parameterized methodological exercises. They assist students in overcoming deficits in mathematical skills, grasping, and applying the course material. Furthermore, bi-weekly extra-credit online quizzes provide an incentive for all students to start early with studying for the course. Yet, in past years students have tended to use JACK only sporadically during the semester and bunch their studying just before the quizzes and exam. Therefore, our experiment was intended to tackle potential behavioral biases, make students more aware of the benefits of the online platform, and ultimately improve their self-regulated learning (Tullis et al., 2013; Tullis & Maddox, 2012). For this purpose, we randomly assigned students to a treatment and control group after the first quiz. The treated students were encouraged to set a goal on how many online exercises they planned to complete in preparation for the next bi-weekly extra-credit quizzes. We analyzed the timing as well as the total amount of individual usage of JACK exercises for both the control and treatment groups, and compared it to the level of success in the final exam. At the end of the semester we elicited a range of demographic and personality indicators such as high-school level grade point average, parental education background, self-control (using the Tangney et al., 2004 scale), and patience. These all help us to decipher heterogeneous treatment effects.

The intervention increased the usage of JACK, the computer-assisted online learning platform, among the treated group, and improved their exam performance. We observe that treated students are 0.19 standard deviations more likely to take the early exam[3] than the control group, they earn 0.19 standard deviations more net points, and achieve better grades (0.16 standard deviations). The results for the intermediary outcomes, i.e. engagement with the online learning platform, corroborate the positive course-performance findings. The treated students complete 0.2 standard deviations more sessions and more unique exercises on JACK. In addition, the relatively large effect sizes suggest that such interventions may be particularly well-suited for diverse student groups.

The results offer several reasons why the intervention worked. First, there was some revealed take-up of goal-setting among the treated. 22% of students in the treatment group set a goal at least once on the interface provided to the treatment group. Second and more pervasively, students beyond the goal-setting group responded to the prompt itself with changes in study behavior. We observe that treated students engage with the online platform earlier in the semester and attempt more exercises. While we cannot disentangle the effects of goal-setting from the prompt sent, this suggests that the main effect may be coming from the prompt itself not the actual goal-setting.

Using causal forests, we reveal heterogeneity in treatment effect estimates (Wager & Athey, 2018). While nearly all treated students benefit from the intervention by earning more net points in the final exam, we show that larger positive effects are found for student populations who the literature has identified to be at risk of falling behind in online education, especially low achievers. The results suggest male students benefiting more from the treatment than female students; yet, the difference is not as pronounced as (Clark et al., 2020) observe in their study. Moreover, while indicators of self-control, patience, and prior achievement are significantly associated with effect heterogeneity, we find that measures capturing pre-determined online learning behavior explain more of the treatment effect heterogeneity. This suggests that early-course behavior can be used to target interventions.

The rest of the paper is structured as follows: Section 2 introduces the experimental design. Section 3 presents and discusses the main results, effects on study timing and usage of the online learning platform as well as goal-setting on the online platform. Section 4 quantifies effect heterogeneity using causal forests, and considers what factors determine this heterogeneity. Section 5 concludes.

2. Experimental design

2.1. Experimental set-up

The intervention is embedded in the "Introduction to Microeconomics" class at the University of Duisburg–Essen. This class is mandatory for study courses with majors or minors in Business and Economics, including teacher-training courses for vocational secondary schools. The curriculum advises students to take it early, i.e. in their second semester at university. For a number of years, the class has been taught by the same professor every other year. In addition to in-person lectures and exercise sessions, the course offers online material on its corresponding Moodle page, the online-learning management platform used by the University of Duisburg–Essen. The Moodle page provides the slides, instruction videos and online exercises facilitated by JACK, an automated online learning system developed by Paluno (University Duisburg–Essen).[4] The class is competence-oriented. To incentivize students to check their skills early on, biweekly quizzes are offered. All students can participate in seven online extra-credit quizzes which are spaced out evenly across the semester. With each quiz, students can earn 2 to 3 points per quiz that are added to their exam scores if they pass.[5] Despite the online exercises that JACK offers and the incentives provided by the quizzes, anecdotical evidence suggests the first-year students typically learn by heart and cram at the end of the semester. This suggests there are behavioral biases such as time-inconsistency and a lack of self-control which inhibit acquiring the skills taught. Furthermore, first-year students may be unaware of successful learning technique for university course work.

To overcome behavioral bias related to time-inconsistency and self-control, such as the cramming that was apparent in previous iterations, we use the blended learning environment to prompt treated students to set a task-based goal prior to each online quiz. Indirectly, the goal-setting prompt also communicates that using JACK is a successful learning technique for the course.

All students in the course who log in to the course Moodle page by the third course week participate in the experiment. Fig. 1 summarizes the timeline of the experiment. In the third week of the course, after the first online quiz, students are randomly assigned to the control or the treatment group.[6] Ten days prior to each online quiz, treated students receive a Moodle message encouraging them to set a goal for their preparation for the online quiz. The goal-setting prompt reads (the main treatment):[7]

[3] In the study program, students have to complete five to six exams per semester. For each exam there are two dates, one immediately after the course program ended and the second two months later. Typically, students choose the early exam if they feel well-prepared and confident in the specific course.

[4] For an overview of this platform visit: https://www.s3.uni-duisburg-essen.de/en/jack/ and Paluno, the developing institute https://paluno.uni-due.de/

[5] These bonus points amount to 4% of all attainable points in the course. They are only granted if the final exam is passed.

[6] Students who have not logged into Moodle by this time are excluded from the analysis, since they would potentially miss the first treatment. This means, they would select into treatment intensity.

[7] Figure A.1 in the appendix provides a screenshot of the original message in German.

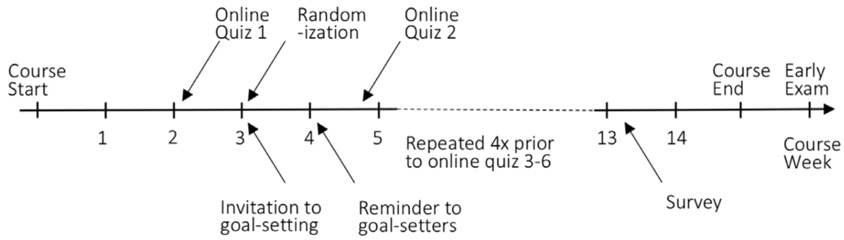

Fig. 1. Timeline of the Experiment.

Set a learning goal in preparation for the 3rd quiz

Dear Student,

Study advice: Did you know that setting concrete goals can increase learning success? With this in mind, set a goal for how many JACK exercises you would like to solve in preparation for the upcoming quiz. A total of 7 JACK exercises are available for the upcoming quiz.

It is all about your **personal learning goal**. No one, neither the instructors nor other students in the course, will see your goal. You will only be reminded of your learning goal with an automated Moodle message.

It is **my goal to complete** . . . prior to the 3rd quiz:
- . . . 1 JACK-exercise. . .
- . . . 2 JACK-exercises. . .
- . . . 3-4 JACK-exercises. . .
- . . . 5 or more JACK-exercises. . .
- I prefer not to indicate anything.
- I do not want to set a goal.

Fig. 2. Screenshot of Goal-Setting Interface prior to Quiz 3. *Notes:* For the original German screenshot of the goal-setting interface, please refer to Figure A.2 in the appendix.

"Dear Students, next week the X. quiz will take place. As always, setting specific goals can improve academic performance. Therefore, we encourage you to visit the Moodle-Website 'Learning goal in preparation for the X. quiz'. This is also possible, if you did not set a goal for the last quiz. The tool is now accessible.
This is an automated message from the Micro I Moodle course".

Fig. 2 depicts an example of how the goal-setting was implemented on a Moodle page. The categories for the goals (1, 2, 3–4, 5+ exercises) roughly represent the quartiles of the number of exercises available for each quiz. Therefore, for all but the last goal-setting interface, we could use the same categories to elicit the goals. Note that, similar to Clark et al. (2020), these goals define very concrete tasks. Further, they fulfill all requirements of Dotson (2016), who established that successful goals should be specific, measurable, attainable, relevant, and time-sensitive.

Only treated students reporting back individual goals are reminded via a Moodle message of the specific goal they set three days in advance of each online quiz ("goal reminder"). This procedure is repeated for all quizzes 2–6.[8] Goal achievement was not available publicly and this was communicated to the students in advance. At the end of the course, a survey elicits self-control, socio-economic characteristics, proxies for ability and study times as well as asking for an informed consent of all students to use their course-related data for research and course improvement.[9]

The messages relating to the experiment are in addition to all normal course communication (Fig. 3). Notably, the control group is informed about the bi-weekly online quizzes just as the treatment group. The number of additional messages (dark blue in Fig. 3) is only a small increase of the total messages in the course. Only those students who used the interface on Moodle to set their individual goals received an additional goal-reminder (light blue).

We choose this experimental design because it fits well with the blended learning course environment and is low-tech and low-cost. Yet, it does not allow us to disentangle the effect of goal-setting from the treatment messages. Goal-setting on the interface was voluntary, hence the sub-sample of students who set a goal there is likely to be selective. This is why, in the spirit of the "planning prompts" literature (see, for instance, Kizilcec et al., 2020; Yeomans & Reich, 2017), we refer to our intervention as a "goal-setting prompt". It is the prompt which we test empirically in the main body of this paper.

2.2. Descriptive statistics

Randomization was successful, as Table 1 shows. Nearly all predetermined characteristics are balanced across the treatment and control group. A joint F-test from a regression of the pre-determined characteristics on the treatment status confirms this (F-statistic: 0.7565 and corresponding p-value: 0.8312). Furthermore, the descriptive statistics are in line with expectations for the study population. Nearly half of students are male, most are younger than 22 years old and within the first four semesters of their university studies. Almost all completed

[8] For quiz 6 no reminder of set goals was sent out.

[9] 557 students participated in the survey. Survey participation was not statistically significantly different across treatment and control group (48.4% versus 50.8%; p-value of t-test on equality 0.448). Among those who participated in the survey, consent rates were 89.3% for the treated and 89.9% for the control group. The p-value of a t-test on equality amounted to 0.801. Hence, the consent rates do not differ statistically significantly across treatment and control groups.

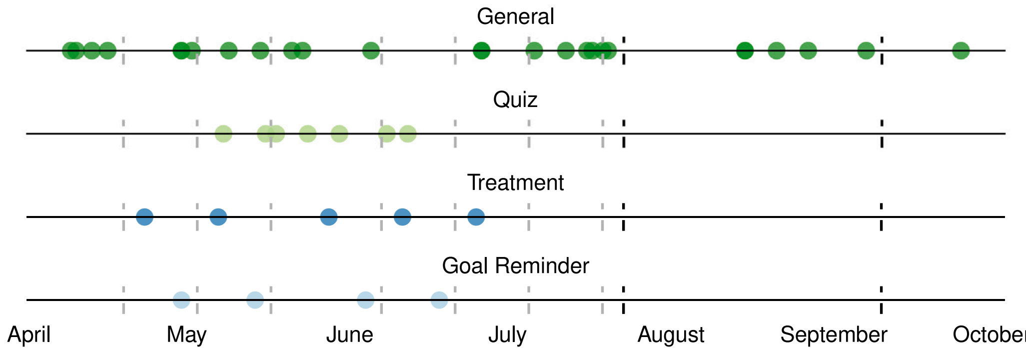

Fig. 3. Timing of course communication by treatment status.
Notes: The figure displays the timing of all course communication. General and quiz-related communication were sent out to all students. The treatment-related messages only targeted the treatment group, and the goal-reminder was only sent to treated students who had set a goal. Gray lines indicate the time quizzes and black lines when exams took place.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Descriptive statistics and balancing tests.

Covariate	Mean control	Mean treated	Difference	*P*-value
Study course				
Business	0.457	0.419	0.038	0.390
Economics	0.194	0.162	0.032	0.351
Math/Business math	0.093	0.137	−0.044	0.126
Business informatics	0.140	0.183	−0.043	0.193
Teaching degree	0.105	0.087	0.018	0.507
Other degree or missing	0.012	0.012	−0.001	0.933
Timing in study				
Semester 1–2	0.430	0.386	0.044	0.315
Semester 3–4	0.450	0.402	0.047	0.288
Semester 5+	0.116	0.195	−0.079	0.016
Semester missing	0.004	0.017	−0.013	0.164
Abitur GPA				
Abi grade very good or good+	0.143	0.149	−0.006	0.851
Abi grade good−	0.209	0.183	0.027	0.453
Abi grade satisfactory+	0.403	0.432	−0.028	0.521
Abi grade satisfactory−	0.136	0.124	0.011	0.711
Abi grade pass	0.054	0.041	0.013	0.504
Other School leaving exam or missing	0.054	0.071	−0.016	0.455
Bundesland of Abitur				
Abi in NRW	0.919	0.892	0.026	0.314
Abi not in NRW or missing	0.004	0.012	−0.009	0.293
Age				
Less than 19 years old	0.143	0.124	0.019	0.536
20 to 21 years old	0.422	0.386	0.037	0.406
22 to 24 years old	0.326	0.336	−0.011	0.803
More than 25 years old	0.105	0.133	−0.028	0.334
Age missing	0.004	0.021	−0.017	0.092

(*continued on next page*)

their Abitur, the German secondary education certificate qualifying them for university, at a school in North-Rhine Westphalia (NRW), the same Bundesland as the University. About a third of the sample comes from an academic parental household.[10] A third of the students receive BAfoeG, i.e. means-tested financial state aid. These shares differ from the German national average, for which 54% of students come from academic households and only 11.5% of students receive BAfoeG (Middendorff et al., 2017), but are in line with the generally more diverse student body of the University of Duisburg–Essen, where 39.1% come from academic households, and 30.3% receive BAfoeG (Ganseuer et al., 2016). The distribution of Abitur grade point averages (gpa) also differ from the average in NRW of 2016/17. In our sample, 14% of students attained a "very good or good +" grade point average, in NRW 28% (Kultusministerkonferenz, 2017). While our sample records a higher share of students with a "good" gpa (21%) than the NRW average (14%), our study sample also has a lower share of students with "satisfactory -" gpa (13.6%) than the NRW average (21%). All in all, this suggests that students in our sample tend to come from more disadvantaged backgrounds and be from the middle of the regional prior-achievement distribution.

The treatment begins in the third week of the semester. Hence, if randomization was successful, all online learning activities until Week 3 should be the same across the treatment and control groups. This is indeed the case, as there are no economically or statistically significant differences in pre-treatment intermediary outcome variables. Students in the treatment and control group used JACK to the same intensity, participated in the first quiz in equal shares, and earned roughly the same number of bonus points.

2.3. *Empirical strategy*

The randomization enables identifying the causal impact of the goal-setting prompt with a simple ordinary least squares regression. Hence,

[10] "Academic" indicates that at least one parent has either a university degree or a degree from a university of applied science.

Table 1 (*continued*).

Covariate	Mean control	Mean treated	Difference	*P*-value
Gender				
Male	0.477	0.481	−0.005	0.919
Gender missing	0.012	0.008	0.003	0.708
Financial aid status				
BAfoeG recipient	0.310	0.290	0.020	0.633
BAfoeG missing	0.039	0.037	0.001	0.934
Parental background				
Academic background	0.322	0.353	−0.031	0.465
Background missing	0.205	0.170	0.035	0.313
Personality traits				
Low patience	0.388	0.361	0.027	0.540
Low self-control	0.341	0.390	−0.049	0.258
Early course behavior (before treatment)				
Participation in test 1	0.934	0.900	0.034	0.174
Score on test 1	1.233	1.261	−0.029	0.650
Share of non-JACK users	0.320	0.328	−0.008	0.855
Number of JACK-exercises attempted	6.450	6.485	−0.036	0.962
Number of unique JACK-exercises attempted	4.891	4.859	0.033	0.948

Notes: The table presents means, differences, and resulting *t*-statistics for the pre-determined characteristics as well as early course behavior, i.e. prior to the first online quiz. "Academic background" indicates students with at least one parent having some form of university degree. Patience is elicited using the 11-Likert-Point SOEP patience question. Self-Control measures the index collected through the 14-item Self-Control Scale (Tangney et al., 2004). Here the mean of binary versions, lower than the lowest tercile, are presented for these personality traits.

for the main analysis, we estimate the effect of encouraging students to set a goal prior to each quiz as follows:

$$Y_i = \alpha + \beta Z_i + \epsilon_i \tag{1}$$

Y_i stands for the main outcomes, exam participation and performance (grade and points net of bonus points), and intermediary outcomes, number of JACK exercises attempted, number of sessions, and total time spent on the online learning platform. Z_i indicates the treatment group status of student i. β captures the causal effect of the random assignment to the goal-setting prompt on the student's outcome. We test whether a pre-specified set of covariates X_i (high school grade point average, gender, socio-economic status, study course, and indicator variables for the point in time when students log in to Moodle for the first time during the course, patience and self-control) renders the estimation more precise.[11]

In our pre-analysis, we conducted a power analysis based on the following assumptions: a sample size of about 900 students, descriptive statistics of JACK-use and average exam points from the previous iteration of the course (in 2016), a significance level of 5%, and statistical power of 80%[12] Ex-ante, this suggests a minimum detectable effect size of a 12% increase in JACK-sessions, a 17% increase in JACK-exercises, and a 6% increase in exam points.

3. Results

3.1. Results on main course outcomes

Encouraging students to set task-based goals positively impacts pass rates, points earned in the exam net of bonus points, and grades (including bonus points). As we expected following the results from previous literature on task-based goals, the intervention does not affect overall exam participation (Table 2).[13] This is not surprising, given that the course is mandatory for most students enrolled. The treatment does increase the likelihood of taking the early exam just after the summer semester (as opposed to later, before the winter semester) by 6.4 percentage points (7% or 0.19 SD more than the control group). Overall, pass rates increase by 10 percentage points. This corresponds to an effect size of 18.9%, relative to the control group (0.20 SD more than the control group). The treatment also improves grades by 0.26 (7.5% and 0.16 SD compared to the control group).[14] Net of bonus points, the treated students earn about 2 points more in the exam than the control group.[15] This corresponds to a relative increase of 6.7% (0.19 SD) compared to the control group. Effect sizes for grades (including bonus points) and net points are nearly the same. This means treated students earn as many bonus points as control group students;[16] but they earn more net points in the exams. Students were able to cooperate on the quizzes, which might be a reason why we do not observe treatment effects for bonus points earned. Yet, the exam outcomes excluding bonus points, are unaffected by potential cooperation. As Fig. 4 shows, the treatment shifts the middle of the distribution of net points to the right and reduces the variance. This suggests that the treatment leads students to have a better understanding of the material.

Our effect sizes are on the upper bound of those found in similar contexts. A well-powered study rooted in the growth mindset literature, which asked students to set goals with open-ended questions and which targeted study time, had a zero effect on performance (Dobronyi et al., 2019). Performance-based goals have regularly shown not to affect average exam performance in Europe and the USA (Clark et al., 2020; van Lent, 2019). Yet, Clark et al. (2020), who's task-based goal

[11] In the pre-analysis plan, we intended to estimate the average treatment effect on the treated (ATT), using an instrumental variable strategy. We outline this analysis in the appendix (A.1). However, we do not include it in the main results section, because only 53 students set a goal once, making the ATT estimates very local and further, the exclusion restriction may not hold. We elaborate on this in A.1.

[12] The descriptive statistics used were: average number of JACK-exercises attempted 3.35 (sd: 3.48), total number of JACK-sessions 9.72 (sd: 6.50), and average number of points achieved in the exam 33.22 (sd: 11.20).

[13] Precision and coefficients barely change when we control for the pre-specified set of covariates, see Table A.1. Therefore, to facilitate interpretation we report the regressions without controls in the main text.

[14] In the German academic grading scheme a lower value indicates a better grade, as the schemes ranges from "1.0 – very good" to "4.0 pass". A "5.0" indicates failing an exam. However, to facilitate interpretation we re-scaled the variable such that 4.0 is the best grade and 1.0 the passing grade. This allows to interpret a positive point estimate as having a positive impact.

[15] The exam had a total of 60 possible points net of bonus points.

[16] Anecdotal evidence suggests that there is cooperation among students during the semester, probably across intervention groups. This may induce positive externalities of the goal-setting prompt even for the control group, reducing the observed differences between the groups in our experiment.

Table 2
Main course outcomes.

	Exam (1)	Early exam (2)	Pass (3)	Grade (4)	Points (net) (5)
Treated	0.006	0.064**	0.102**	0.255*	1.956**
	(0.025)	(0.028)	(0.046)	(0.143)	(0.915)
Constant (Control Group Mean)	0.911***	0.864***	0.540***	1.598***	29.370***
	(0.018)	(0.022)	(0.033)	(0.102)	(0.679)
Standard deviation of control group	0.29	0.34	0.50	1.57	10.41
Observations	499	456	456	456	456
Adjusted R^2	−0.002	0.009	0.009	0.005	0.008

Note: *p<0.1; **p<0.05; ***p<0.01.
The table presents the coefficients of simple linear regressions of the treatment assignment on the outcome variables. This means the constant can be interpreted as the mean outcome in the control group. Column 1–3 are binary variables indicating exam participation, participation in the first exam round and passing the exam. Columns 4 and 5 are continuous variables indicating the grade and net points. Robust standard errors are reported in parentheses. See Table A.1 in the appendix for the regression results with covariates and Table A.2 for a robustness check using the sample of the heterogeneity analysis.

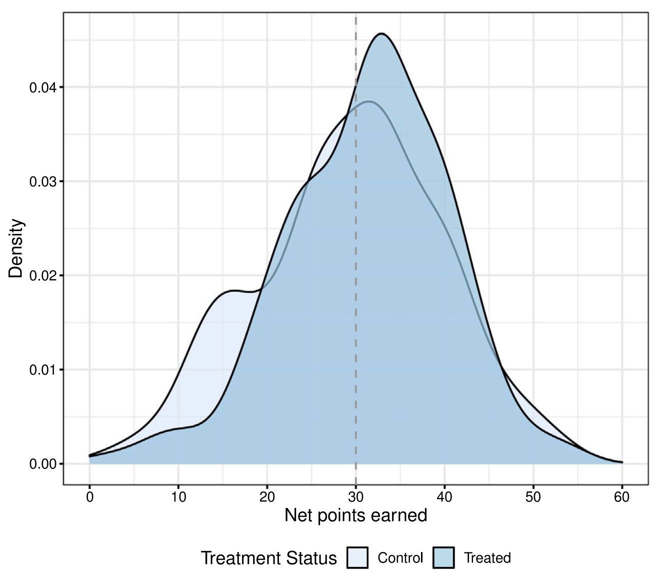

Fig. 4. Distribution of Net Points by Treatment Status.
Notes: Distribution of net points by treatment status. The dashed line indicates 30 points, the total points needed to pass the exam.

experiment is closest to ours, reach similar effect sizes. They report an 0.07 SD increase in average exam points for the treatment group, and a 0.10 SD increase in the average number of practice exams—their intermediary outcome. They find that their male students reacted more strongly to the treatment, reaching 0.16 SD more points on the exam and completing 0.19 SD more practice exams than the control group.

While differences in effect sizes may be purely random, some contextual aspects and design specifics may also play a role (Kizilcec et al., 2020). In contrast to Clark et al. (2020), we measure the number of exercises and sessions for our intermediary outcomes. These are more fine grain measures of exerted effort and may therefore pick up behavioral changes quicker. Further, the populations are different not only with respect to the country, but also with respect to the part of the distribution from which students are drawn. While in Clark et al. (2020), the student body comes from a “top-ranked public university”, the socio-demographic characteristics (Table 1) suggest that our sample stems from more diverse backgrounds and has a lower share of top-achieving students compared to the regional Abitur grade point average. Hence, the scope for improvement may have been larger for our study’s sample than that of Clark et al. (2020).

3.2. Results on study timing and study effort on the online learning platform

The treatment tends to induce more usage of the online learning platform JACK, both at the extensive and the intensive margin. Fig. 5 shows that JACK usage was the same prior to the treatment. But with each invitation to set a goal in preparation for the next quiz, the share of JACK users[17] and the average cumulative number of exercises attempted rises more among the treated than the control group. By the end of the semester, the share of JACK users among the control group is as high as that in the treatment group. Thus, the JACK usage gap at the extensive margin is closed. Yet, the gap in usage intensity is never closed; throughout the semester, treated students submit more exercises on average than the control group. Further, the gap at the intensive margin seems to widen as the exam approaches: for exam preparation, the treated students tend to turn to JACK earlier and more intensively than the control group.

The regression results for the intermediary outcomes quantify these positive effects (Table 3). Treated students complete 9.8 more exercises and attempt 2.9 more unique exercises. This corresponds to effect sizes of 12.2% (0.18 SD) and 7.9% (0.19 SD) compared to the control group. Treated students log in for about 5.5 more sessions (13.1% and 0.20 SD) than the control group.[18] Adding the duration of all sessions together, treated students spend about 31 minutes longer on the online learning platform than the control group, who spend 4.1 h. This is a relative increase of 12.2% (0.17 SD) compared to the control group. Yet, this point estimate is insignificant. In short, the prompt affects effort and timing of learning on the online platform. As we expected, the prompt seems to make effective study strategies so salient that students indeed distribute their study effort, i.e. start studying earlier in the semester, and consult the exercises on the online learning platform more to better grasp the material. We think this is the main reason for the better course outcomes.

One concern may be that the intervention led to an increase in time investment of the treated students in the microeconomics course at the expense of other courses. Using approximate measures of mean time invested, in order to conduct t-tests between the treatment and control group confirm this finding. The survey elicited time input as a categorical value.[19] For the t-tests, we assigned the midpoint of the

[17] We define “JACK users” as students who submit at least one exercise on the learning platform.

[18] “Sessions” are approximated using the time stamp recorded for each exercise submission. One session is a collection of timestamps with pauses shorter than 15 minutes.

[19] The distribution of these categorical values and a more detailed discussion is in the appendix A.2.

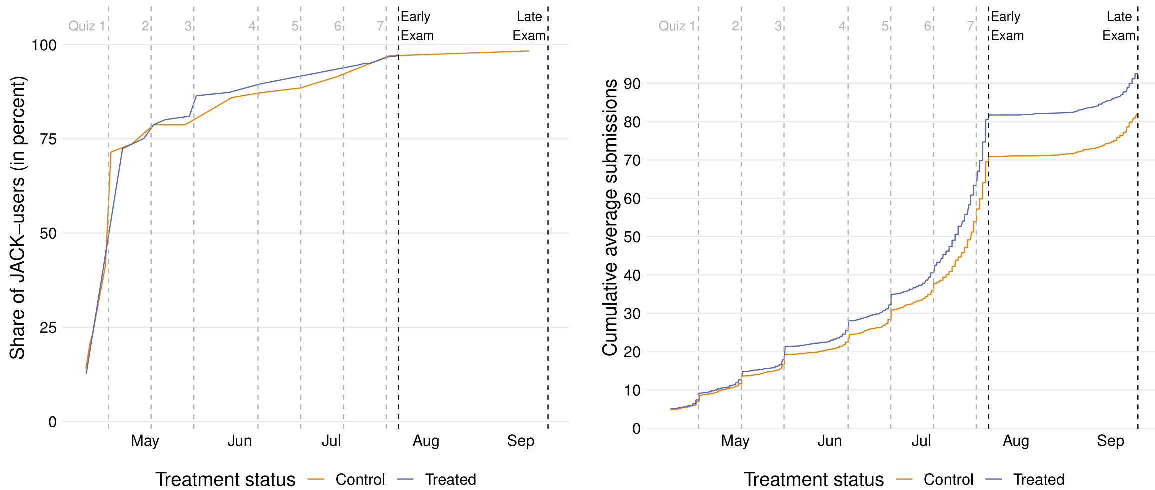

Fig. 5. Timing of JACK-Usage by Treatment Status.
Notes: The figures show the timing when the control and treatment groups make use of the JACK online learning platform. The left figure shows the share of JACK-users by treatment status (the extensive margin). The right figure shows the intensity of JACK-usage, i.e. the average submissions by treatment status (intensive margin). The dashed gray lines mark the points in time of the extra-credit online quizzes where students could earn bonus points. The dashed black lines indicate the early exam (in July) and the late exam (in September).

Table 3
Intermediary outcomes.

	Exercises (1)	Exercises (unique) (2)	Number of sessions (3)	Time on platform (4)
Treated	9.819** (5.004)	2.859** (1.245)	5.537** (2.578)	30.095* (16.804)
Constant (Control Group Mean)	80.574*** (3.441)	36.136*** (0.899)	42.403*** (1.807)	246.721*** (11.646)
Standard Deviations of control group	54.32	15.42	27.78	179.47
Observations	456	456	447	447
Adjusted R^2	0.006	0.009	0.008	0.005

Notes: *p<0.1; **p<0.05; ***p<0.01.
This table presents results of simple linear regressions of the treatment on intermediary outcome variables for students who participated in the exam. This means the constant can be interpreted as the mean outcome in the control group. In column 1, the outcome is the number of JACK-exercises attempted and in column 2, it is the number of unique JACK-exercises attempted. Column 3 captures the number of sessions. Sessions are constructed using consecutive time stamps that have interruptions which are no longer than 15 min. In column 4, "time on platform" is measured by summing all sessions per individual. Note the number of observations in Column 3 and 4 are lower because nine students only have one time stamp. Robust standard errors reported in parentheses. See Table A.3 in the appendix for the regression results with covariates, and Table A.5 in the appendix for the sample used in the heterogeneity analysis.

categorical time variable which the student indicated as the numeric value for time investment. This transformation allows us to determine a proxy for average time input for the control and the treatment group. The p-values from this t-test are: (1) 0.76, (2) 0.93, (3) 0.56, (4) 0.21. In short, time investment patterns are all very similar. This means that the treatment promoted more productive study activities, rather than increasing the overall time input.

3.3. Goal-setting on the online platform

We refer to our intervention as a "goal-setting prompt"[20] and cannot disentangle actual goal-setting from the messages sent to students. However, we can provide descriptive evidence on how the goal-setting interface was used and this usage may contribute to the main effects found. In the treatment group 53 students (22%) interacted with the online goal-setting page on Moodle at least once. Table 4 shows the distribution of all goals set by these students. Most students opted for the category "3-4 JACK exercises" as their goal to prepare for the online quiz. This corresponds to about half of the JACK exercises available for the respective quiz study material.

Table 4
All goals disclosed on the online learning platform.

Goal	Numbers of observations
... 1–2 JACK exercises.	13
... 3–4 JACK exercises	42
... 5 or more JACK exercises	27
I do not want to set a goal	3
I do not want to make a statement	4

Notes: Prior to quiz 2 to 5 the online goal-setting page provided the following categories: 1, 2, 3–4, 5 or more JACK exercises, "I do not want to set a goal", and "I do not want to make a statement". For quiz 6 there were only 3 exercises available on JACK to complete in preparation, so the numeric categories were simplified to 1, 2, and 3 JACK exercises.

[20] We choose this wording following the spirit of the "planning prompts" literature (Kizilcec et al., 2020; Yeomans & Reich, 2017).

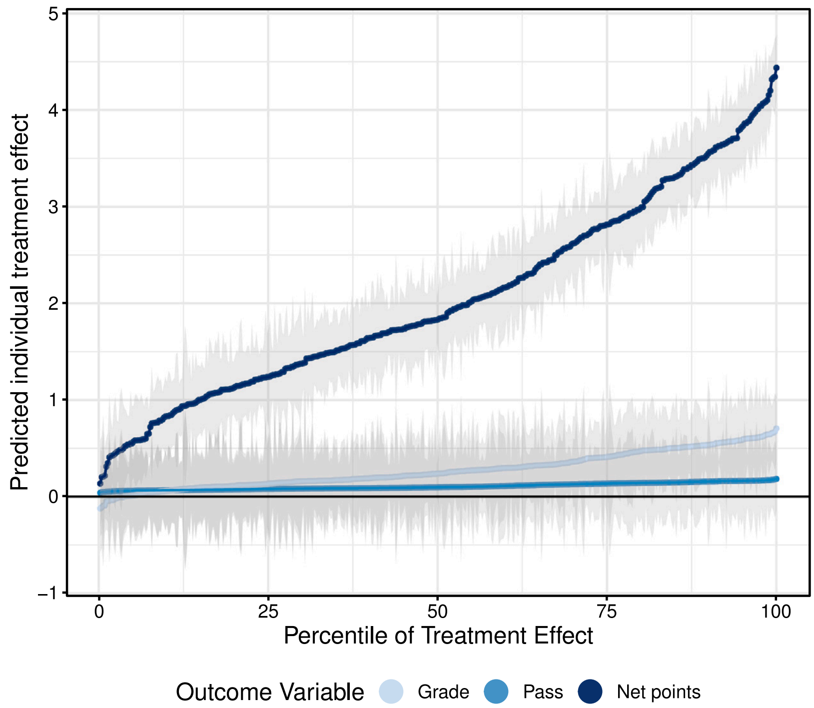

Fig. 6. Distribution of Treatment Effects.
Notes: The figure shows the distribution and the 95%-confidence interval of treatment effects for the tree outcomes: net points, grade, and passing the exam. The individual treatment effects are sorted in ascending order for each outcome separately.

While the actual goal-setting of the small self-selected group impacts the main results, the effects are not isolated to the students who disclose goals on the platform. In the appendix, Table A.4 reproduces the main results of Table 2 additionally controlling for an indicator variable for those 53 students who disclose a goal at least once. The estimates for the treated, who did not set a goal on the platform, point into the same direction but are slightly smaller than for the full treatment group. This may be due to a lack of statistical power. The coefficients for "set goal once" suggest that in addition to the overall treatment effect, this self-selected group experiences a boost in their course outcomes. From this analysis we conclude, that actual goal-setting and an additional reminder of the goal set contribute to the positive outcomes but may not be the only factor.

4. Treatment effect heterogeneity

4.1. Estimating treatment effect heterogeneity

We use causal random forests to estimate the distribution of treatment effects and their heterogeneity (Wager & Athey, 2018). With this method we can estimate $\hat{\tau}(X_i)$ the individual causal effect by averaging over the difference in outcomes $\bar{Y}_1 - \bar{Y}_0$ in the terminal nodes over many trees. The trees are constructed via recursive partitioning, where splits are set such that the variance between leaves is maximized. This also maximizes effect heterogeneity (Athey & Imbens, 2016). Forests decorrelate trees by introducing two sources of randomness; a random subsample is used to construct each tree and only a random subset of covariates enter the algorithm. Further, causal forests hinge on the honesty principle. For this purpose, the training data is split into two parts, one to build the forest and a second to populate it. This means the estimation of the treatment effects is independent of model building. This honesty principle is at the core of the assumptions for deriving conditions for consistency and asymptotic normality of causal forests which allow estimating confidence intervals (Wager & Athey, 2018).

Using the *grf* R-package, we build causal random forests for three outcome variables, net points, grade, passing the exam, using 9000 trees and otherwise tuning the parameters in a data-driven manner. A total of 43 pre-treatment variables enter the causal forest. Of these, 15 variables are JACK-related and 36 are demographic variables. Figure A.3 shows an example of a tree that entered the causal forest. Table A.6 lists the 25 most important variables, i.e. those that are most predictive of treatment effect differences. Despite many more demographic variables entering into the algorithm, 14 of the top 25 important variables are JACK-related.

Fig. 6 shows the distribution of treatment effects for the tree outcomes: net points, grade, passing the exam. For net points earned in the exam, the treatment effects range between 0.14 and 4.44, for grades between −0.12 and 0.71, and for passing the exam, between 4.02% and 18.34%. Using the plotted 95%-confidence intervals as guidance, we do not observe statistically significant effect heterogeneity for passing or grades earned. This is not surprising, given the narrow span of these variables. For net points earned as an outcome variable, however, the confidence intervals of the bottom and top terciles of the distribution do not overlap. Hence, for net points, we can visually reject the null hypothesis of no effect heterogeneity.

4.2. What determines effect heterogeneity?

The literature often identifies time-inconsistency coupled with a low level of self-control, gender, and prior ability as driving factors for treatment effect heterogeneity in the context of both online education and behavioral interventions such as goal-setting on student outcomes (Clark et al., 2020; Figlio et al., 2013). We use survey proxies for these concepts in order to understand to what extend they are associated with the estimated effect heterogeneity.[21] We also study

[21] Our survey, which among other things, elicited personality traits took place after the intervention. However, these kinds of traits remain rather constant for at least a four-year period (Cobb-Clark & Schurer, 2012, 2013). For our specific question on patience, recent research suggests that even large unforeseen events as COVID-19 tend to have limited effects (Frondel et al.,

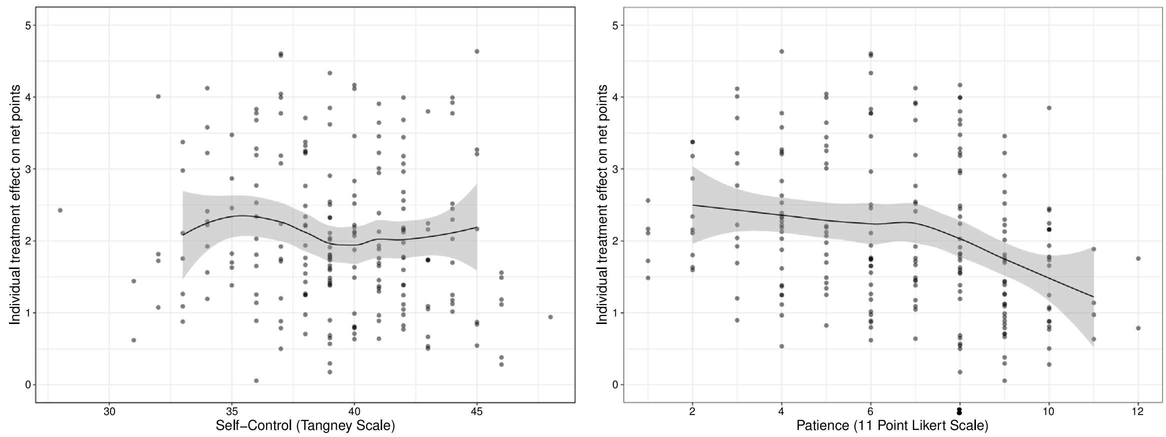

Fig. 7. Predicted Treatment Effects by Self-Control and Patience Levels.
Notes: The figures show the predicted individual treatment effects on net points for the treated subsample (N = 216) on the *y*-axis, and the survey indicators of self-control (Tangney Scale) and patience (11-point Likert scale) on the *x*-axis. The figure also includes a local regression (loess), which for Self-Control only considers observations above 33 and below 46.

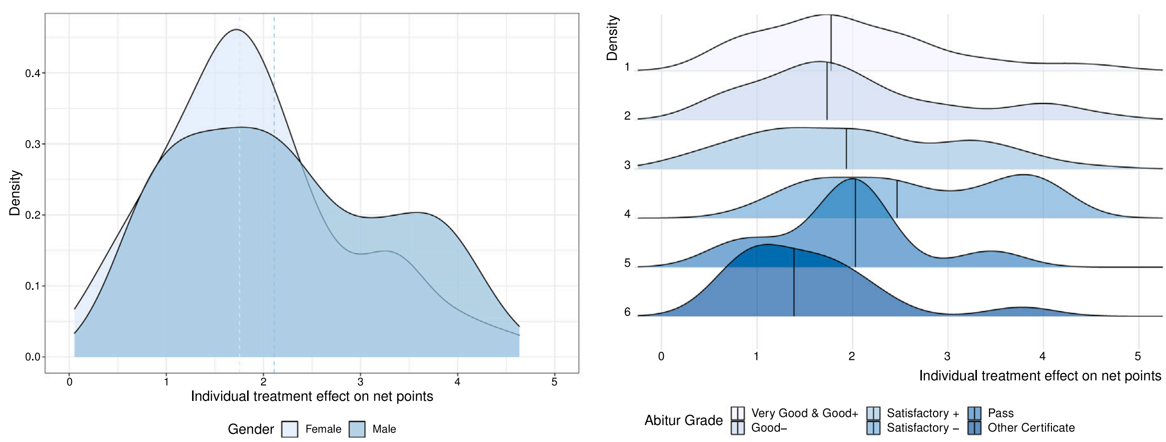

Fig. 8. Predicted Treatment Effects Distribution by Gender and Prior Achievement levels.
Notes: The figures show the distribution of predicted individual treatment effects on net points for the treated subsample (N = 216) disaggregated by gender and different grades levels of the Abitur, the German university entrance certificate. "Very good and good+" refers to grades 0.7 to 2.0, "Good-" to 2.1–2.3, "Satisfactory +" to 2.4 to 3.0, "Satisfactory -" to 3.1 to 3.3, "Pass" to 3.4–4.0. "Other Certificate" captures all those who enter with different qualifications, e.g. with foreign university entry certificates. The vertical lines indicate medians.

early course behavior and how it is associated with effect heterogeneity. While personality traits and other student characteristics might not always be available, in blended learning settings, practitioners have the advantage of observing early course behavior.

The bivariate plot of the estimated treatment effects on net points for self-control does not suggest a clear association (Fig. 7). For patience - a proxy for time-inconsistency (Vischer et al., 2013) - we do observe a slight negative association, the direction one would expect theoretically. Thus, the predicted treatment effects tend to be higher the more impatient students are. Fig. 8 suggests that predicted treatment effects follow a more bipolar distribution for male students than for female ones. This means that there is a group of male students for whom the treatment is predicted to have a larger effect on net points than for female students. The distribution of treatment effects of students with "satisfactory" Abitur grades, i.e. a bit below average, exhibit more mass for higher treatment effects than other achievement levels (Fig. 8).

In a multiple regression of these variables on the predicted treatment effects (see Table 5), only the coefficients on patience and prior ability reach statistically significant levels. However, the adjusted R-squared only amounts to 0.125. Accordingly, these and other demographic variables such as age and field of study do not explain much of the variation in treatment effect heterogeneity.

We now turn to the variables capturing early behavior in the course, to see how they are associated with the treatment effect heterogeneity. Students who did not use the online learning platform prior to the first test have higher predicted treatment effects (Fig. 9). Further, Fig. 10 shows that higher predicted treatment effects on net points are concentrated at zero or low numbers of unique exercises attempted. This means that the treatment had substantial effects on the extensive margin of JACK-use during the semester. While the share of JACK users is the same in the treated and control groups at the beginning and at the end of the semester, this is not the case at all times during the semester. As we saw in Fig. 5, during the semester, the JACK user share is higher among the treatment group than in the control group. Yet, beyond the large extensive-margin effect, the predicted treatment effect on net points stays well above 1. This suggests that there is a positive effect on the intensive margin, no matter how intensive the JACK usage was prior to the treatment.

The predicted treatment effect distribution of students who earn one or no bonus point on the first test (out of two maximum) tend to bunch at higher individual treatment effects. This suggests that the

2021). Therefore, given our six-month time frame, we regard these personality traits as rather constant.

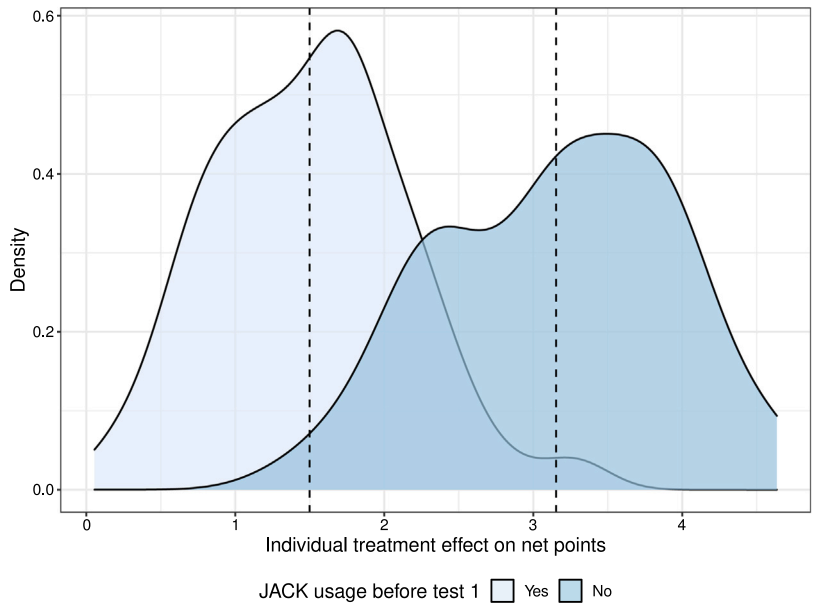

Fig. 9. Predicted Treatment Effects and JACK Usage Before First Test.
Notes: The figures show the predicted individual treatment effects on net points for the treated subsample (N = 216) disaggregated by JACK usage before the first test. The number of observations are as follows: 143 treated students used JACK prior to the first test, 73 treated students did not. The dashed lines indicate the average treatment effect for each group (1.46 for those who used JACK before the first test and 3.18 for those who did not.

intervention was more effective for students who performed below average early in the course.[22]

We also investigate, in a multiple regression framework, how early course behavior is associated with the predicted treatment effects on net points (Table 5, Column 2).[23] The main conclusions discussed above hold. While not all of the course behavior variables reach statistically significant levels, the adjusted R-squared skyrockets. This means that the pre-treatment course behavior explain much of the predicted treatment variation. When we include both the personality, prior achievement, and demographic variables as well as the early course behavior variables in one regression, self-control, patience, and prior achievement levels are statistically significant.[24] This provides suggestive evidence that both behavioral biases and lower prior achievement levels are challenges which our prompting intervention helps to overcome. However, the variation in early course behavior explains much more of the predicted treatment effect heterogeneity.

5. Conclusion

Based on a randomized natural field experiment in a blended learning environment, we showed that sending out prompts to students to set task-based goals regularly positively affects engagement with the computer-assisted online learning platform and course performance. Students in the treatment group submit 9.8 (0.18 SD of the control group) more exercises. They attempt three more unique exercises

Table 5
Regression on predicted treatment effects on net points.

	Predicted treatment effects		
	(1)	(2)	(3)
Self-control	−0.017		−0.022***
	(0.012)		(0.006)
Patience	−0.072***		−0.060***
	(0.016)		(0.008)
Male	0.118		−0.093**
	(0.088)		(0.039)
“Very good” &	−0.030		0.046
“good+” Abitur	(0.123)		(0.060)
“Good−” Abitur	−0.050		0.017
	(0.122)		(0.055)
“Satisfactory−” Abitur	0.395***		0.247***
	(0.137)		(0.058)
“Pass” Abitur	−0.105		0.041
	(0.149)		(0.099)
Other Certificate	−0.106		0.074
	(0.164)		(0.072)
No JACK activity		1.194***	1.159***
		(0.111)	(0.098)
Unique exercises		−0.018	−0.032***
		(0.012)	(0.010)
Mean activity		−0.018	−0.019**
		(0.011)	(0.009)
Test 1 participation		0.300***	0.220***
		(0.096)	(0.083)
Test 1 score		−0.111***	−0.117***
		(0.038)	(0.032)
Observations	447	447	447
Adjusted R^2	0.125	0.725	0.827

Note: Robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01. Column 1 also controls for: age, academic parental household, other university entry certificate, field of study, semester of study. Column 2 also controls for (all prior the first test): count of correctly answered questions, median time of day logged in and the standard deviation of the median time of day, time difference between the first login to Moodle and the first test, time difference between last login to the online learning platform and the first test, number of sessions on the online learning platform, average number of submissions during a session and its standard deviation. Column 3 controls for the union of Column 1 and 2 variables.

[22] 1.25 points was the average earned in the first test.

[23] We include 12 of the 14 early behavioral variables that are among the 25 “important” variables in the causal random forest (see Table A.6 in the Appendix). To avoid multicollinearity in the multiple regression, we exclude duration and its standard deviation, as well as the overall score, since these variables are by construction linear combinations of other variables such as sessions and correctly answered questions.

[24] The coefficient for the male dummy declines in size and changes signs. This is due to holding all other factors constant—especially whether students are active on JACK prior to the first test. More male students (38.6%) are actually not active before the first test than female students (26.8%).

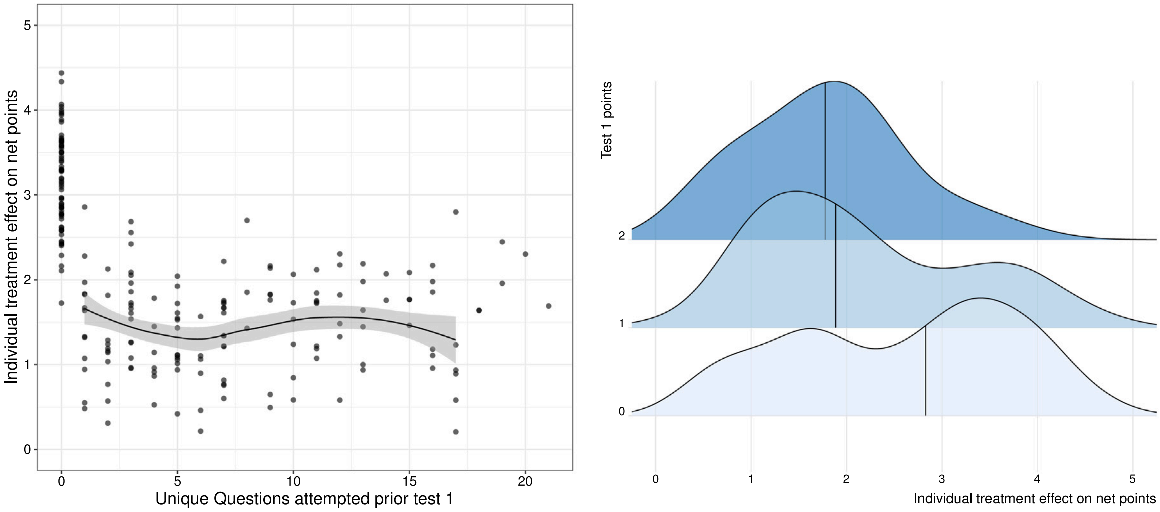

Fig. 10. Predicted Treatment Effects and Number of Unique Exercises Attempted and Score on First Test.
Notes: The figures show the predicted individual treatment effects on net points for the treated subsample (N = 216) for the number of unique questions answered prior to the first test. The figure on the left also includes a local regression (loess). We adjust the interval of the loess regression to range 1 to 17 unique questions answered prior to test 1 to ensure it is not driven by the bunching at zero or the few outlier observations which showed intensive pre-treatment JACK usage. The figure on the right shows the distribution of predicted treatment effects and the points students earned on the first test and the median predicted treatment effect. The number of observations are as follows: 0-points group N = 34 (among those 20 did not participate in the test), 1-point group N = 87, 2-points group N=95.

(0.19 SD of the control group), and log in for 5.5 (0.20 SD) more sessions than the control group. The treatment group does not spend significantly more study time on the course overall. However, as we expected following the previous literature, the intervention induces treated students to start studying earlier, hence distributing studying more evenly over the semester, and to attempt more exercises on the online learning platform. These positive effects on study timing and engagement with the online learning platform translated into higher early exam participation (0.19 SD more than the control group), higher passing rates (0.20 SD), and more net points earned (0.19 SD). In short, treated students outperform the control group in the course.

These effect sizes are on the upper bound of what similar interventions have shown. We argue that this may be because the student body is more demographically diverse and drawn from the middle of the high-school achievement distribution, hence missing the top students. Therefore, the scope for improvement may have been larger in our context than in other contexts. As also pointed out by Kizilcec et al. (2020), in what way the effects of prompts depend on the specific context is an interesting avenue for future research.

While we cannot disentangle the effects of goal-setting from the prompt sent, the results suggests that the main effect may be coming from the prompt itself not as much from actual goal-setting. About a fifth of students in the treatment group set a goal at least once on the online learning platform. Yet, beyond the goal-setting we find the prompt itself led students to engage with the online platform earlier in the semester and attempt more exercises.

At a more general level, we see two main takeaways. First, our results suggest how prompts embedded in a blended learning settings can help mitigate poor course performance by design. In our setting, the goal-setting prompt led students to study earlier and more actively by solving exercises on the online learning platform JACK. While the platform was available to all students, in both the control and treatment groups, it seems that control group students were less aware of the effectiveness of this study method. Hence, inviting students to set goals on how they want to engage with the online platform can make it more salient earlier in the semester, that such learning techniques are more effective because they facilitate active learning through self-testing opportunities.[25]

Second, the effect heterogeneity analysis shows that poor early-course behavior is associated with higher predicted treatment effects. Similarly, students who the literature has identified as at risk of falling behind in an (online) university context because they have behavioral biases, i.e. lower levels of self-control and patience, or lower prior achievement levels showed higher predicted treatment effects. Yet, it was the pre-treatment course behavior that explained most of the predicted treatment variation. Future research can shed more light on how these behavioral biases and early-course behavior are linked causally.

CRediT authorship contribution statement

Erwin Amann: Conceptualization, Writing – review & editing. **Sylvi Rzepka:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Gunther Bensch, Marco Caliendo, Christoph Hanck, Hannah Schildberg-Hörisch and seminar participants at University of Potsdam and RWI Leibniz-Institute for Economic Research for helpful comments and suggestions. We are grateful for outstanding research assistance by Kristina Nieswand and Michael Striewe for data support. The trial is registered in the AEA RCT registry, RCT ID AEARCTR-28790 (https://doi.org/10.1257/rct.2928-1.0).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econedurev.2022.102331.

[25] Using a crude measure of time-investment for other courses, we found no evidence that increased use of the online exercise tool in the microeconomics class crowded out study time in other courses. It seems there was an efficiency gain. It is up to future research to determine whether this finding withstands more fine-grained measurements.

References

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*(3), 369–378.

Clark, D., Gill, D., Prowse, V., & Rush, M. (2020). Using goals to motivate college students: Theory and evidence from field experiments. *The Review of Economics and Statistics*, 648–663.

Cobb-Clark, D. A., & Schurer, S. (2012). The stability of big-five personality traits. *Economics Letters*, *115*(1), 11–15.

Cobb-Clark, D. A., & Schurer, S. (2013). Two economists' musings on the stability of locus of control. *The Economic Journal*, *123*(570), F358–F400.

Damgaard, M. T., & Nielsen, H. S. (2018). Nudging in education. *Economics of Education Review*, *64*, 313–342.

Dobronyi, C. R., Oreopoulos, P., & Petronijevic, U. (2019). Goal setting, academic reminders, and college success: A large-scale field experiment. *Journal of Research on Educational Effectiveness*, *12*(1), 38–66.

Dotson, R. (2016). Goal setting to increase student academic performance. *Journal of School Administration Research and Development*, *1*(1), 44–46.

Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature*, *58*(4), 897–996.

Figlio, D., Rush, M., & Yin, L. (2013). Is it live or is it internet? Experimental estimates of the effects of online instruction on student learning. *Journal of Labor Economics*, *31*(4), 763–784.

Frondel, M., Osberghaus, D., & Sommer, S. (2021). *Corona and the stability of personal traits and preferences: evidence from Germany*: *ZEW-Centre for European Economic Research Discussion Paper*, (21–029).

Ganseuer, C., Linder, A., & Stammen, K.-H. (2016). Vierte zentrale Studieneingangsbefragung Wintersemester 2015/2016 bachelorstudiengänge und staatsexamen medizin. https://panel.uni-due.de/assets_websites/18/Ganseueretal_2016_ErgebnisberichtUDEPanelWiSe20152016.pdf.

Himmler, O., Jäckle, R., & Weinschenk, P. (2019). Soft commitments, reminders, and academic performance. *American Economic Journal: Applied Economics*, *11*(2), 114–142.

Kerdijk, W., Cohen-Schotanus, J., Mulder, B. F., Muntinghe, F. L. H., & Tio, R. A. (2015). Cumulative versus end-of-course assessment: effects on self-study time and test performance. *Medical Education*, *49*(7), 709–716.

Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., Turkay, S., Williams, J. J., & Tingley, D. (2020). Scaling up behavioral science interventions in online education. *Proceedings of the National Academy of Sciences*, *117*(26), 14900–14905.

Kultusministerkonferenz (2017). Abiturnoten im Ländervergleich. https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik/abiturnoten.html. (Accessed 25 January 2021).

Küpper-Tetzel, C. E. (2014). Strong effects on weak theoretical grounds: Understanding the distributed practice effect. *Z. F. Psychol*, *222*, 71–81.

Lavecchia, A. M., Liu, H., & Oreopoulos, P. (2016). Chapter 1 - behavioral economics of education: Progress and possibilities. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education*: *vol. 5*, (pp. 1–74). Elsevier.

Middendorff, E., Apolinarski, B., Becker, K., Bornkessel, P., Brandt, T., Heißenberg, S., & Poskowsky, J. (2017). Die wirtschaftliche und soziale lage der studierenden in Deutschland 2016. 21. Sozialerhebung des deutschen studentenwerks – durchgeführt vom Deutschen Zentrum für Hochschul- und Wissenschaftsforschung. Bundesministerium für Bildung und Forschung. URL http://www.sozialerhebung.de/archiv/soz_21_auszaehlung.

O'Connell, S. D., & Lang, G. (2018). Can personalized nudges improve learning in hybrid classes? experimental evidence from an introductory undergraduate course. *Journal of Research on Technology in Education*, *50*(2), 105–119.

OECD (2019). *Education at a glance 2019* (p. 520). (Accessed 26 November 2020).

Oreopoulos, P., Patterson, R. W., Petronijevic, U., & Pope, N. G. (2019). Low-touch attempts to improve time management among traditional and online college students. *Journal of Human Resources*, 0919–10426R1.

Stinebrickner, R., & Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, *32*(3), 601–644.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429–442.

Tullis, J. G., & Maddox, G. B. (2012). Self-reported use of retrieval practice varies across age and domain. *Metacognition and Learning*, *15*, 129–154.

van Lent, M. (2019). Goal setting, information, and goal revision: A field experiment. *German Economic Review*, *20*(4), e949–e972.

Vischer, T., Dohmen, T., Falk, A., Huffman, D., Schupp, J., Sunde, U., & Wagner, G. G. (2013). Validating an ultra-short survey measure of patience. *Economics Letters*, *120*(2), 142–145.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Yeomans, M., & Reich, J. (2017). Planning prompts increase and forecast course completion in massive open online courses. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 464–473). ACM.