CEPA DP No. 25

FEBRUARY 2021

The Effect of Goal-Setting Prompts in a Blended Learning Environment – Evidence from a Field Experiment

Erwin Amann
Sylvi Rzepka

CEPA Discussion Papers

Center for Economic Policy Analysis
https://www.uni-potsdam.de/cepa

University of Potsdam
August-Bebel-Straße 89, 14482 Potsdam
Tel.: +49 331 977-3225
Fax: +49 331 977-3210
E-Mail: dp-cepa@uni-potsdam.de

ISSN (online) 2628-653X

CEPA Discussion Papers can be downloaded from RePEc
https://ideas.repec.org/s/pot/cepadp.html

Published online at the Institutional Repository of the University of Potsdam
https://doi.org/10.25932/publishup-49347

CEPA Discussion Paper No. 25 February 2021

The Effect of Goal-Setting Prompts in a Blended Learning Environment – Evidence from a Field Experiment*

Erwin Amann
University Duisburg-Essen

Sylvi Rzepka
University of Potsdam

ABSTRACT

We investigate how inviting students to set task-based goals affects usage of an online learning platform and course performance. We design and implement a randomized field experiment in a large mandatory economics course with blended learning elements. The low-cost treatment induces students to use the online learning system more often, more intensively, and to begin earlier with exam preparation. Treated students perform better in the course than 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. There is no evidence that treated students spend significantly more time, rather they tend to shift to more productive learning methods. The heterogeneity analysis suggests that higher treatment effects are associated with higher levels of behavioral bias but also with poor early course behavior.

Keywords: natural field experiment, blended learning, behavioral economics, goal-setting
JEL Codes: I21, I23, C93, D91

Corresponding author:
Erwin Amann
Universität Duisburg-Essen
Fakultät für Wirtschaftswissenschaften
Professur für Mikroökonomik
Universitätsstraße 12
45117 Essen
GERMANY
E-mail: erwin.amann@uni-due.de

* 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).

1 Introduction

Higher education is 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 biases 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 current pandemic accelerating technological change in the educational sphere some form of online education is likely to stay present in higher education. The question is how to make it work to the benefit of students who struggle with higher and online education.

In this paper, we tackle this question in a blended learning environment. Specifically, we test whether encouraging students to set task-based goals for their engagement with the exercise material on the online learning platform helps them to perform better in the exam. Moreover, we investigate the study effort of the students during the semester on the online learning platform. Not only do we examine these direct effects; but, we also seek to understand through which channels the encouragement to set individual goals may lead to better academic performance: Is it an increase in study time or more efficient learning? What kind of effect heterogeneity can we identify?

We ran a randomized controlled trial in the field, a large introductory class in economics with a diverse student body. The class had several blended learning elements. Most importantly students had access to JACK, a computer-assisted online learning platform which 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 provided an incentive for all students to start early with their studying in the course. Yet, in passed years students tended to use JACK only sporadically during the semester. Therefore, our experiment intended to make students more aware of benefits of the online platform and improve self-regulated learning (Tullis & Maddox, 2012; Tullis et al., 2013). For this, 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. 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 help us to decipher heterogeneous treatment effects.

The intervention increased the usage of JACK, the computer-assisted online learning platform, among the treated and improved their exam performance. We observe that treated

students are 0.19 standard deviations more likely to take the early exam than the control group, they earn 0.19 standard deviations more net points, and achieve better grades (−0.16 standard deviations).[1] The results on the intermediary outcomes, 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. These results show that making productive study tasks salient is effective in improving class performance in a blended learning setting, beyond sheer access to material. In addition, the relatively large effect sizes suggest that such interventions may be particularly suited for diverse student groups.

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 on the final exam, we show that the 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. We also find that measures capturing pre-determined online learning behavior is most successful in explaining treatment effect heterogeneity. This suggests that early course behavior can be used to target interventions. However, indicators for self-control, patience, and prior achievement are also significantly associated with effect heterogeneity. This provides further tentative evidence that encouraging task-based goal-setting helps overcome behavioral biases and unfavorable starting conditions.

We contribute to two strands of literature. First, behavioral economics of education[2] has seen a surge in experimental studies that have shown mixed results. Studies on task-based goal-setting (Clark et al., 2020) and reminders (O'Connell & Lang, 2018) have shown to be effective in improving course-level performance. Yet, studies which tackle study time or test goal-setting with a broader perspective, e.g., not linked to just one course but to academic and personal life in general (Dobronyi et al., 2019; Oreopoulos et al., 2019), sent non-course specific reminders for staying on track (Himmler et al., 2019), or asked for grade-based goals for a course (Clark et al., 2020; van Lent, 2019) record null results. Building on these results we design an intervention that invites students to set task-based goals in a course with blended learning elements. This intervention aims at making productive learning strategies salient for the specific course. Further, our blended learning setting allows us to study whether there really is treatment-induced increase in study effort and whether there is effect heterogeneity due to behavioral biases or prior academic achievement.

Second, we add to the literature on the effectiveness of educational technology, especially computer-assisted learning programs (for a review see Escueta et al., 2020). Our interventions

[1] Note that 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".

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

uses educational technology to communicate two major findings of the psychological literature on effective studying strategies to students in an action-driven way: spacing learning and self-testing. While cramming may be effective for academic performance with respect to memorizing, it has proven ineffective for long-term retention of material (Kerdijk et al., 2015; Carpenter et al., 2012). Social psychology has shown that setting goals may help students focus and put more effort into their studying and therefore, lead to more success in university (Locke & Latham, 2002). Our intervention aims at funneling students' efforts into active learning and testing practices, i.e. attempting more math exercises and thus identifying deficits early on. This self-testing learning strategy is deemed very effective by psychologists; but, students may need guidance to implement it (Tullis et al., 2013). Our results suggest using task-based goal-setting is a promising strategy in this respect, since students in the treatment group indeed use the platform more often and outperform the control group in the exam.

The rest of the paper is structured as follows: Section 2 introduces the experimental design. Section 3 presents and discusses the main results for exam performance and usage of the online learning platform. Section 4 quantifies effect heterogeneity using causal forests and studies what determines this heterogeneity. Section 5 concludes.

2 Experimental Design

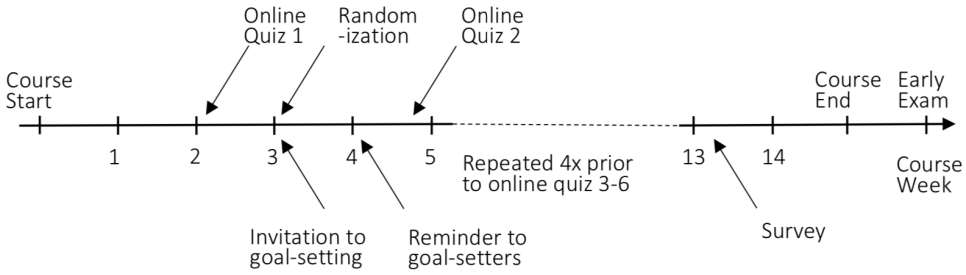
2.1 Experimental set-up

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 curricula suggest 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 the University.[3] Quiz participation is incentivized. All students can participate in 7 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.[4] In previous years, students tended to use the JACK exercises only just before these quizzes. Hence,

[3]For an overview on this platform visit: https://www.s3.uni-duisburg-essen.de/en/jack/

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

Figure 1: Timeline of the Experiment



they did not take full advantage of this tool, which provides effective self-testing exercises for students.

All students in the course who login to the in Moodle space of the course participate in the experiment. Figure 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.[5] 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 invitation to the goal-setting reads:[6]

"Dear students, next week the X. quiz will take place. As always, setting concrete 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."

Figure 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 amount 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 worked out that successful goals should be specific, measurable, attainable, relevant, and time sensitive.

Treated students who set a goal are reminded via a Moodle message of the goal they set three days in advance of each online quiz. This procedure is repeated for all quizzes 2-6.[7] Goal achievement was not monitored and this was communicated to the students in advance.

[5]Students who have not logged into Moodle by this time are excluded from the analysis since they select into treatment intensity.
[6]Figure A.1 in the appendix provides a screenshot of the original message in German.
[7]For quiz 6 no reminder of set goals was sent out.

Figure 2: Screenshot of Goal-Setting Interface prior to quiz 3

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.[8]

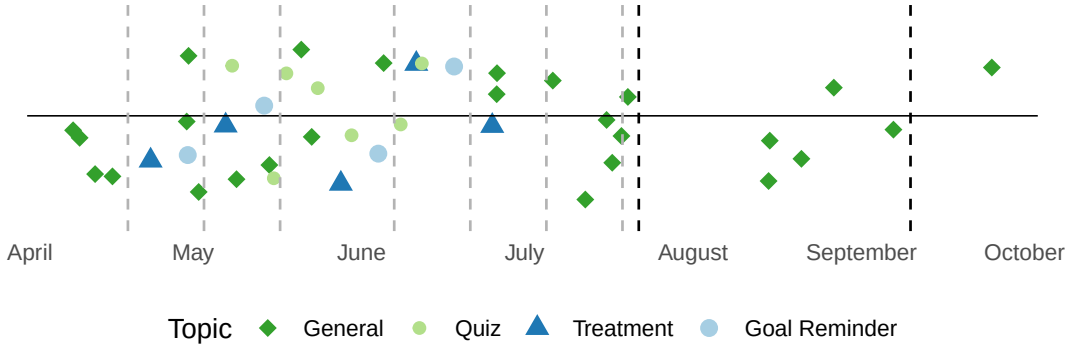
The messages of the experiment are on top of all normal course communication. Notably, the control group should be as informed about the bi-weekly online quizzes as the treatment group. Figure 3 summarizes the number and the timing of Moodle messages sent out to all course participants. Prior to the early exam in July, all participants received 20 general course messages (green diamond) and seven quiz-related messages (light green, small circle). The treated received an additional 5 messages (blue triangle), and students who set goals received up to 4 extra messages (light blue circle).

2.2 Descriptive Statistics

Randomization was successful as Table 1 shows. Nearly all pre-determined characteristics are balanced across the treatment and control group. Furthermore, the descriptive statistics are in line with expectations for the study population. Nearly half of students are male, the majority is younger than 22 years old and within the first 4 semesters of their university studies. Almost all completed 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

[8] 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 statistically significantly diff across treatment and control groups.

Figure 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.

household.[9] A third of the students receive BAfoeG, i.e. means-tested financial aid. These shares differ from the German national average, where 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 and 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 tend to be from 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 group. 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 in the same intensity, participated in the first quiz in equal shares, and earned roughly the same amount of bonus points.

[9] Academic household means at least one parent has either a university degree or a degree from a university of applied science.

Table 1: Descriptive Statistics and Balancing Tests

Covariate	Mean Control	Mean Treated	Difference	T-statistic	P-value
Study Course					
Business	0.457	0.419	0.038	0.860	0.390
Economics	0.194	0.162	0.032	0.933	0.351
Math/ Business Math	0.093	0.137	-0.044	-1.533	0.126
Business Informatics	0.140	0.183	-0.043	-1.304	0.193
Teaching Degree	0.105	0.087	0.018	0.664	0.507
Other degree or missing	0.012	0.012	-0.001	-0.084	0.933
Timing in Study					
Semester 1-2	0.430	0.386	0.044	1.006	0.315
Semester 3-4	0.450	0.402	0.047	1.063	0.288
Semester 5+	0.116	0.195	-0.079	-2.425	0.016
Semester missing	0.004	0.017	-0.013	-1.396	0.164
Abitur GPA					
Abi grade very good or good+	0.143	0.149	-0.006	-0.188	0.851
Abi grade good-	0.209	0.183	0.027	0.751	0.453
Abi grade satisfactory +	0.403	0.432	-0.028	-0.643	0.521
Abi grade satisfactory -	0.136	0.124	0.011	0.370	0.711
Abi grade pass	0.054	0.041	0.013	0.668	0.504
Other Certificate / missing	0.054	0.071	-0.016	-0.748	0.455
Bundesland of Abitur					
Abi in NRW	0.919	0.892	0.026	1.007	0.314
Abi not in NRW or missing	0.004	0.012	-0.009	-1.053	0.293
Age					
Less than 19 years old	0.143	0.124	0.019	0.620	0.536
20 to 21 years old	0.422	0.386	0.037	0.831	0.406
22 to 24 years old	0.326	0.336	-0.011	-0.249	0.803
More than 25 years old	0.105	0.133	-0.028	-0.968	0.334
Age missing	0.004	0.021	-0.017	-1.690	0.092
Gender					
Male	0.477	0.481	-0.005	-0.102	0.919
Gender missing	0.012	0.008	0.003	0.375	0.708
Financial Aid Status					
BAfoeG recipient	0.310	0.290	0.020	0.477	0.633
BAfoeG missing	0.039	0.037	0.001	0.082	0.934
Parental Background					
Academic Background	0.322	0.353	-0.031	-0.730	0.465
Background missing	0.205	0.170	0.035	1.009	0.313
Personality Traits					
Low Patience	0.388	0.361	0.027	0.613	0.540
Low Self-Control	0.341	0.390	-0.049	-1.133	0.258
Early Course Behavior (before treatment)					
Participation in test 1	0.934	0.900	0.034	1.361	0.174
Score on test 1	1.233	1.261	-0.029	-0.455	0.650
JACK-exercises attempted	6.450	6.485	-0.036	-0.047	0.962
JACK-exercises (unique)	4.891	4.859	0.033	0.066	0.948
Number of observations	258	241			

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 where at least one parent has 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.

2.3 Empirical Strategy

The randomization allows to identify the causal impact with a simple ordinary least squares regression. Hence, 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 treatment group assignment 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 time when students login to Moodle for the first time during the course, patience and self-control) renders the estimation more precise.[10]

3 Main Results

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). However, it does not affect overall exam participation (Table 2).[11] The latter is not surprising given this course is mandatory for most students enrolled. The treatment does increase the likelihood of early exam participation 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 -.16 SD compared to the control group).[12] Net of bonus points, the treated students earn about 2 points more on the exam than the control group.[13] 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;[14] but, they earn more points on their own. As Figure 4 shows the treatment shifts the middle of

[10] In the pre-analysis plan we intended to estimate the average treatment effect on the treated (ATT) using an instrumental variable strategy. We sketch this analysis in the appendix (A.1), however, do not include it into the main results section especially because the exclusion restriction may not hold.

[11] 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.

[12] In Germany lower grades are better, e.g. "1" is the best you can achieve compared to a "4" pass and "5" fail.

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

[14] Anecdotal evidence suggests that there was cooperation among students during the online quizzes, possibly across intervention groups. This may be a reason why we do not observe treatment effects for bonus points earned.

Table 2: Main Outcomes

	Exam	Early Exam	Pass	Grade	Points (net)
	(1)	(2)	(3)	(4)	(5)
Treated	0.006	0.064**	0.102**	−0.255*	1.956**
	(0.025)	(0.028)	(0.046)	(0.143)	(0.915)
Constant	0.911***	0.864***	0.540***	3.402***	29.370***
	(0.018)	(0.022)	(0.033)	(0.102)	(0.679)
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

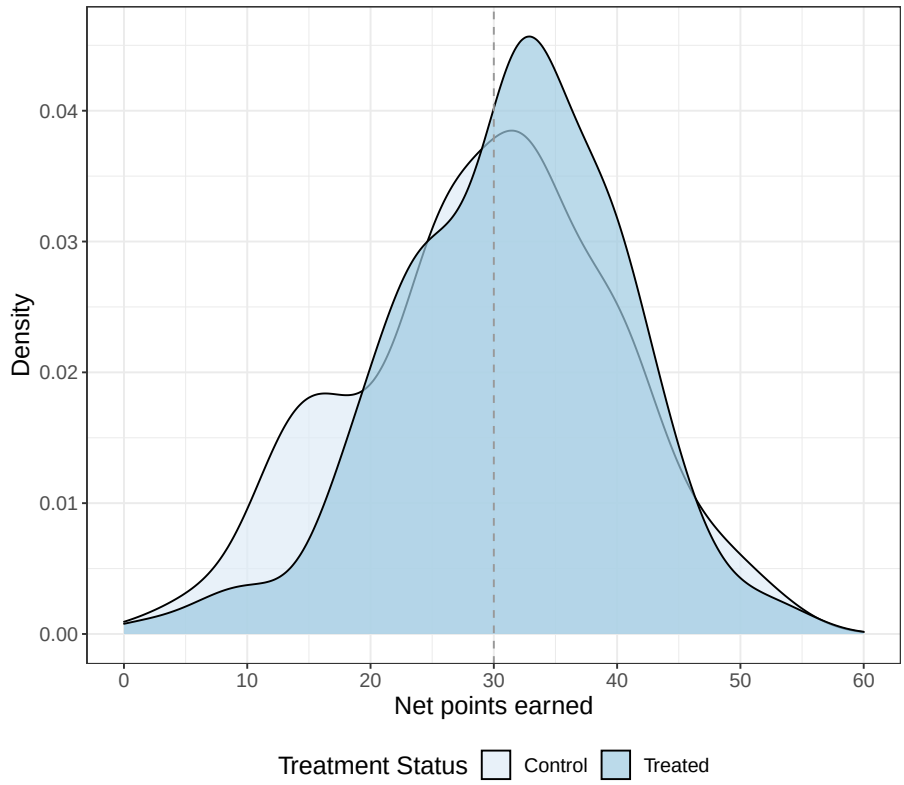
The table presents the coefficients of simple linear regressions of the treatment assignment on the outcome variables. Column 1 - 3 are binary variables indicating exam participation, participation in the first exam round and passing the exam. Columns 4 and 5 are continuous variables indicating the grade and net points. Robust standard errors 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. The standard deviations of the control group are: 0.29 for exam participation, 0.34 for early exam participation, 0.50 for passing, 1.57 for the grade, 10.41 for net points.

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.

The treatment tends to induce more intensive usage of the JACK, the online learning platform. Figure 5 shows that JACK use was the same prior to the treatment. But with each invitation to set a goal in preparation for the next quiz the average cumulative number of exercises attempted rises more among the treated than the control group. Further, the treated students tend to turn to JACK more intensively and earlier than the control group for exam preparation. 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% (.18 SD) and 7.9% (.19 SD) compared to the control group. Treated students log in for about 5.5 more sessions (13.1% and .20 SD) than the control group.[15] 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 hours. This is a relative increase of 12.2% (.17 SD) compared to the control group. Yet, this point estimate is insignificant. In short, the treatment affects effort and timing of learning on the online platform. This rationalizes why treated students outperform control group students in the exam – they grasp the material better because they practice more throughout the semester and they start earlier with their exam preparation.

[15]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.

Figure 4: Distribution of Net Points by Treatment Status

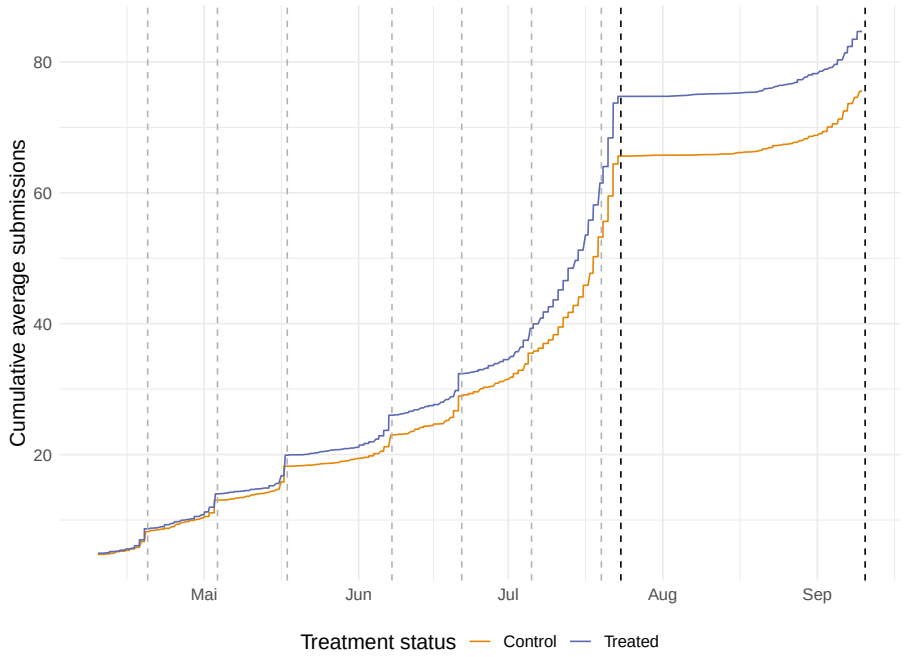


Notes: Distribution of net points by treatment status. The dashed line indicates 30 points, the total points needed to pass the exam.

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 to not affect average exam performance in Europe and the USA (van Lent, 2019; Clark et al., 2020). Yet, Clark et al. (2020), who's task-based goal 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 stronger 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 specifics in design 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 students are drawn from. While in Clark et al.

Figure 5: Timing Exercise Submission by Treatment Status



Notes: The figure shows the timing of submissions on the JACK online learning platform by treatment status. The dashed gray lines mark the times 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).

(2020) the student body comes from a "top-ranked public university", the socio-demographic characteristics (Table 1) suggest 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).

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. Yet, as Figure 6 shows there are only small and no systematic differences in time input. All in all, the treatment did not substantially change the overall time investment. For instance, the treated tend to participate in more in-class activities in all their courses. They report a higher share of being in class up to 24 hours a week than the control group (15.0% versus 9.7% of the control group). The treated invest a bit more time in individual study than the control group. In microeconomics the treated indicate more often studying individually up to 3 hours and up to 6 hours than the control group. To some extent this is also true for other courses because treated report studying up to 12 hours more often than the control group. However, the control group reports higher shares in studying up to 6, 18, or 30 hours more often than the treatment group. All four chi-square tests of independence do not reach significant levels. Using crude measures of mean time invested to conduct t-tests between

Table 3: Intermediary Outcomes

	Exercises	Exercises (unique)	Number of session	Time on platform
	(1)	(2)	(3)	(4)
treated	9.819**	2.859**	5.537**	30.095*
	(5.004)	(1.245)	(2.578)	(16.804)
Constant	80.574***	36.136***	42.403***	246.721***
	(3.441)	(0.899)	(1.807)	(11.646)
Observations	456	456	447	447
Adjusted R^2	0.006	0.009	0.008	0.005

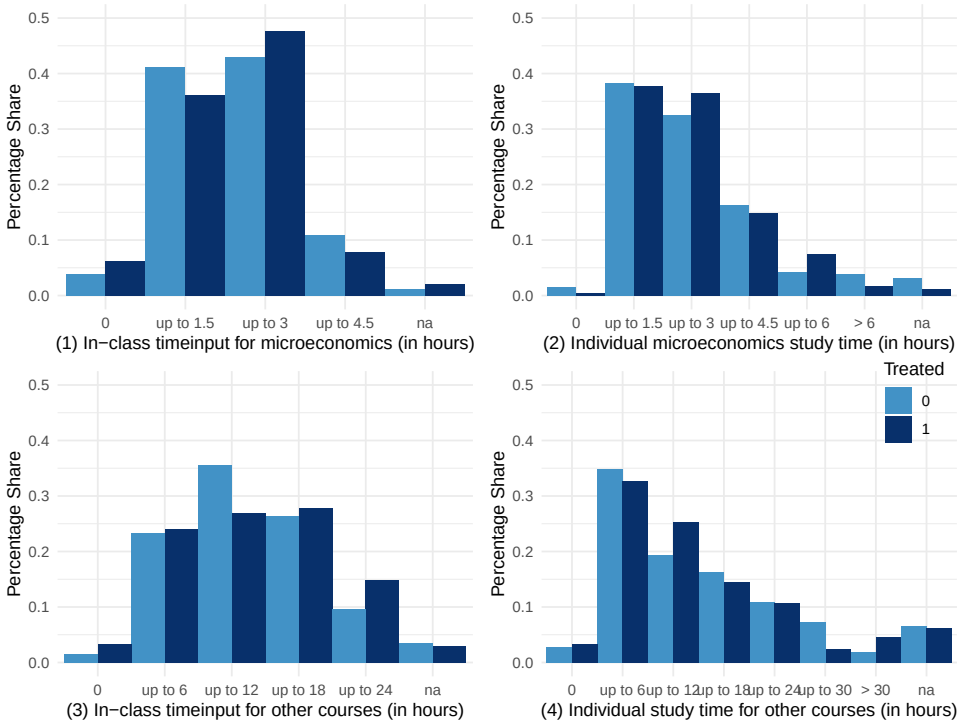
Note: *p<0.1; **p<0.05; ***p<0.01

Notes: This table presents results of simple linear regressions of the treatment on intermediary outcome variables for students who participated in the exam. 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 minutes. In column 4 "time on platform" is measured by summing up 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.4 in the appendix for the sample used in the heterogeneity analysis. The corresponding control group standard deviations are as follows: Exercises 54.32, Exercises (unique) 15.42, Sessions 27.78, Time on platform 179.47.

the treatment and control group confirm this finding.[16] In short, time investment patterns are all very similar. This means, the treatment promoted more productive study activities rather than increasing the overall time input.

[16]For these t-tests, we assigned the midpoint of the categorical time variable the student indicated as the numeric value for time investment. This transformation allowed 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.

Figure 6: Time Input by Treatment Status



Notes: This figure depicts the time investment in studying in-class and individually for the microeconomics class and all other subjects taken during the semester. The bins reflect the categories used in the survey which elicited time use. P-values of chi-square tests for independence are 0.33 (for 1), 0.19 (for 2), 0.17 (for 3), 0.13 (for 4).

4 Treatment Effect Heterogeneity

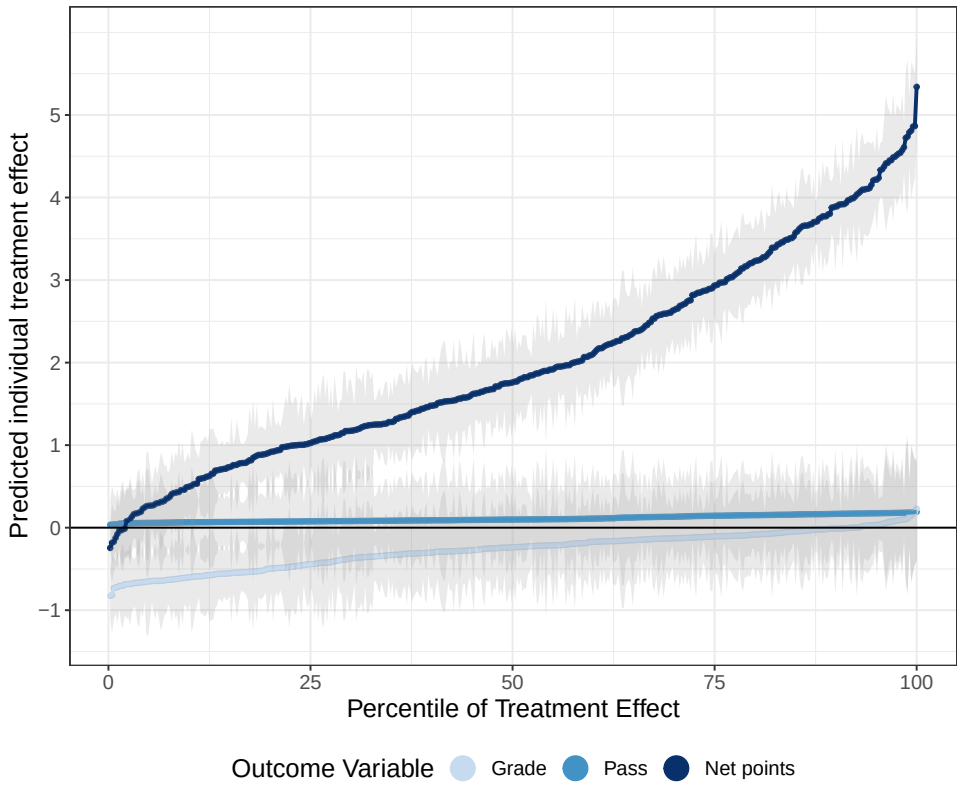
4.1 Estimating Treatment Effect Heterogeneity

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, the training data is split into two parts: One part to build the forest and a second part to populate the forest. This means the estimation of the treatment effects is independent of model building. This honesty principle is at the core of the assumptions to derive conditions for consistency and asymptotic normality of causal forests which allow to estimate 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 way. 43 pre-treatment variables enter the causal forest. Of these 15 variables are JACK-related and 36 are demographic variables. Figure A.2 shows an example of a tree that entered the causal forest. Table A.5 lists the 25 most important variables, i.e. the variables that are most predictive of treatment effect differences. Despite many more demographic variables entering into the algorithm, 14 out of the top 25 important variables are JACK-related variables.

Figure 7 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 -0.24 and 5.23, for grades between -0.48 and -0.09, and for passing the exam between 4.83% and 18.26%. 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.

Figure 7: 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 ascendingly for each outcome separately.

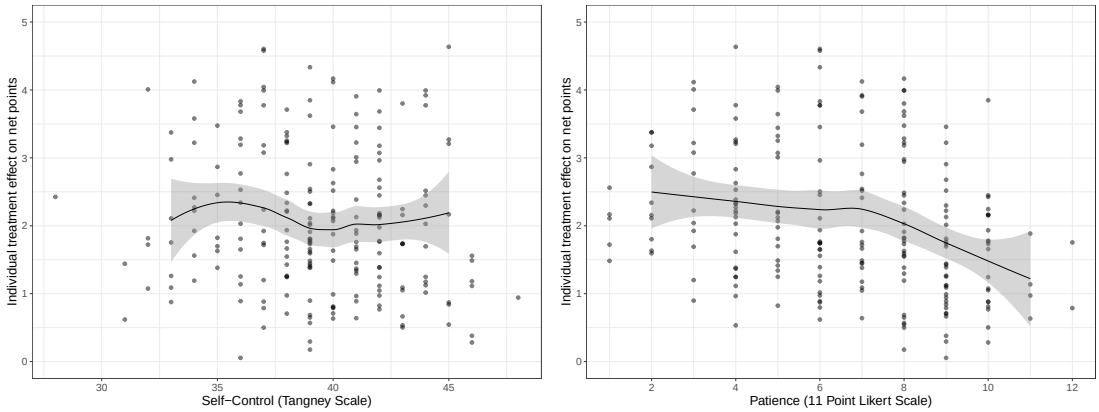
4.2 What determines Effect Heterogeneity?

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 to understand how much they are associated with the estimated effect heterogeneity. We also study 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 (Figure 8). For patience - a proxy for time-inconsistency (Vischer et al., 2013) - we do observe a slight negative association, the direction one would expect theoretically. This means the predicted treatment effects tend to be higher the more impatient students are. Figure 9 suggests that predicted treatment effects follow a more bipolar distribution for male students than for female students. This means there is a group

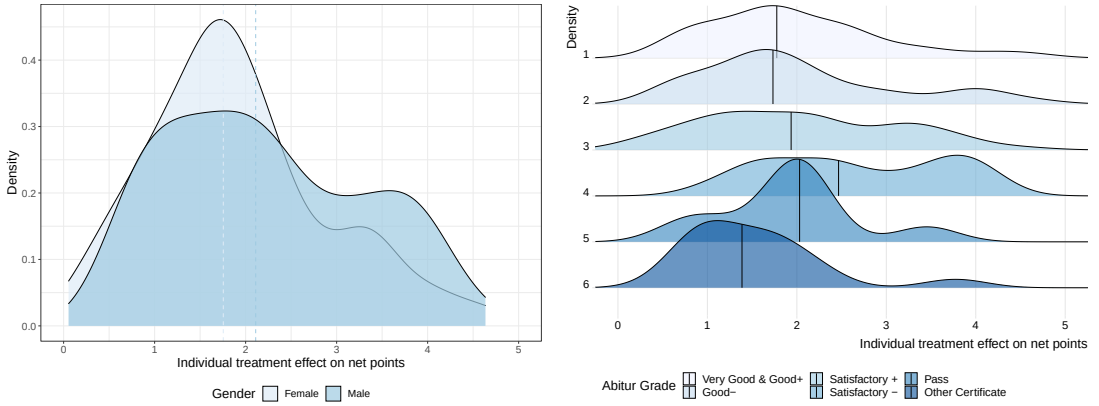
of male students for whom the treatment is predicted to have a larger effect on net points than for female students. The distribution of treatment effects of students with "satisfactory" Abitur grades, i.e. a bit below average, exhibit more mass for higher treatment effects than other achievement levels (Figure 9).

Figure 8: 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.

Figure 9: 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.

In a multiple regression of these variables on the predicted treatment effects (see Table 4) only the coefficients on patience and prior ability reach statistically significant levels. However, the adjusted R-squared only amounts to 0.125. This means, 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 (Figure 10). Further, Figure 11 shows that higher predicted treatment effects on net points are concentrated at zero or low numbers of unique exercises attempted. 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 intervention was more effective for students who performed below average early in the course.[17] We also investigate how early course behavior is associated with the predicted treatment effects on net points in a multiple regression framework (Table 4, Column 2).[18] The main conclusions discussed above hold. While not all of the course behavior variables reach statistical significant levels; the adjusted R-squared, however, skyrockets. This means, the pre-treatment behavioral variables can explain a lot 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.[19] This provides suggestive evidence that both behavioral biases and lower prior achievement levels are channels through which the task-based goal setting intervention works. However, the variation in early course behavior explains much more of the predicted treatment effect heterogeneity.

[17] 1.25 points was the average points earned on the first test.

[18] We include 12 of the 14 early behavioral variables that are among the 25 "important" variables in the causal random forest (see Table A.5 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.

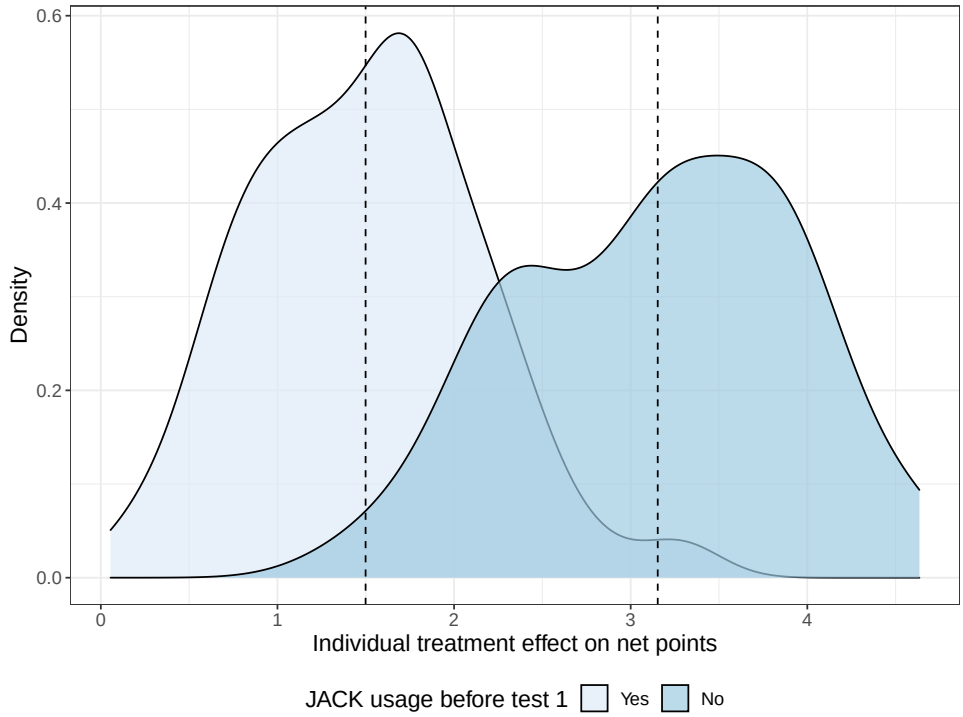
[19] The coefficient for the male dummy drops in size and changes signs. This is due to holding all other factors constant - especially whether students are active on JACK prior to the first test. More male students (38.6%) are actually not active before the first test than female students (26.8%).

Table 4: 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

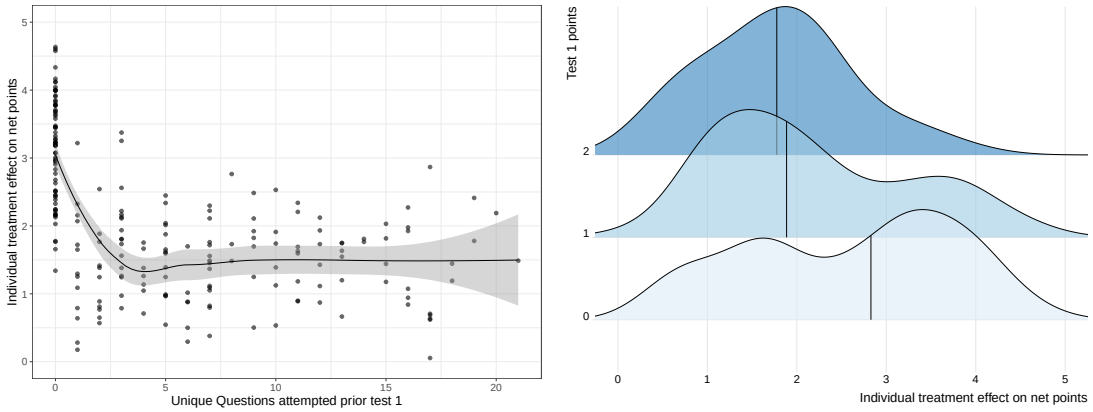
Note: Robust standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01. Column 1 also controls for: age, academic parental household, other university entry certificate, field of study, semester of study. Column 2 also controls for (all prior the first test): count of correctly answered questions, 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.

Figure 10: 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.

Figure 11: 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). The figure on the right shows 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.

5 Conclusion

Based on a randomized natural field experiment, we showed that a low-cost intervention inviting students to set task-based goals affects engagement in 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 (0.19 SD of the control group), and login for 5.5 (0.20 SD) more sessions than the control group. While not spending significantly more study time on the course, the treated students shift to more productive study modes and start their exam preparation earlier. These positive effects on engagement with the online learning platform and study timing translated into higher earlier exam participation (0.19 SD), higher passing rates (0.20 SD), 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 context. As also pointed out by Kizilcec et al. (2020), in what way, the effects of goal-setting depend on the concrete context is an interesting alley for future research.

On a more general stance, we see two main takeaways. First, our results suggest how blended learning settings can help mitigate poor course performance by design. In our setting, the invitation to goal-setting nudged students to study earlier and more actively by solving exercises on the online learning platform JACK. This proved to be a more effective learning strategy. While the platform was available to all students, both in the control and treatment group, 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 that such learning strategies more are effective because they facilitate active learning and self-testing opportunities.

Second, the effect heterogeneity analysis showed that those students who exhibit larger behavioral biases, i.e. lower levels of self-control and patience, as well as lower prior achievement levels were more positively affected by the treatment. This means that this low-touch intervention helped those most who the literature has identified as being at risk of falling behind in a (online) university context. While variables such as self-control, impatience, and prior achievement may not always be available to educators in the field, our results show that early course behavior, which is observable in online or blended learning settings, can be used to identify students who would benefit from an intervention as tested in this paper. This means that early online course behavior can be used to target interventions.

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, (pp. 648–663).

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.

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), 114–42.

Julie Tibshirani, R. F. V. H. D. H. L. M. E. S. S. W. M. W., Susan Athey (2020). grf: Generalized Random Forests. R package version 1.2.0.
URL https://github.com/grf-labs/grf

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. 117(26), 14900–14905.

Kultusministerkonferenz (2017). Abiturnoten im Ländervergleich. https://www.kmk.org/dokumentation-statistik/statistik/schulstatistik/abiturnoten.html. Accessed: 2021-01-25.

Lavecchia, A. M., Liu, H., & Oreopoulos, P. (2016). Chapter 1 - behavioral economics of education: Progress and possibilities. vol. 5 of Handbook of the Economics of Education, (pp. 1 – 74). Elsevier.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. American Psychologist, 57(9), 705–717.

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. Accessed: 2020-11-26.

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, (pp. 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.

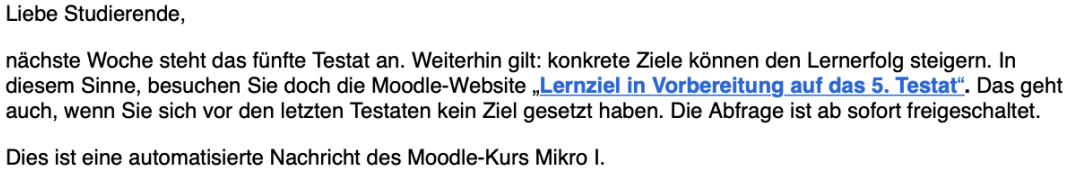
van Lent, M. (2019). Goal setting, information, and goal revision: A field experiment. German Economic Review, 20(4), e949–e972.

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.

A Appendix

Figure A.1: Wording of Goal-setting Invitation


Liebe Studierende,

nächste Woche steht das fünfte Testat an. Weiterhin gilt: konkrete Ziele können den Lernerfolg steigern. In diesem Sinne, besuchen Sie doch die Moodle-Website „Lernziel in Vorbereitung auf das 5. Testat". Das geht auch, wenn Sie sich vor den letzten Testaten kein Ziel gesetzt haben. Die Abfrage ist ab sofort freigeschaltet.

Dies ist eine automatisierte Nachricht des Moodle-Kurs Mikro I.

Table A.1: Main Outcomes with Pre-Specified Covariates

	Exam	Early Exam	Pass	Grade	Points (net)
	(1)	(2)	(3)	(4)	(5)
treated	−0.001	0.065**	0.096**	−0.244*	1.982**
	(0.026)	(0.029)	(0.045)	(0.137)	(0.862)
Constant	0.928***	0.864***	0.508***	3.643***	28.002***
	(0.039)	(0.041)	(0.067)	(0.204)	(1.232)
Observations	499	456	456	456	456
Adjusted R^2	-0.012	0.012	0.081	0.130	0.129

Note: *p<0.1; **p<0.05; ***p<0.01

Column 1 to 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, the net points. The following pre-specified covariates are included: Binary indicator of first Moodle login time, field of study, gender, indicator variables for high school GPA, financial aid recipient status, indicator variables for patience and self-control.

Table A.2: Main outcomes with sample used in heterogeneity analysis

	Early Exam	Pass	Grade	Points (net)
	(1)	(2)	(3)	(4)
treated	0.060**	0.108**	−0.270*	2.071**
	(0.028)	(0.046)	(0.144)	(0.902)
Constant	0.870***	0.550***	3.374***	29.623***
	(0.022)	(0.033)	(0.103)	(0.675)
Observations	447	447	447	447
Adjusted R^2	0.008	0.010	0.006	0.009

Note: *p<0.1; **p<0.05; ***p<0.01

Column 1 - 2 are binary variables indicating participation in the first exam round and passing the exam. Columns 3 and 4 are continuous variables indicating the grade and the net points.

Table A.3: Intermediary Outcomes with Pre-Specified Covariates

	Exercises	Exercises (unique)	Sessions	Time on platform
	(1)	(2)	(3)	(4)
treated	10.822**	3.121**	5.669**	32.214*
	(4.974)	(1.245)	(2.490)	(16.972)
Constant	76.187***	35.867***	39.603***	249.958***
	(7.196)	(1.832)	(3.426)	(25.502)
Observations	456	456	447	447
Adjusted R^2	0.047	0.015	0.087	0.032

Note: *p<0.1; **p<0.05; ***p<0.01

The outcome variable in column 1 is the number of JACK-questions answered and in column 2 the number of unique JACK-questions answered. Column 3 captures the number of sessions measured by time stamps that have interruptions shorter than 15 minutes. Column 4 adds all session times per individual. Note, the number of observations in Column 3 and 4 are lower because nine students only have one time stamp. The following pre-specified covariates are included: Binary indicator of first Moodle login time, field of study, gender, indicator variables for high school GPA, financial aid recipient status, indicator variables for patience and self-control.

Table A.4: Intermediary Outcomes (Sample used for Heterogeneity Analysis)

	Exercises	Exercises (unique)	Number of session	Time on platform
	(1)	(2)	(3)	(4)
treated	10.516**	3.136***	5.537**	30.095*
	(4.967)	(1.160)	(2.578)	(16.804)
Constant	81.970***	36.762***	42.403***	246.721***
	(3.429)	(0.858)	(1.807)	(11.646)
Observations	447	447	447	447
Adjusted R^2	0.008	0.014	0.008	0.005

Note: *p<0.1; **p<0.05; ***p<0.01

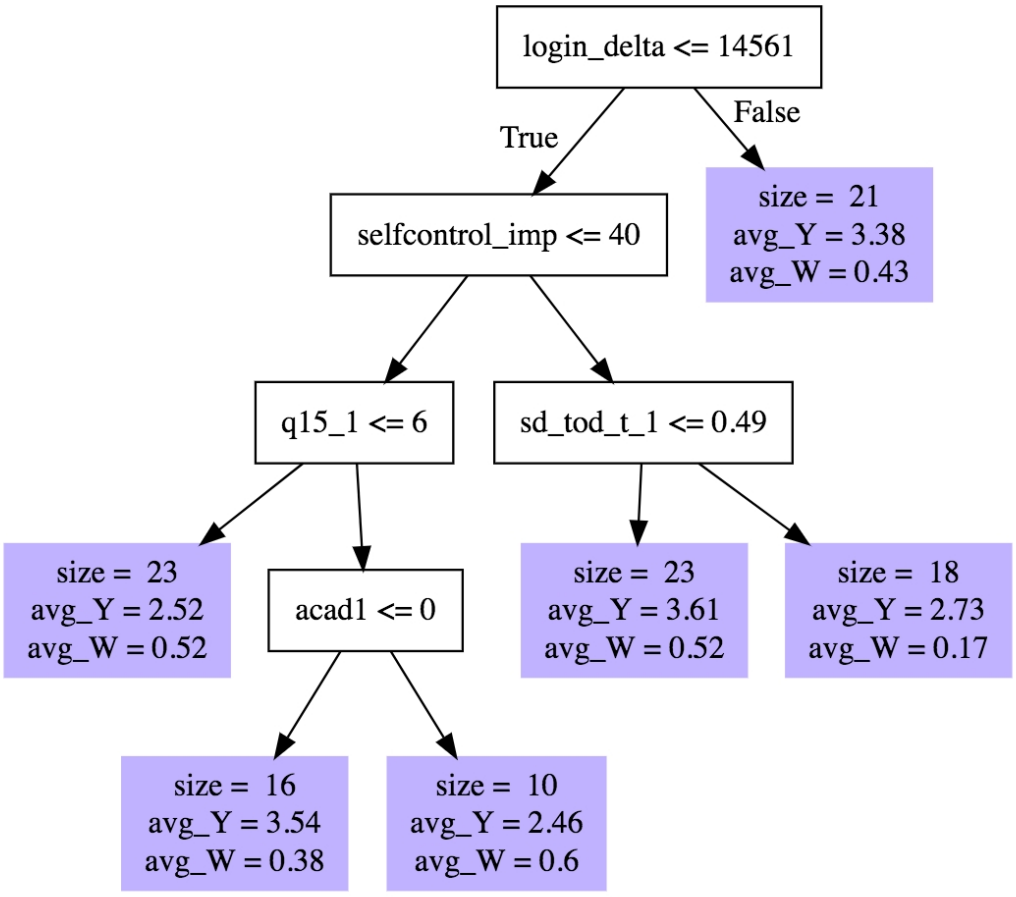
Notes: This tables shows the simple regression results for the intermediary outcomes using just the sample of students who participated in the exam and used the online learning platform at least once. It is this sample that is used for the heterogeneity analysis in Section 4.

Table A.5: Top 25 important variables

Variables	**JACK-related and other learning behavior**
Last login to JACK prior first test	yes
First Moodle login timing	yes
Median time of day logged into JACK & standard deviation	yes
Total score on JACK prior to first test	yes
Mean submissions per session & standard deviation	yes
Number of unique exercises attempted	yes
Total Duration on platform & standard deviation	yes
Number of sessions	yes
Count correct exercises	yes
No JACK activity prior to test 1	yes
Test 1 points	yes
Patience measure	no
Semester 1-2 & Semester 3-4	no
Self-control	no
Field of Study: Business	no
Field of Study: Economics	no
20 or 21 years old & 21 or 22 years old	no
Academic Parental Household	no
Male	no
Abitur GPA between 2.4 – 3.0	no

Notes: This table lists the 25 most important variables. In the *grf* R-package importance is "the weighted sum of how many times (a) feature (...) was split on at each depth in the forest" (Julie Tibshirani, 2020).

Figure A.2: Tree Example for the Causal Forest



Notes: This is an example of the trees used in the causal forest for net points as an outcome. Here the first split is done at a continuous variable (login_delta) that measures the timing of the first login to Moodle measured in seconds before the first online test. Further splits are done at self-control, patience (q_15_1), the standard deviation of the time of day students logged in to JACK prior to the first test (sd_tod_t_1), and whether students come from an academic parental household (acad1).

A.1 Average Treatment Effect on the Treated estimation

In the treatment group, 53 students (22%) interacted with the online goal-setting page on Moodle at least once. Table A.6 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.

In the main part of the paper we view "invitation to goal-setting" as the treatment. However, one could also view the actual goal-setting as the treatment and use the random assignment as an instrument for the one-sided non-compliance, i.e., estimate an average treatment effect on the treated (ATT). Results for this are reported in Table A.7. These results suggest that effects tend to be large for those students who select into goal-setting. However, these estimates are very local effects, since the absolute number of compliant students was small (53). Hence, it may also be the sheer mechanics behind this instrumentation that inflate the ATT results. Hence, the ATT results do not lend themselves for generalization. Furthermore, there are two reasons why the exclusion restriction may not hold. First, treated students who did not interact with the online goal-setting page may have set a goal in some other way. Second, the invitation message itself may have had an effect. Both aspects imply a potential violation of the exclusion restriction, since they mean that the randomization may have had other effects on the outcome variables than the observable goal-setting.

Table A.6: All goals set

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.

Table A.7: Average Treatment Effect on the treated

	First-stage	Exam	First stage	Early Exam	Pass	Net Points	Grade
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treated	0.220***		0.217***				
	(0.027)		(0.028)				
Set goal		0.028		0.294**	0.470**	9.004**	−1.174*
at least once		(0.114)		(0.135)	(0.214)	(4.294)	(0.663)
Constant	−0.000	0.911***	−0.000***	0.864***	0.540***	29.370***	3.402***
	(0.000)	(0.018)	(0.000)	(0.022)	(0.033)	(0.678)	(0.102)
Observations	499	499	456	456	456	456	456
Adjusted R^2	0.125	-0.004	0.123	-0.060	-0.023	-0.033	-0.010
F Statistic	72.443***		64.916***				

*p<0.1; **p<0.05; ***p<0.01

Notes: Column 1 and 3 present the first stage results. Columns 2 and 4-7 provides the second stage results after instrumenting for non-compliance.