INSTITUTE OF ECONOMICS
Centre for Economic and Regional Studies

Közgazdaság- és Regionális Tudományi Kutatóközpont
KÖZGAZDASÁG-TUDOMÁNYI INTÉZET

CERS-IE WORKING PAPERS | KRTK-KTI MŰHELYTANULMÁNYOK

Not just words! Effects of a light-touch randomized encouragement intervention on students' exam grades, self-efficacy, motivation, and test anxiety

TAMÁS KELLER – PÉTER SZAKÁL

CERS-IE WP – 2021/21

May 2021

https://kti.krtk.hu/wp-content/uploads/2021/05/CERSIEWP202121.pdf

ABSTRACT

Motivated by the self-determination theory of psychology, we ask how simple school practices can forge students' engagement with the academic aspect of school life. We carried out a large-scale preregistered randomized field experiment with a crossover design, involving all the students of the University of Szeged in Hungary. Our intervention consisted of an automated encouragement message that praised students' past achievements and signaled trust in their success. The treated students received encouragement messages before their exam via two channels: e-mail and SMS message. Control students did not receive any encouragement. Our primary analysis compared the end-of-semester exam grades of the treated and control students, obtained from the university's registry. Our secondary analysis explored the difference between the treated and control students' self-efficacy, motivation, and test anxiety, obtained from an online survey before students' exams. In the whole sample, we did not find an average treatment effect on students' exam grades. However, in the subsample of those who answered the endline survey, the treated students reported higher self-efficacy than the control students. The treatment affected students' motivation before their first exam—but not before their second—and did not affect students' test anxiety. Our results indicate that automated encouragement messages sent shortly before exams do not boost students' exam grades. Nevertheless, since occasionally received light-touch encouragement messages instantly increased students' self-efficacy even before an academically challenging exam situation, we conclude that encouraging students systematically and not just shortly before their exams might lead to positive emotional involvement and help create a school climate that engages students with the academic aspect of school life.

JEL codes: I23, I21, C93, D91

Keywords: Preregistered randomized field experiment, encouragement message, exam grades, test anxiety, self-efficacy, motivation

Tamás Keller
Centre for Social Sciences: Research Center for Educational and Network Studies
Research Center for Economic and Regional Studies: Institute of Economics
TÁRKI Social Research Institute, Budapest
keller.tamas@tk.hu

Péter Szakál
University of Szeged
Szakal.Peter@rekt.szte.hu

Nem csak szavak! Egy randomizált bátorító üzenet hatása a diákok vizsgaeredményeire, önhatékonyságára, motivációjára és vizsgaszorongására

KELLER TAMÁS – SZAKÁL PÉTER

ÖSSZEFOGLALÓ

A pszichológia önmeghatározás-elmélete által motiválva azt a kérdést tesszük fel, hogy egyszerű iskolai gyakorlatok hogyan fokozhatják a diákok iskolai teljesítés iránti elkötelezettségét. Egy nagymintás, előregisztrált, randomizált terepkísérletet végeztünk, amelyben a Szegedi Tudományegyetem összes hallgatója részt vett. A beavatkozásunk egy automatizált bátorító üzenetből állt, melyben a hallgatókat úgy bátorítottuk, hogy feleleventettük a korábbi sikeres vizsgáikat és kifejeztük a további sikerükbe vetett bizalmunkat. A kísérleti csoportba sorolt hallgatók a félévvégi vizsgájuk előtt két csatornán keresztül kaptak bátorító üzeneteket: e-mailben és SMS-ben. A kontroll csoportba sorolt hallgatók nem kaptak semmilyen bátorító üzenetet. Elsődleges elemzésünkben összehasonlítottuk a kísérleti és kontroll csoportba sorolt hallgatók félév végi vizsgaeredményeit. Másodlagos elemzésünkben a kísérleti és kontroll csoportba sorolt hallgatók saját képességükről vallott vélekedését hasonlítottuk össze, nevezetesen az önhatékonyságukat, motivációjukat és vizsgaszorongásukat. A bátorító üzenetek nem befolyásolták a hallgatók vizsgaeredményeit. A kísérleti csoportba sorolt hallgatók körében azonban magasabb önhatékonyság-indexet mértünk, mint a kontroll csoportba sorolt társaiknál. A kezelés hatással volt a diákok motivációjára, de csak az első vizsgájuk előtt, és a második vizsga előtt már nem. Végül a kezelés nem befolyásolta a diákok vizsgaszorongását. Eredményeink azt mutatják, hogy a röviddel a vizsgák előtt küldött automatikus bátorító üzenetek nem növelik a diákok vizsgaeredményeit. Ugyanakkor, mivel az alkalomszerűen küldött bátorító üzenet is növelte a hallgatók önhatékonyság-indexét, arra a következtetésre jutottunk, hogy a diákok szisztematikus és nem csak röviddel a vizsgák előtt történő bátorítása hozzájárulhat olyan iskolai légkör kialakításához, amely fokozhatja a diákok iskolai teljesítés iránti elkötelezettségét.

JEL: I23, I21, C93, D91

Kulcsszavak: Előregisztrált randomizált terepkísérlet, bátorító üzenet, vizsgajegyek, önhatékonyság, motiváció, vizsgaszorongás

Not just words! Effects of a light-touch randomized encouragement intervention on students' exam grades, self-efficacy, motivation, and test anxiety

Tamás Keller
1. Centre for Social Sciences: Research Center for Educational and Network Studies
2. Research Center for Economic and Regional Studies: Institute of Economics
3. TÁRKI Social Research Institute, Budapest
keller.tamas@tk.hu

Péter Szakál
4. University of Szeged
Szakal.Peter@rekt.szte.hu

Abstract

Motivated by the self-determination theory of psychology, we ask how simple school practices can forge students' engagement with the academic aspect of school life. We carried out a large-scale preregistered randomized field experiment with a crossover design, involving all the students of the University of Szeged in Hungary. Our intervention consisted of an automated encouragement message that praised students' past achievements and signaled trust in their success. The treated students received encouragement messages before their exam via two channels: e-mail and text message. Control students did not receive any encouragement. Our primary analysis compared the exam grades of the treated and control students, obtained from the university's registry. Our secondary analysis explored the difference between the treated and control students' self-efficacy, motivation, and test anxiety, obtained from an online survey before students' exams. In the whole sample, we did not find an average treatment effect on students' exam grades. However, in the subsample of those who answered the endline survey, the treated students reported higher self-efficacy than the control students. The treatment affected students' motivation before their first exam—but not before their second—and did not affect students' test anxiety. Our results indicate that automated encouragement messages sent shortly before exams do not boost students' exam grades. Nevertheless, since occasionally received light-touch encouragement messages instantly increased students' self-efficacy even before an academically challenging exam situation, we conclude that encouraging students systematically and not just shortly before their exams might lead to positive emotional involvement and help create a school climate that engages students with the academic aspect of school life.

.

Keywords

Preregistered randomized field experiment, encouragement message, exam grades, test anxiety, self-efficacy, motivation

I. Introduction

Students' engagement with the academic aspect of school life is based on positive emotional involvement in initiating and carrying out learning activities (Skinner & Belmont, 1993). Engaged students develop skills and abilities that help them to adjust to school: they maintain positive beliefs about their competence, are self-determined, and report a low level of anxiety (Miserandino, 1996). Therefore, students' engagement affects school achievement (Appleton et al., 2008) and is one of the major components in understanding dropout and promoting school completion (Christenson et al., 2012).

The self-determination theory explicates the motivational foundation of students' engagements and posits that self-motivated and self-determined behavior hinges on fulfilling fundamental needs of autonomy, competence, and relatedness (Deci, 1998; Ryan & Deci, 2000). The theory points out that contextual factors under schools' control can facilitate student self-determination and promote internalization of external school expectations (Appleton et al., 2008). By contrast, if schools have deficient practices that lead to unsuccessful school outcomes, this decreases students' self-esteem and ensures problematic behaviors that further encourage unsuccessful school outcomes (Finn, 1989). In short, specific school practices can foster engaging school climates. Supportive school practices are especially important in older age, when students have already accumulated some bad experiences that they need to overcome (Appleton et al., 2008).

This paper investigates a particular school practice introduced on an experimental basis at a Hungarian university (the University of Szeged) to develop a student-friendly university climate. We investigated first whether a light-touch intervention—an automated encouragement message—can induce an exogenous change in students' ability-beliefs, and second how much the induced change translates to a gain in students' school performance measured by their end-of-semester exam grades. We focus on three specific beliefs that express students' perceptions of their ability to some extent.

The first belief we focus on is self-efficacy: a persons' confidence in their own ability to complete a particular task (Bandura, 1977). Students' self-efficacy in regulating their learning and mastering their academic activities determines their aspirations and level of motivation (Bandura, 1993). It activates students' belief in their competence (Wigfield et al., 2015), fuels their expectancy of success (Eccles & Wigfield, 2002), regulates the amount of effort students invest in a given task, and determines how long they persevere (Bandura, 2001). Therefore, self-efficacy directly influences students' learning outcomes (Barrows et al., 2013). Furthermore, research in educational psychology has shown that self-efficacy reduces emotional stress and might have a beneficial indirect effect on students' performance (Ringeisen et al., 2018).

The second belief is the motivational belief in students' own readiness to perform a given behavior. This belief ultimately rests on trust in one's own ability. In his seminal work, Ajzen (1991) describes a similar concept—behavioral intention—which hinges on the perceived control over the intended behavior. In Ajzen's theory, behavioral intention regulates how hard students try and how much effort they exert in performing a goal. Therefore, students who intend to succeed in an exam may, in fact, be more likely to achieve success, since the stronger the intention to engage in a behavior, the more likely its realization.

The last belief we focus on is test anxiety, which is a worrisome belief students hold about their own failure (Mandler & Sarason, 1952), fueled by negative beliefs about their own ability (Cassady & Finch, 2020). Test anxiety hinders individual learning and blocks students from

presenting already acquired knowledge. Test anxiety therefore reduces academic performance (Pritchard & Wilson, 2003), as worrying about failure prevents students from concentrating on the exam (Pekrun, 2001).

Our interest in evaluating the effects of a treatment targeting these beliefs is motivated by research in economics and educational psychology showing how beliefs related to academic success are malleable (Heckman 2000; Heckman and Rubinstein 2001). Various interventions have successfully improved students school performance by developing their mindfulness (Zenner et al., 2014), social skills (Lösel & Beelmann, 2003), social-emotional competencies (Durlak et al., 2011), or self-concept (Flay et al., 2001).

Nevertheless, in educational practice, the implemented programs differ in intensity. For example, the 2-year long xl club program focused on improving students' confidence, self-esteem, and motivation (Holmlund & Silva, 2014). The intensive development of these skills implemented in small groups brought about a rise in these skills in the program.

Alongside long and intensive interventions, simple behavioral procedures like the Good Behavior Game (Dolan et al., 1993; Embry, 2002) successfully spur students' self-regulation by introducing regular routines in the daily operation of education. Participants in the program scored higher test scores in reading and mathematics than students in the matched control group (Weis et al., 2015).

Light-touch encouragements can induce a change in students' test results, particularly by targeting self-confidence and test anxiety. A small randomized trial at the University of St. Gallen in Switzerland revealed that students whose teacher read aloud a standard positive affirmation message before their exam scored higher in tests than those who had not received the positive affirmation (Behncke 2012). Furthermore, Deloatch et al. (2017) documented that highly test-anxious students who could read their Facebook-friends' affirmation messages before an exam situation scored similarly to low test-anxious students.

Still, there are at least three concerns that prevent the overgeneralization of these positive results. First, prior meta-analyses show significant heterogeneity in the effect sizes; larger studies report a smaller effect size (Lösel & Beelmann, 2003). Programs introduced in education are particularly prone to a negative correlation between sample size and effect size (Slavin & Smith, 2009). Therefore, well-executed large-scale studies that employ an experimental design and impact students' achievement via their noncognitive skills often report limited or no findings (Feron & Schils, 2020; Oreopoulos & Petronijevic, 2019). This suggests that small case studies are insufficient to determine a particular educational program's scientific validity and practical utility. Therefore, upcoming large-scale studies should corroborate the explorative results of small-scale experiments and produce conclusive evidence of the effectiveness of a given program.

Second, the efficacy of the developmental programs in education hinges on teachers' understanding of the program and their capacity to implement it (Villase, 2014). These programs either require a change in teachers' daily school routines or endow teachers with new skills . Altering teachers' daily school routines can increase teachers' workload. Teachers may thus become less motivated to implement these programs, ultimately inhibiting the program's efficacy. Integrating developmental programs into teachers' training systems and thus endowing teachers with new skills slow down the interventions' return process (Duckworth et al., 2009). Only a scant number of studies propose light-touch interventions that are ready to be integrated into educational practice without requiring teachers' motivation or experience.

Third, studies often fail to detect the particular belief or non-cognitive skill that could potentially induce the change in the targeted cognitive skills (Heckman et al., 2013). This

shortcoming is especially problematic if the intervention does not directly influence students' cognitive skills. This lack of knowledge about the treatment mechanisms could lead to an underrating of the programs' general importance, making it more difficult for future research to improve the intervention (Holmlund & Silva, 2014).

This paper advances our understanding of each of these concerns. First, we have conducted a large-scale, well-powered, and preregistered randomized field experiment that involved all the students of the University of Szeged (N = 15,539) in Hungary. Thus, our study is not specific to a particular subpopulation of students but is well powered to detect small effect sizes and capable of exploring treatment heterogeneity.

Second, we have developed an easily scalable light-touch intervention that does not require teachers' attention. Students received an encouragement message before their exam—via e-mail and text message—from the Head of the Directorate of Education at the university.

Third, we focus on particular mechanisms proposed by Behncke (2012): self-efficacy,[1] motivation, and test anxiety. Identifying the treatment mechanisms promotes innovative and more effective future treatments (Kraemer et al., 2002).

Specifically, our intervention consisted of an automated message that the treated students received before their exam. The language of the message praised students' past achievements and signaled trust in their success. Thus, we targeted students' ability beliefs by empowering them, so that our treatment could increase students' self-efficacy and motivation and decrease their test anxiety. We randomized whether students received the treatment before their first or second exam. Therefore, we could observe each student when they received and did not receive the treatment, enabling us to compare students to themselves under different conditions.

We evaluated the treatment's effect on our primary outcome—exam grades, which we assessed from the university's register. Furthermore, we investigated the treatment effect in various secondary outcomes such as self-efficacy, motivation, and test anxiety. These measures were collected via an online survey that both treated and control students filled in before the exam, and thus data on the secondary outcomes are available for a subsample of the students.

Our results show that the encouragement message had no effect on students' average exam grades (primary outcome) in the whole sample. Initially more able students, however, did achieve higher grade scores if they were encouraged. Out of our three secondary outcomes, we find a positive treatment effect in one outcome variable (self-efficacy). Specifically, treated students reported higher self-efficacy than control students. Concerning the two other secondary outcomes: in the case of students' motivation, the treatment effect is most evident in students' first exam but is attenuated in their second exam. The treatment did not translate into a significant decrease in students' test anxiety.

Based on our results, we argue that encouraging students has its own value even if it is not the appropriate tool to increase students' average exam grades. Receiving empowerment from the university contributes to feelings of importance and acknowledgment that are necessary factors in preventing university dropout (Montecel et al., 2004). Furthermore, prior research argues that many students leave university after their perception of their ability is

[1] In our pre-analysis plan, we used the term self-control instead of self-efficacy. In this paper, we use the term self-efficacy, and we refer to the same measured concept. We have changed the term we use but not the underlying empirical concept since our measure refers to the confidence in a specific task (successful exam). As Bandura (1997:382) says: "Confidence is a nondescript term that refers to strength of belief but does not necessarily specify what the certainty is about […] self -efficacy […] includes both an affirmation of a capability level and the strength of that belief."

affected by their recently awarded grades (Stinebrickner & Stinebrickner, 2012). Providing positive feedback to students may thwart these processes and contribute to a school climate that engages students.

We conclude that students' self-efficacy is sensitive to encouraging words, even if students receive them on an occasional basis shortly before an academically challenging exam situation. Encouraging students systematically and not just shortly before their exams is a possible school practice that can forge positive emotional involvement and engagement with the academic aspect of school life. Therefore, light-touch encouragement interventions might have substantial significance in themselves, even though these interventions do not directly affect students' exam grades.

II. Design, data, and method

II. 1. Preregistration

Our coding choices and statistical analysis closely follow our detailed pre-analysis plan, which we archived at the registry for randomized controlled held by the American Economic Association (https://doi.org/10.1257/rct.5155-1.1) before the end of the fieldwork and before receiving any kinds of endline data.

We archived supplementary materials, data and analytic scripts on the project's page on the Open Science framework: https://osf.io/qkfe4/. The study was reviewed and approved by the IRB office at the Centre for Social Sciences, Budapest.

II. 2. The field experiment

We conducted our field experiment at the University of Szeged (SZTE), which is the second-largest Hungarian university. The study program was initiated by the Directorate of Education of the university to develop a low-cost and easily scalable tool for decreasing dropout. The program was approved by the rector and senate of the university.

Our target population was those students engaged in full or correspondence-based education at SZTE, enrolled in the fall semester of the academic year 2019/2020, and attending classes taught in Hungarian (some students have classes taught in English). We only treated students in one study program (e.g., sociology) if they were involved in many programs (e.g., sociology and economics).

We preregistered 16,992 students at the university who met these criteria. After preregistration, 1,453 students (8.5%) changed their active status; as we could not treat them, they were excluded from the analysis. Our target population therefore contained 15,539 students. The median age of the students was 22.2, and 57% were female.

Our sample size is powered to detect a Cohen's *d* effect size of 0.03 with an 80% chance.[2] Thus, the sample is large enough to detect even a substantially small effect.

[2] We had preregistered to have at least 10,000 students in the analytical sample; therefore, the power calculation for the Cohen's *d* effect size was 0.05 in the pre-analysis plan.

II. 3. The encouragement intervention

We treated students with an intervention that consisted of an encouragement message that students received before their end-of-semester exam.[3] Treated students received an e-mail and an SMS (text) message. The email message consisted of encouragements followed by a link that prompted students to participate in the endline survey. The SMS message consisted of only the encouragement message without the link to the endline survey. The control students received an e-mail that asked them to participate in the endline survey (without the encouragement message). They did not receive an SMS message.

The English translation of the Hungarian text that treated students received in the e-mail message was as follows: "*Dear Student! The fact that you will soon take your exam proves that you already have many successful exams behind you! I truly hope that you will succeed in the next one as well, and I wish you every success! Please follow this link and answer three simple questions before your next exam. We will distribute vouchers worth a total of 100,000 HUF* [4] *that can be redeemed at the SZTE Gift Shop among the respondents*[5]*. Winners will be notified via e-mail. In the name of the Head of the Directorate of Education Péter Szakál.*"[6] Our treatment message used a very similar sentence that Behncke (2012) used successfully.[7]

The first sentence of the e-mail message praises students for their prior achievements ("you already have many successful exams behind you"). The sentence confirms students' competence, and empowers them by pointing to their successes rather than their challenges. This sentence, therefore, is intended to raise students' self-efficacy as, according to Bandura, (1977), accomplishments of past performance and verbal persuasions are important sources of self-efficacy. The sentence also aims to influence students' test anxiety since positive affirmation messages decrease students' worries (Deloatch et al., 2017). The sentence is valid for all students, since students have already taken successful exams to be admitted to the university.

The second sentence signals trust in students' success ("I truly hope that you will succeed"). The sentence is designed to be a self-fulfilling prophecy (Rosenthal & Jacobson, 1968). It is intended to affect students' behavioral intention (Ajzen, 1991) by evoking their motivation to fulfill the meaning of the sentence (Friedrich et al., 2015; Rosenthal & Jacobson, 1968).

Students in the control group received an e-mail directing their attention to the endline survey and lottery without encouragement. They received the following message: "*Dear Student! Please follow this link to answer three simple questions before your next exam. We will distribute vouchers worth a total of 100,000 HUF that can be redeemed at the SZTE Gift*

[3] These were not low-stakes exams, in contrast to the situation in similar earlier research (Behncke, 2012).
[4] About 350 USD.
[5] Students could buy various products branded with the SZTE logo in the SZTE gift shop, like office supplies, mugs, t-shirts, sweatshirts, etc. The price of an average product is under 10,000 HUF. More information: https://szteshop.hu/en/
[6] Kedves Hallgató! Az, hogy Ön hamarosan vizsgázik, annak bizonyítéka, hogy számtalan sikeres vizsga áll már Ön mögött. Őszintén bízom benne, hogy a soron következőt is sikeresen fogja teljesíteni. Ehhez sok sikert kívánok! Kérem, ezen a linken válaszoljon három egyszerű kérdésre. A válaszadók között összesen 100 ezer forint értékben sorsolunk ki az SZTE Ajándékboltba szóló utalványokat. A nyerteseket emailben értesítjük.
Az Oktatási Igazgatóság nevében Szakál Péter
[7] "I am sure that you will solve the given problems very well. You have already taken tests in the past with success; otherwise you would not be here."

Shop among the respondents. Winners will be notified via e-mail. In the name of the Head of the Directorate of Education Péter Szakál."

The sentence about the lottery in both the treated and control students' e-mail aims to motivate students to fill in the endline questionnaire. The wording of the sentence prompts students to win vouchers by making a small effort and answering just three questions.

In addition to the e-mail message, treated students received a text message before their exam on their mobile device. Similar to the e-mail, the SMS messages contain the same elements (praise for past achievements and trust) in a more condensed form. The English translation of the Hungarian SMS sentence is as follows: "*We wish you good luck in your next exam since, during your educational career, you have already successfully proved your aptitude! SZTE Education Directorate*"[8] Students in the control group did not receive any text messages on their mobile devices.

We sent out the treated and control e-mail messages at 8 pm the day before the students' exam. The treatment SMS was sent out at 7 am on the day of the exams.

The motivations behind sending out the treatment message via two channels were threefold. First, our aim was to strengthen the treatment effect by sending out the encouragement twice, while varying the language and the channel of the message. Second, we aimed to encourage students relatively close to their exams, but we could only customize sending text messages (but not e-mails). Third, we aimed to collect endline data before students' exams. Nevertheless, students are unlikely to answer a questionnaire just before their exam. Therefore, only the e-mail contained the link to the online questionnaire.

We do not know exactly when students read the treatment messages—that is, how long before the exam. Nevertheless, the date when students filled in the endline survey indicates when they might have read the e-mail. Figure 1 shows when students completed the endline survey relative to the corresponding exam. On average, students filled in the questionnaire 13 hours before their exam. This means that the treatment e-mail targeted the students a couple of hours before their exam.

[8] Soron következő vizsgájához sok sikert kívánunk, hiszen eddig tanulmányai során is eredményesen bizonyította rátermettségét! SZTE Oktatási Igazgatóság

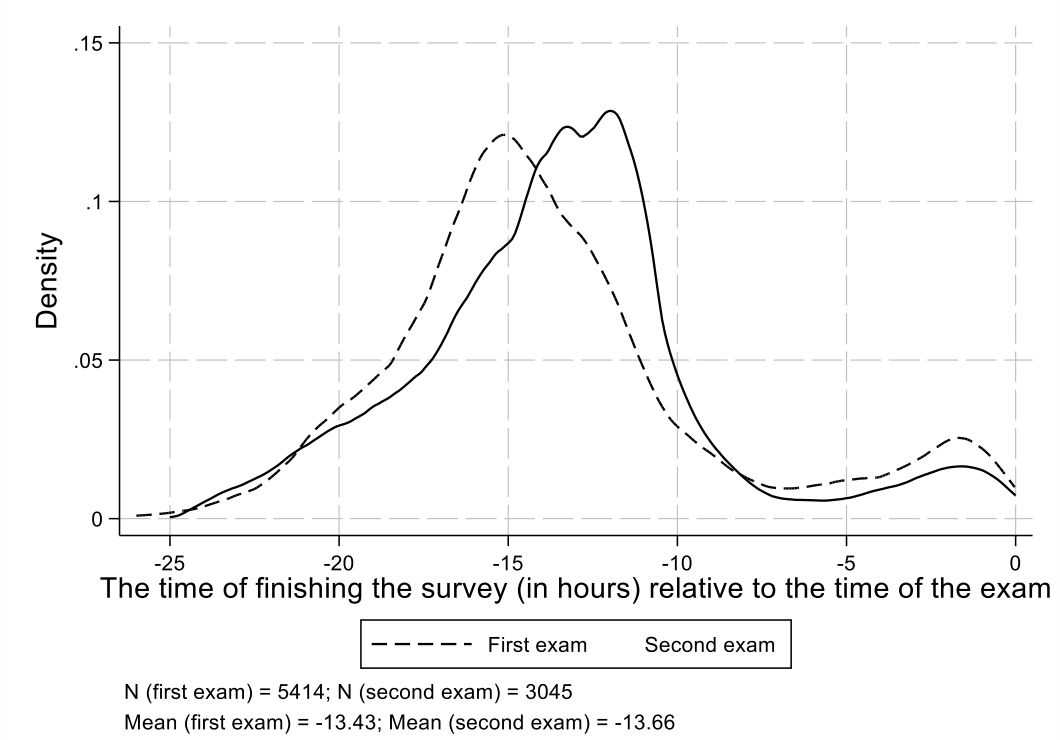


Figure 1: The relative time difference in hours between finishing the survey and the beginning of the exam

Figure 2 shows the time (in hours) relative to the exam when the treatment SMS was sent out to students' mobile devices. The majority of students (66%) received the treatment SMS 3 hours before the exam, indicating that we encouraged students shortly before their exams.

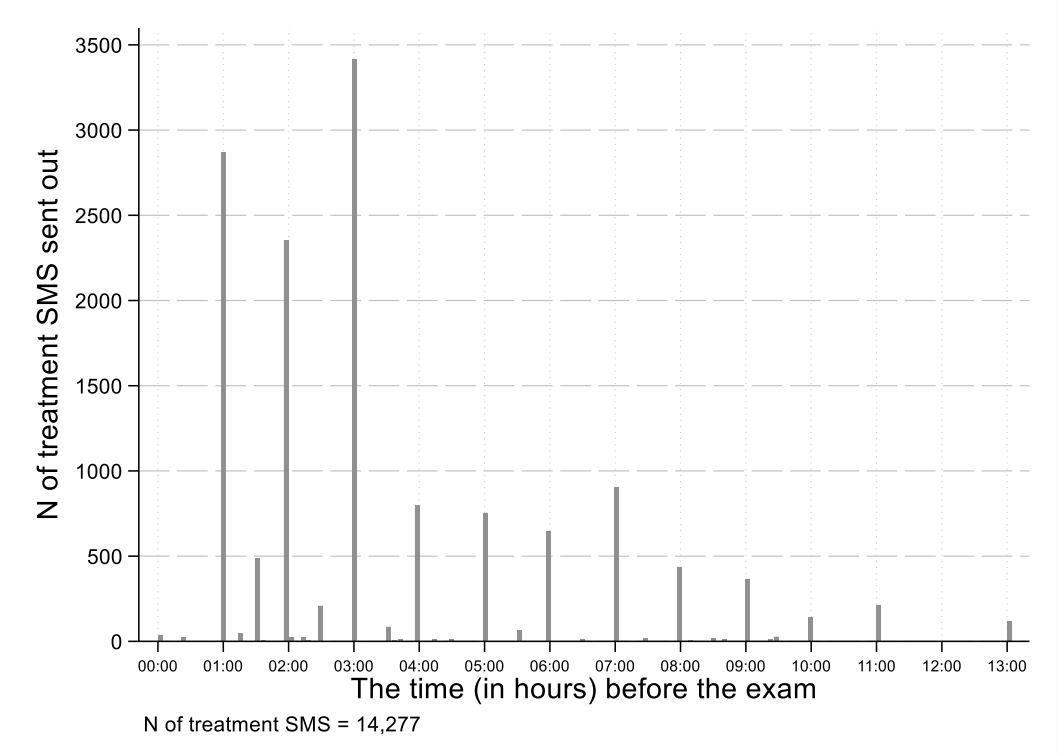


Figure 2. The time (in hours) relative to the exam when the treatment SMS was sent

Figure 3 shows the numbers of total treatment messages (e-mails and SMS) that we sent out to students taking exams on the corresponding calendar date. Approximately 80% of treatment messages were sent out in the first ten days of the campaign. This indicates a condensed treatment period, mainly concentrated in the first few days of the exam period.

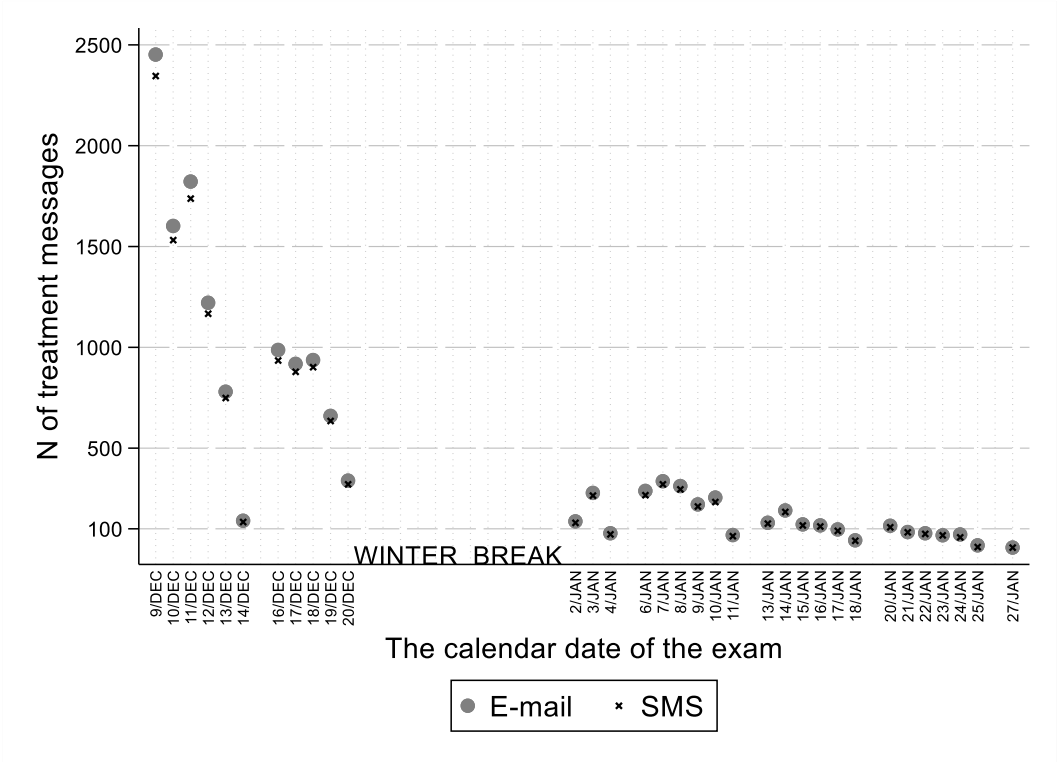


Figure 3. The total number of treatment messages (e-mail and SMS) corresponding to an exam on a particular calendar date

Note: E-mail messages were sent out at 8 pm the day before the exam.
Text messages (SMS) were sent out at 7 am on the day of the exam.
N of treatment e-mail = 14,974
N of treatment SMS = 14,277

See Appendix A for details on students' perception of the intervention.

II. 4. Study design and randomization

We designed a crossover randomized field experiment in which students act as their own control (Brown, 1980). We randomized the ordering of the treatment (at the student level), e.g., when students received the treatment. Students randomized to *Group A* received the treatment before their first exam. Students randomized to *Group B* received the treatment before their second exam[9]. Specifically, we allocated students to Group A/B based on pair-matched randomization (Imai et al., 2009)[10].

In the analytic sample of students (N = 15,539), there are 7,771 students (50.01%) who were allocated to Group A and randomized to receive the encouragement message before the

[9] The first and second exams are in different subjects—this difference is controlled for in the analysis.
[10] First, we sorted the data file according to the following baseline variables: the study program in which the student is enrolled, the level of training, the type of training, the financial form of training, students' gender, and students' ability. In the sorted data file, students who followed each other were alike. We next identified the most similar two students: students who followed each other in the data file. In the next step, within each pair, we randomly assigned students to Group A or Group B based on the value of a randomly generated number.

first exam. There are 7,768 students (49.99%) randomized to group B to receive the message before the second exam.

The design enabled us to observe all students under two conditions: when they received and when they did not receive the encouragement message. Note that we intended to treat both groups of students (A and B) for ethical reasons, but the sequences of treated and control conditions differed across the two groups. Therefore, we intended to send each student two messages (one treatment and one control message).

We re-examined the treatment status after randomization at the end of the treatment period, when all messages had been sent out. We discovered that every student had received at least one e-mail message (before their first or second exam), but not every student had received the encouragement message (e.g., they only received the control message).

Students did not receive the treatment message if their teachers entered the exam in question in the university's registry after the exam had happened. In this case, we were not able to send students the encouragement message, since the corresponding exam was not listed in the university's registry at that time. In sum, 3.65% of students (N = 565) did not receive an encouragement message. Our analysis is, therefore, an intention-to-treat (ITT) analysis.

II. 5. Balance test

Randomization resulted in groups that are well balanced with respect to the baseline covariates. Table 1 shows the differences in means between students allocated to Group A or Group B in each baseline covariate separately.

The mean difference between students in Group A (minus) those in Group B is quite small. There are only a few baseline variables (marked with bold) where the difference in means exceeds +/- 5 percentage points. Most notably, none of the differences between the two groups are statistically significant based on two-tailed t-tests.

[Table 1 about here]

II. 4. Measures

II. 4. 1. The outcome variables (Y)

The primary outcome variable is students' exam grades, measured in integers between 1 and 5. Grade 1 means fail. Other grades are equivalent to passing the exam, and in ascending order they express the quality of students' performance, with 5 as the best.

The distribution of exam grades varies according to the subject of the exam. However, pooling all exam grades at the university level, the distribution of the grades is not a bell curve. For example, in our data, the mode of the exam grade was 5 concerning students' first and second exam, which grades were received by 37% (first exam) and 34% (second exam) of students, respectively. Thus, students' GPA is usually skewed to the left. Therefore, even though relative grading is used in Hungary—that is, there is no absolute benchmark to which teachers relate students' performance—our intervention has a chance to find an effect on students' exam grades since teachers do not grade students on a curve.

In Hungary, like in many other countries, university students are required to take exams at the end of the semester. Exams can be either written or oral in nature. Students have to register for the exams on the university's online platform. They can change their registration up to 24 hours before an exam. Students who do not show up for an exam automatically fail unless a medical doctor certifies that the student was ill on the day of the exam. Therefore, the primary outcome has a missing value if a student did not show up to the exam and a medical doctor certified that he or she was ill[11]. Missing values were not replaced.

The source of the primary outcome is the university's registry. We have information on the exam grades that students were awarded in a particular subject at a particular time and date.

The secondary outcome variables are self-efficacy (1), motivation (2), and test anxiety (3) .[12] We measured these variables with three single-item questions on a scale ranging from 0 to 10. The source of the secondary outcome variables is the endline questionnaire that treated and control students voluntarily answered before their exam[13]. Figure 4 summarizes the questions we asked in the endline questionnaire and lists how the single-item measures correspond to the deployed secondary outcomes.[14]

In the table below, you can read statements that people often use to describe how they feel. Please read each statement independently and use the scale that is provided to indicate how much you consider these statements to correspond to your current mood. Your answers will not influence the grade of your exam.

	Not at all => => => => => => => => Very much											Referred as
1) I am confident to have a successful exam tomorrow	0	1	2	3	4	5	6	7	8	9	10	Self-efficacy
2) I intend to do well in the exam tomorrow	0	1	2	3	4	5	6	7	8	9	10	Motivation
3) I am anxious because of the exam tomorrow	0	1	2	3	4	5	6	7	8	9	10	Test anxiety

[11] Highly anxious students with low self-confidence might be more likely to report illness, which could cause selective attrition in the primary outcome. We tested these hypotheses in a study-program fixed effect bivariate linear probability model. We found that neither baseline text anxiety ($p = 0.7$) nor baseline self-confidence ($p = 0.28$) is associated with missingness in the primary outcome.

[12] In order to ensure a high response rate in the endline questionnaire, we only deployed a few questions. Since students were asked to respond to the survey in a stressful situation—before their exam—we could not deploy detailed measures containing many items in the questionnaire. However, we tested how our simple single-item endline variables correlated with the more detailed baseline variables. We found a pairwise correlation of 0.48 ($p<0.01$) between students' endline test anxiety measured by the single-item question, and the more detailed measure for baseline test anxiety using items from the Sarason (1980) test anxiety scale. Similarly, we found a pairwise correlation coefficient of 0.42 ($p<0.01$) between the single-item endline measure of endline self-efficacy and the more detailed measure of baseline self-confidence, deploying the academic subscale items from Shrauger and Schohn's Personal Evaluation Inventory (1995). The other correlation coefficients between various baseline and endline measures showed the expected sign. Thus, we concluded that our endline questions are good proxies of the underlying ability beliefs. Full correlation coefficients are shown in Appendix Table A1.

[13] We preregistered to delete those answers that were answered after the corresponding exam. We deleted 2,940 answers since approximately 25% of the answers to the endline questionnaire were provided after the exam.

[14] The original Hungarian version of these questions was as follows: A következő táblázatban különböző érzéseket leíró mondatokat olvashat. Kérem, egy 0-tól 10-ig terjedő skálán jelölje be, hogy a mostani érzésit figyelembe véve Ön mennyire érzi azt, hogy az adott kijelentés megfelel annak, amit most érez. A 0 azt jelenti, hogy Ön a jelenlegi érzései alapján egyáltalán nem így érez. A 10-es azt jelenti, hogy Ön jelenlegi érzései alapján teljes mértékben így érez. A 0 és 10 közötti számokkal árnyalhatja véleményét. [1] Biztos vagyok benne, hogy holnap sikeresen fogok vizsgázni; [2] Szeretnék jól teljesíteni a holnapi vizsgámon; [3] Izgulok a holnapi vizsgám miatt.

.

Figure 4. Questions students answered before their first and second exam

As students voluntarily answered the endline questionnaires, the secondary outcomes are available for a subsample of students. Appendix B summarizes the differences between the composition of various sub-samples with three highlights. First, in the subsample of those who answered the endline questionnaire, the share of students allocated to Group A versus Group B was the same as in the whole sample. Therefore, randomization was maintained with no differential selection between Groups A and B. Second, the treatment status significantly decreased students' willingness to answer the endline questionnaire; only directing students' attention to the lottery increased participation in the endline survey. Third, the subsample of students that filled in the survey was more advantaged. It contains younger and more able students who are more likely to be enrolled in full-time and state-financed education, and female students are also over-represented among them. Because the subsample of those with secondary outcomes is more advantaged, we warn against generalizing the results of the secondary outcomes to the entire analytic sample.

Since our primary outcome can take only five values, the chances to find significant treatment effects on students' exam grades are smaller than finding significant treatment effects on the secondary outcomes since these variables range between 0 and 10.

Descriptive statistics of the outcome variables in the whole sample, and in the subsample of those who answered the endline questionnaire, are summarized in Appendix C.

II. 4. 1. Treatment variable (T)

The treatment variable (T) is a 0/1 variable that indicates whether the student received the encouragement message (T=1), i.e., an e-mail and SMS before the exam. The treatment variable is coded as zero (T=0) if students received the control message, which is an e-mail without encouragement, before their exam.

II. 4. 3. The exam (E) and carry-over effects (T×E)

Students' first and second exams are from different subjects which may differ in format, scope, and difficulty. We captured these differences with a dummy variable (E) indicating whether the corresponding exam was student's first (E=0) or second exam E(=1).

The interaction of T and E indicates the carry-over effect, indicating whether the ordering of the treatment influences the outcome variables.

II. 4. 4. Control variables (X)

The preregistered control variables and their coding are listed in Appendix D.

II. 4. 5. Variables exploring treatment heterogeneity (Z)

We preregistered to explore treatment heterogeneity concerning the following baseline variables: self-confidence (1), students' ability (2), parental education (3), test anxiety (4), external control (5), students' status as a first-year student (6), students' gender as female (7), students' possession of a mobile phone number that was entered in the university's registry (8), the day (calculated from the beginning of the campaign) on which students received the message[15] (9), and difficulty of the exam (10).

III. Empirical analysis and hypothesis

III. 1. Testing the main effects (Eq.1.)

In our primary analysis, we hypothesize that receiving an encouragement message would increase students' grades in the exam.

To assess the treatment effect, we preregistered to use the following multilevel random-effects model:

$$Y_{ied} = \beta_0 + \beta_1 T_{ied} + \beta_2 E_{ied} + \beta_3 T_{ied} \times E_{ied} + \beta_4 X_{ied} + \varphi_{ied} + \mu_i + \epsilon_{ied} \qquad \text{(Eq.1.)}$$

where Y_{ied} is the *i*-th students' grade in exam *e* on day *d*. Variable T is the treatment (0/1). The variable E refers students' second exam (first=0/second=1) and controls for differences between students' first and second exams[16]. Variable X captures students' baseline variables measured before the treatment, obtained from the university's registry. We employ study-program-fixed-effect (φ_{ied}) and student-random-effect (μ_i) effects.

In our secondary analysis, we substitute Y_{ied} in Eq.1. with one of the corresponding secondary outcomes—for example, on self-efficacy, motivation, and test anxiety, respectively.

The coefficients in Eq.1 are unstandardized regression coefficients. The coefficient β_1 identifies the causal treatment effect. The coefficient is the mean difference in the first exam grades between students in the treated minus the control condition.

The coefficient β_2 identifies the period effect, i.e., the difference in exam grades between the first and second exams. The coefficient does not have a causal interpretation, since the ordering of students' exams was not randomized. The coefficient is the mean difference in control students' exam grades (the difference in mean grades control students earned at the second minus the first exam).

The coefficient β_3 identifies the carry-over effect, i.e., the difference in exam grades between the students in the treated and control conditions in the first and second exams. The coefficient is the difference of two mean-differences, i.e., the mean difference of exam grades

[15] This is a number ranging from 1 (the first day of the campaign) until the last day when students are treated.

[16] We could not include exam-subject fixed effects in the preregistered equation since this would be collinear with the preregistered exam dummy (E) indicating students' first and second exam. As a robustness check in the Appendix, we show alternative models using exam-subject fixed effects and restricting the sample to students' first (Table A3) and second exams (Table A4). Furthermore, we deploy students' fixed effects to capture all unobserved differences at the student level (Table A5). Our results are qualitatively the same as the results calculated with the preregistered model.

between treated and control students in the second exam minus the mean difference of exam grades between treated and control students in the first exam.

If there is a statistically significant $\beta_3$coefficient, students' treatment before their first exam has a long-lasting effect or long wash-out period. In other words, a significant carry-over effect reflects that encouraging students before their first exam affects their grades at the second exam; thus, the ordering of the treatment matters. A significant carry-over effect biases the estimation of the average treatment effect (Piantadosi, 2005).

Our hypothesis on the main treatment effect will be confirmed if we obtain a positive coefficient for β_1, and if we do not have a carry-over effect—i.e., if the main treatment effect concerning students' first and second exams do not differ statistically. We preregistered to use the 5% significance level concerning the primary outcome. Since we have three secondary outcomes, in the secondary analyses, we preregistered here the family-wise error rate to deal with multiple testing error (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001). We preregistered the following rules of decisions. We ordered p-values from low to high. With three secondary outcomes and the significance level of 0.05, the critical p-value would be 0.0167 for the coefficient with the lowest p-value (0.05* 1/3); this is the same as the Bonferroni correction. For the coefficient with the second-lowest p-value, the critical p-value would be 0.033 (0.05*2/3). For the coefficient with the highest p-value, the critical p-value would be 0.05 (0.05*3/3).

III. 2. Testing treatment heterogeneity (Eq.2.)

We hypothesized a greater treatment effect for students with: low self-confidence (1), a lower level of initial ability (2), and students whose parents do not have a university education (3).

We hypothesized a higher treatment effect for: anxious students (4), students with external control (5), first-year students (6), female students (7), students who had a phone number and thus received the text message parallel to the e-mail message (8), students who received the encouragement message at a later day calculated from the beginning of the campaign (9), and students who took a difficult exam (10).

In order to explore treatment heterogeneity, we included the preregistered baseline variables (Z_{ied}) in Eq.1. and in separate models, and we tested the two-way interaction of each of the Z variables with the treatment (T).

We estimated the following multilevel random-effects model to explore treatment heterogeneity:

$$Y_{ied} = \beta_0 + \beta_1 T_{ied} + \beta_2 E_{ied} + \beta_3 T_{ied} \times E_{ied} + \beta_4 X_{ied} + \beta_5 Z_{ied} + \beta_6 T_{ied} \times Z_{ied} + \varphi_{ied} + \mu_i + \epsilon_{ied} \quad \text{(Eq.2.)}$$

In Eq.2. the coefficient β_6 shows the treatment heterogeneity.

III. 3. The preregistered mediation analysis

We preregistered a mediation analysis that aimed to explore the mechanism through which the encouragement message influences exam grades. Since the main treatment effect was not

significant in any subsamples, we do not show the results in the paper. The results of the preregistered models are, however, available in the Appendix (Table A11).

IV. Results

IV. 1. Bivariate raw results

Figure 5 visualizes the unconditional raw mean of primary and secondary outcome variables in treated and control groups with 95% confidence intervals.

It is notable that in the case of three of the four outcome variables, the mean values are slightly above the theoretical middle point of the measurement scales' range. Students' motivation is the only outcome variable where the means are close to the theoretical maximum of the measurement scale, suggesting that all students were highly motivated. Thus, the potential to change students' motivation by a light-touch intervention might be limited.

The differences between the means are statistically significant in the case of self-efficacy ($p < 0.005$[17]) and motivation ($p = 0.032$) and are not statistically significant in the case of exam grades and test anxiety. All the differences are quite small.

The raw differences between the means of the treated and control groups are induced by our light-touch randomized encouragement treatment that students received a short time before their exams. Our multivariate analyses will go behind these raw differences.

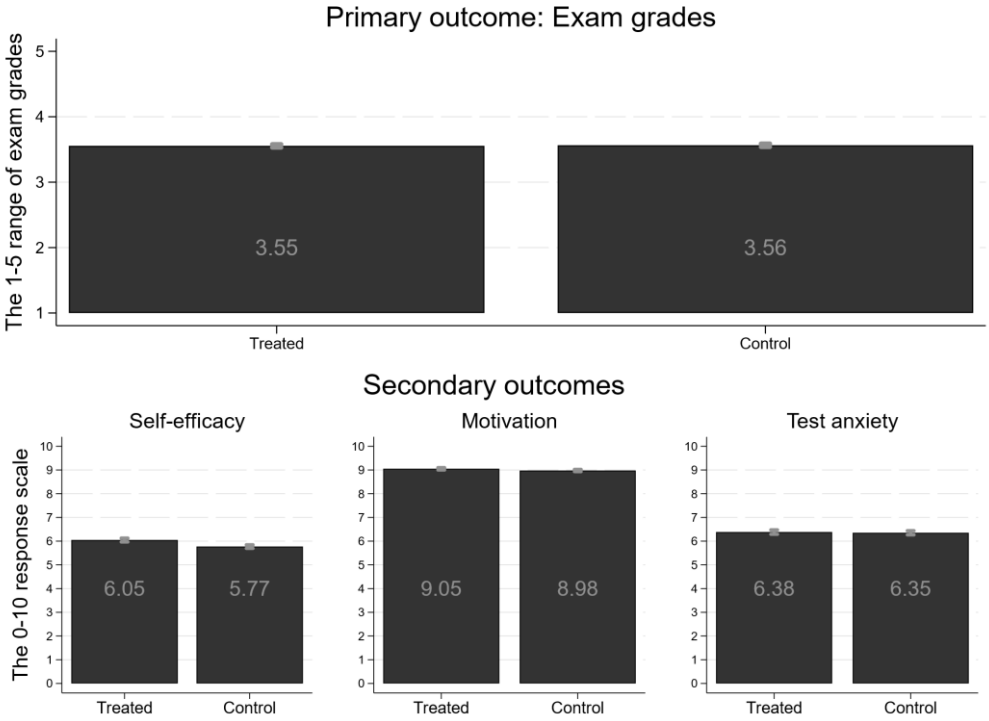


Figure 5. The unconditional raw mean of primary and secondary outcome variables in treated and control groups with 95% confidence intervals

[17] $p = 0.000000729017$

IV. 2. Exam grades

Table 2 summarizes the results for the exam grades. Column 1 shows that students who received the encouragement message did not gain higher exam grades at their first exam (β_1 = 0.017; p = 0.418). The positive treatment effect suggests that receiving the encouragement message empowered students; however, it did not significantly increase their exam grades. The Cohen's *d* effect size[18] of the treatment is small (0.011).

Students performed worse in their second exam (β_2 = -0.075; p < 0.001) than in their first exam. The results show no carry-over effect (β_3 = -0.040; p = 0.208); thus, the treatment effect was similar at students' first and second exams. In other words, receiving the encouragement message before the first exam did not have an enduring effect on students' exam grades and the instant effect of the encouragement message disappeared quickly between the two treatment points.

As Column 2 indicates, we explored treatment heterogeneity in students' baseline ability (β_6 = 0.033; p = 0.040). More able students gained a larger increase in their grades. Since we hypothesized that students with lower ability would gain more from the treatment, the result contradicts our preregistered hypothesis.

We did not find any other treatment heterogeneity regarding students' exam grades.

[Table 2 about here]

Figure 6 shows the treatment heterogeneity based on students' baseline ability. For example, among those students whose baseline ability was one standard deviation higher than the average, the encouragement message induced an increase (coef. = 0.049; p = 0.056) in their exam grades, which is statistically marginally significant.

[18] This is the treatment effect expressed in standard deviation units of the outcome variable.

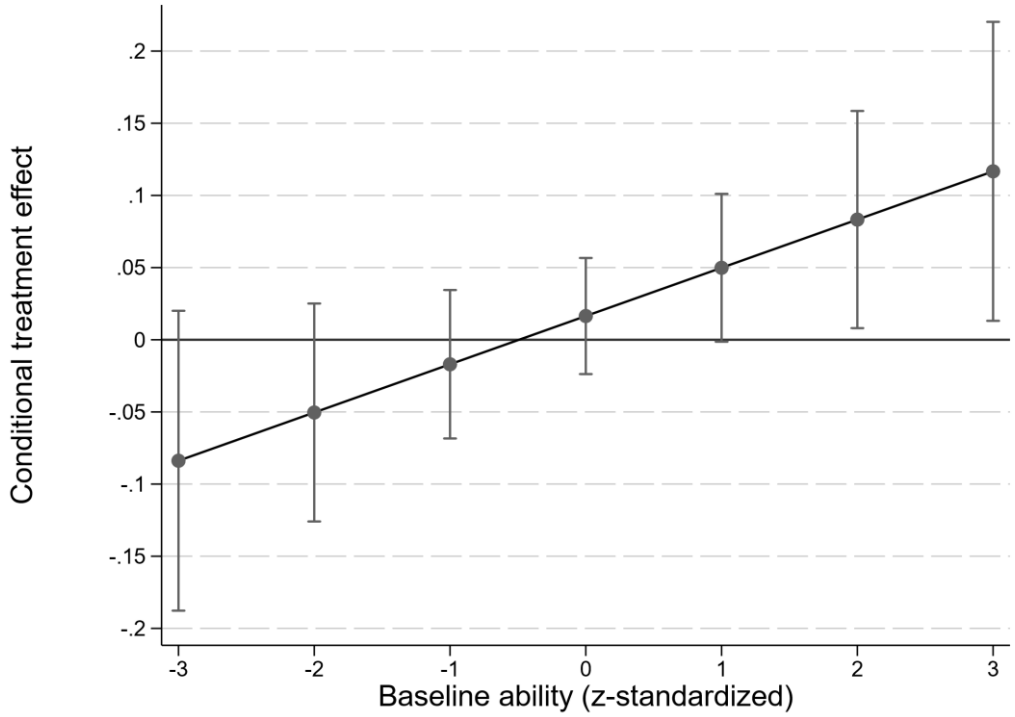


Figure 6: Conditional treatment effect of receiving the encouragement message on students' endline exam grades, based on students' baseline ability

IV. 3. Self-efficacy

Column 1 in Table 3 experimentally confirms a significant positive treatment effect on students' self-efficacy. Receiving the encouragement message increased students' self-efficacy by β_1 = 0.304 ($p < 0.001$[19]) unit, which is a Cohen's *d* effect size of 0.12.

Students reported less self-efficacy before their second exam (β_2 = -0.193; p = 0.008). The treatment effect did not differ between students' first and second exams, (β_3 = 0.027; p = 0.818), which suggests there was no carry-over effect. Therefore, the treatment had the same effect on students' self-efficacy before their first and second exams.

The main treatment effect was somewhat lower (Column 8; β_1 = 0.285; p = 0.031) in the sample of those who have baseline survey data than in the full sample (Column 1). The difference in the treatment effect between the full and restricted samples was not statistically significant (p = 0.860).

There is no treatment heterogeneity in the full sample (Columns 2–7). In the restricted sample, however, the encouragement message increased anxious students' self-efficacy (Column 9; β_6 = 0.227; p = 0.011) and also the self-efficacy of those students' whose baseline self-confidence was low (Column 10; β_6 = -0.171; p = 0.051).

[Table 3 about here]

[19] p = 0.00000322319. We preregistered here to use the significance level of 0.0167 to correct for multiple testing in the secondary outcomes. Thus, this coefficient is highly significant.

IV. 4. Motivation

Column 1 in Table 4 shows how encouragement messages influenced students' motivation to do well in the exam. We have experimentally confirmed that those who received the encouragement message experienced a 0.101 unit increase in their motivation (p = 0.013[20]), equivalent to a Cohen's *d* effect size of 0.066. Results show no difference in students' self-reported motivation (β_2 = 0.000, p = 0.995) between the first and second exam.

The marginally significant carry-over effect (β_3 = -0.121 p = 0.093) shows that the difference between treated and control students' motivation before the second exam was smaller than the same difference in students' first exam. Even though the carry-over effect was marginally significant, we suggest a cautious interpretation of the treatment effect since the encouragement message did not affect students' motivation before their second exam (0.101 + (-0.121) = -0.020, p = 0.703). Thus, the encouragement only affected students' first exam and was not replicated in the second exam.

Compared to the full sample, the treatment effect is estimated to be smaller in the restricted sample among those who filled in the baseline background questionnaire. The difference between the effects (Column 1 and Column 8) is not statistically significant (-0.045 ;p = 0.506).

As shown in Column 3, the treatment had a larger effect for older students (β_6 = 0.149; p = 0.001) and had no impact for first-year students (0.149 + (-0.151) = -0.001, p = 0.977). These findings contradict our hypothesis that first-year students, who were actually taking their first university exam and thus lacked prior experience with university exams, would gain more benefit from the encouragement campaign. The results indicate, however, that those older students who have possibly acquired a set of good/bad exam experiences are those who need encouragement to spur their motivation.

[Table 4 about here]

IV. 5. Test anxiety

Table 5 (Column 1) shows that there was no treatment effect on students' test anxiety (β_1 = -0.053; p = 0.480) concerning their first exam. Those students who received encouragement messages reported lower test anxiety than students in the control group who did not receive encouragement messages. The differences are not, however, statistically significant.

Students reported less test anxiety before their second exam than they did before their first exam. Differences in test anxiety between students' first and second exams are not, however, statistically significant (β_2 = -0.072; p = 0.387).

The carry-over effect is statistically not significant (β_3 = 0.092; p = 0.492). Thus, the ordering of the treatment (e.g., whether students received the encouragement message before their first or second exam) does not generate differences in students' test anxiety after.

[20] p = 0.0131090106. We preregistered here to use the significance level of 0.033 to correct for multiple testing in the secondary outcomes. Thus, this coefficient is significant.

The main effect (β_1 = -0.161; p = 0.297) of the treatment is somewhat larger (more negative) in the restricted sample (among those students who filled in the baseline questionnaire (Column 8). The difference in the treatment effect between the full and restricted samples is statistically not significant (p = 0.994).

The treatment effect increases (becomes more negative) during the intervention period, as indicated by the negative interaction coefficient in Column 6. With each day that is spent relative to the beginning of the treatment period, the (negative) effect increases by β_6 = -0.01 (p = 0.036). Thus, receiving the encouragement message decreases students' test anxiety significantly from the middle of the treatment period.

As hypothesized, the treatment decreased test anxiety for students in the restricted sample with average (or below average) self-confidence (Column 10). Since the interaction coefficient is positive (β_6 = 0.261 p = 0.014), if students' baseline self-confidence increases, the negative treatment effect gradually diminishes. Among students with high self-confidence, the treatment has, however, no effect.

[Table 5 about here]

IV. 6. Summary of treatment heterogeneity

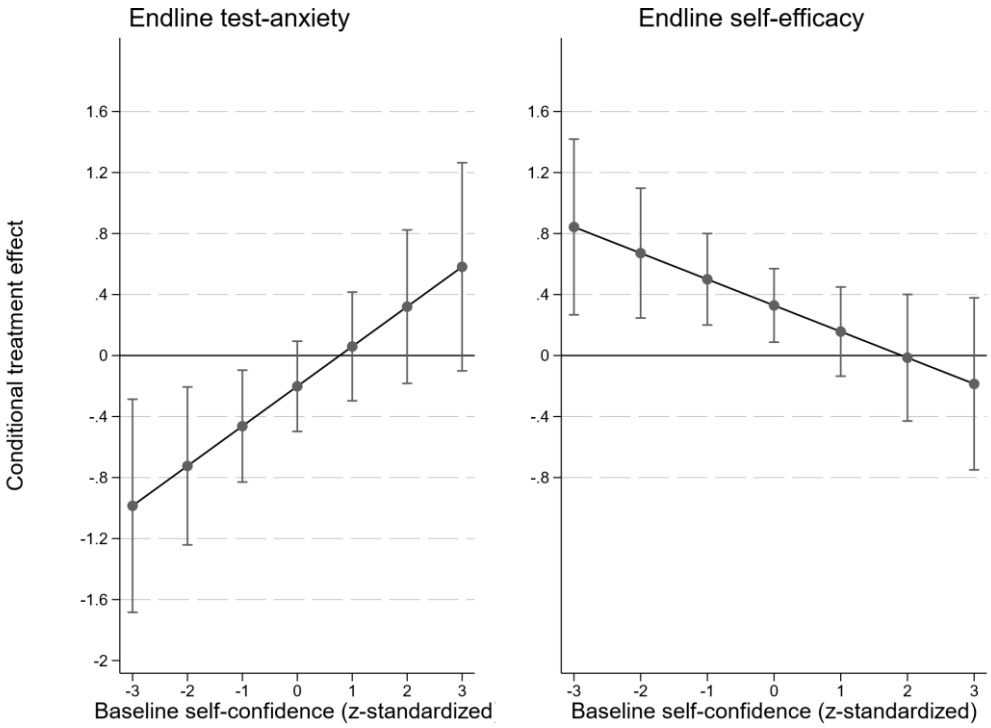
We summarized the preregistered hypothesis about treatment heterogeneity in Table 6. Most of the hypotheses were not supported since the corresponding coefficient was not statistically significant (marked with NS in the table).

We found treatment heterogeneity by students' baseline ability in our primary outcomes but did not explore any other heterogeneous effect. The result is exploratory due to multiple testing.

Significant interaction coefficients occurred sporadically across the three secondary outcome variables without a systematic pattern. Since the numbers of performed tests were large, significant interaction coefficients might have occurred by chance due to multiple testing. In other words, our results on treatment heterogeneity are exploratory, and future experimental research should confirm our exploratory results.

[Table 8 about here]

In the secondary outcomes, we can at best claim treatment heterogeneity according to students' baseline self-confidence. Here we found significant interaction coefficients for two out of the three secondary outcome variables (test anxiety and self-efficacy), as shown in Figure 8.



The left panel: corresponds to Model 10 in Table 5, N of observations = 2,014; N of respondents = 1,590
The right panel: corresponds to Model 10 in Table 6, N of observations = 2,016; N of respondents = 1,594

Figure 7: The explored conditional heterogeneous treatment effect (y-axis) on students' endline test anxiety (left panel) and endline self-efficacy (right panel), based on students' baseline self-confidence (x-axis)

V. Discussion

We carried out a large-scale, preregistered, randomized field experiment at the University of Szeged in Hungary (N=15,539 students). We tested the impact of a light-touch automated encouragement message that praised students for their past achievements. Encouragement messages were sent out via two channels: e-mail and SMS text messages.

The field experiment had a cross-over design: The treatment and control conditions varied within the same students. A random half of the students received the encouragement message before their first exam and the control message before the second exam. The other half of the students received the same message before their second exam and the control message before the first exam.

Our primary outcome variable was students' exam grades, obtained from the university's registry. We collected secondary outcome variables via an endline survey that both the treated and control students voluntarily answered before their exam. The subsample of students who answered the endline survey consisted of a more advantaged group of students regarding their baseline data, e.g., in terms of students' ability. Since we found little treatment heterogeneity in the secondary outcomes according to students' baseline variables, the potential main treatment effect in the whole analytic sample may have a similar size to the effects we observed among the more advantaged subsample of those who answered the endline survey.

Overall, our analysis provides new answers in several aspects. First, we revealed that encouraging students shortly before their exams with automated messages praising their past achievements influenced students' self-efficacy but had no or limited effect on their test anxiety and motivation. Therefore, our results suggest that self-efficacy is malleable and can be impacted by the positive feedback received independently of one's performance (Bouffard-Bouchard 1990; Tenney et al. 2015). However, the development of students' test anxiety or motivation requires a different treatment.

Second, induced self-efficacy does not translate to higher exam grades—probably because the distribution of the grades is determined, e.g., it set to a constant (grading on a curve effect). We precisely estimated zero treatment effect concerning exam grades by estimating a treatment effect close to zero with small standard errors. Thus, our results conflict with prior findings concluding that experimentally induced test anxiety, self-confidence, and motivation increase students' test performance (Behncke, 2012; Deloatch et al., 2017). The results are in line with findings that show that remedial self-esteem intervention (Holmlund & Silva, 2014) or experimentally induced optimism (Tenney et al., 2015) do not lead to higher performance.

Third, scaling up similar encouragement campaigns might have limitations since it only impacts more able students' exam grades. Thus, the success of prior interventions with a similar scope among a specific group of students cannot be generalized to the average student (Behncke, 2012; Deloatch et al., 2017).

Our findings have two important implications that warrant further consideration. First, encouraging words boost students' self-efficacy. Before exams, students receive different "messages" from their teachers, parents, and peers, concerning their ability, performance, and chances of success. Depending on the tone of these messages, each of them might increase or decrease students' self-efficacy. Our experiment reveals that students are sensitive to these words. Therefore, teachers, parents, and peers should be careful with their statements and words since these words are not *just* words but also affect students' self-efficacy.

Second, academic performance among students with initially low ability cannot be raised merely by encouragement. The encouragement instead provides a small lift in more able students' exam grades.

There are several possible explanations why we found that the intervention only affected the exam grades of more able students. More able students might be more motivated (Duckworth et al., 2012). By contrast, less able students may be less interested in gaining a good grade at the exam, and therefore not sensitive to the treatment.

Another possible reason is that students with lower baseline abilities may have less confidence in their abilities (Wigfield, 1994). Therefore, they might not believe that the encouragement message is addressed to them. In particular, students with lower ability may achieve lower grades at university. They could falsely conclude that they are not successful and regard the message as not relevant. By contrast, more able students who achieve better grades might subjectively rate themselves as more successful and therefore place greater trust in the encouragement message.

Finally, the encouragement message might help students to better recall the knowledge they have already obtained. Since students received the message shortly before the exam, it could not have increased their effort to acquire more knowledge, but the message could fine-tune how students access their existing knowledge. More able students might be better prepared for the exam and have more knowledge to mobilize when they receive encouragement. By contrast, students with initially lower ability may be less prepared and have less knowledge to

recall. Therefore, the existing difference in students' knowledge might explain how much benefit they could gain from the encouragement.

We interpret our results on the main treatment effects within the framework proposed by Jacob et al. (2019) of learning from null results. First, one should consider the typical potential growth in students' exam grades over the intervention period. In our case, the intervention period is a couple of hours (i.e., the time elapsed between the time students received the message and the exam). Within such a short period, one should not expect large changes in students' knowledge (that could be translated into higher grades). Therefore, the impact of any intervention (and not just particularly our encouragement campaign) that targets students a couple of hours before their exam might have a limited effect on students' exam grades. Thus, the precisely estimated zero results in exam grades, which suggests that the intervention had no practical significance for students' exam grades, could be attributed to the short period of time and (in addition) the light-touch (nonintensive) intervention.

Second, one should consider the theory behind the outcomes. In our case, any change in students' exam grades can be solely attributed to the change in a student's ability belief targeted by the encouragement message. By contrast, changes in the secondary outcomes can be attributed to the encouraging words that students received in the treatment message. Therefore, our results indicate that the positive beliefs we experimentally induced by the encouragement intervention do not translate into higher cognitive performance in the short run. Nevertheless, encouraging words do affect self-efficacy.

Lastly, one should consider the cost of the treatment. A low-cost intervention with a small impact might be considered successful despite the size of its impact, specifically due to the low costs. We invested about 210 USD[21] in sending out the text messages; sending out the e-mails had no incidental costs. For this level of investment, a short-lived gain in students' self-efficacy is a substantial achievement. Further, the implementation of the intervention does not require additional human effort; it could be scaled up to a virtually unlimited number of students. These features suggest that similar interventions can be worthwhile despite not directly boosting students' exam grades.

Nevertheless, policymakers and educational planners should investigate other ways to motivate low-ability students, as their exam grades seem to be resistant to encouragement. Providing useful information for organizational and time management (Abikoff et al., 2013) and gamification (Lister, 2016) are techniques that have been successfully used in prior practice and research to target students with initially low ability. Based on our findings, we recommend investing effort in future research developing new light-touch interventions in which the targeted ability belief translates into an increase in students' exam performance.

In sum, we conclude that automated encouragement messages shortly before students' exams are not a panacea for increasing students' academic achievement. However, students' self-efficacy is sensitive to encouraging words, even if these words arrive shortly before an academically challenging exam situation. Thus, encouraging students on a systematic rather than occasional basis might be a cost-effective tool for boosting students' self-efficacy. Therefore, encouragement interventions might help to create a school climate that boosts students' self-determination in the academic side of school life. They may thus have their own substantive importance (Appleton et al., 2008; Christenson et al., 2012).

[21] 60,000 HUF

References

Abikoff, H., Gallagher, R., Wells, K. C., Murray, D. W., Huang, L., Lu, F., & Petkova, E. (2013). Remediating organizational functioning in children with ADHD: Immediate and long-term effects from a randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *81*(1), 113–128. https://doi.org/10.1037/a0029648

Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179–211.

Ajzen, I. (2002). Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior. *Journal of Applied Social Psychology*, *32*(4), 665–683. https://doi.org/10.1111/j.1559-1816.2002.tb00236.x

Andrisani, P. J. (1977). Internal-External Attitudes, Personal Initiative, and the Labor Market Experience of Black and White Men. *The Journal of Human Resources*, *12*(3), 308. https://doi.org/10.2307/145493

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*(5), 369–386. https://doi.org/10.1002/pits.20303

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1993). Perceived Self-Efficacy in Cognitive Development and Functioning. *Educational Psychologist*, *28*(2), 117–148. https://doi.org/10.1207/s15326985ep2802_3

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. W. H. Freeman.

Bandura, A. (2001). Social cognitive theory: an agentic perspective. *Annual Review of Psychology*, *52*, 1–26. https://doi.org/10.1146/annurev.psych.52.1.1

Barrows, J., Dunn, S., & Lloyd, C. A. (2013). Anxiety, Self-Efficacy, and College Exam Grades. *Universal Journal of Educational Research*, *1*(3), 204–208. https://doi.org/10.13189/ujer.2013.010310

Behncke, S. (2012). How Do Shocks to Non-Cognitive Skills Affect Test Scores? *Annals of Economics and Statistics*, *107/108*, 155. https://doi.org/10.2307/23646575

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing Author. *Journal of the Royal Statistical Society. Series B*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Bouffard-Bouchard, T. (1990). Influence of Self-Efficacy on Performance in a Cognitive Task. *The Journal of Social Psychology*, *130*(3), 353–363. https://doi.org/10.1080/00224545.1990.9924591

Brown, B. W. (1980). The Crossover Experiment for Clinical Trials. *Biometrics*, *36*(1), 69. https://doi.org/10.2307/2530496

Cassady, J. C., & Finch, W. H. (2020). Revealing Nuanced Relationships Among Cognitive Test Anxiety , Motivation , and Self-Regulation Through Curvilinear Analyses. *Frontiers in Psychology*, *11*(June), 1–13. https://doi.org/10.3389/fpsyg.2020.01141

Christenson, S. L., Reschly, A. L., & Wylie, C. (2012). Handbook of Research on Student

Engagement. In *Handbook of Research on Student Engagement*. Springer. https://doi.org/10.1007/978-1-4614-2018-7

Deci, E. L. (1998). The relation of interest to the motivation of behavior: A self-determination theory perspective. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 43–70). Hillsdale. https://doi.org/10.4324/9781315807430-12

Deloatch, R., Bailey, B. P., Kirlik, A., & Zilles, C. (2017). I Need Your Encouragement! Requesting Supportive Comments on Social Media Reduces Test Anxiety. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 736–747. https://doi.org/10.1145/3025453.3025709

Dolan, L. J., Kellam, S. G., Brown, C. H., Werthamer-Larsson, L., Rebok, G. W., Mayer, L. S., Laudolff, J., Turkkan, J. S., Ford, C., & Wheeler, L. (1993). The short-term impact of two classroom-based preventive interventions on aggressive and shy behaviors and poor achievement. *Journal of Applied Developmental Psychology*, *14*(3), 317–345. https://doi.org/10.1016/0193-3973(93)90013-L

Duckworth, A. L., Quinn, P. D., & Seligman, M. E. P. (2009). Positive predictors of teacher effectiveness. *The Journal of Positive Psychology*, *4*(6), 540–547. https://doi.org/10.1080/17439760903157232

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology*, *104*(2), 439–451. https://doi.org/10.1037/a0026280

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The Impact of Enhancing Students' Social and Emotional Learning: A Meta-Analysis of School-Based Universal Interventions. *Child Development*, *82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Embry, D. D. (2002). The Good Behavior Game: A best practice candidate as a universal behavioral vaccine. *Clinical Child and Family Psychology Review*, *5*(4), 273–297. https://doi.org/10.1023/A:1020977107086

Feron, E., & Schils, T. (2020). A Randomized Field Experiment Using Self-Reflection on School Behavior to Help Students in Secondary School Reach Their Performance Potential. *Frontiers in Psychology*, *11*(June), 1–18. https://doi.org/10.3389/fpsyg.2020.01356

Finn, J. D. (1989). Withdrawing from School Author. *Review of Educational Research*, *59*(2), 117–142.

Flay, B. R., Allred, C. G., & Ordway, N. (2001). Effects of the positive action program on achievement and discipline: Two matched-control comparisons. *Prevention Science*, *2*(2), 71–89. https://doi.org/10.1023/A:1011591613728

Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K., & Trautwein, U. (2015). Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement. *Contemporary Educational Psychology*, *41*, 1–12. https://doi.org/10.1016/j.cedpsych.2014.10.006

Goldsmith, A. H., Veum, J. R., & William, D. (1996). The impact of labor force history on

self-esteem and its component parts, anxiety, alienation and depression. *Journal of Economic Psychology*, *17*(2), 183–220. https://doi.org/10.1016/0167-4870(96)00003-7

Heckman, J., Pinto, R., & Savelyev, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, *103*(6), 2052–2086. https://doi.org/10.1257/aer.103.6.2052

Holmlund, H., & Silva, O. (2014). Targeting Noncognitive Skills to Improve Cognitive Outcomes: Evidence from a Remedial Education Intervention. *Journal of Human Capital*, *8*(2), 126–160. https://doi.org/10.1086/676460

Imai, K., & Jiang, Z. (2018). A sensitivity analysis for missing outcomes due to truncation by death under the matched-pairs design. *Statistics in Medicine*, *37*(20), 2907–2922. https://doi.org/10.1002/sim.7802

Imai, K., King, G., & Nall, C. (2009). The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, *24*(1), 29–53. https://doi.org/10.1214/08-STS274

Jacob, R. T., Doolittle, F., Kemple, J., & Somers, M. A. (2019). A Framework for Learning From Null Results. *Educational Researcher*, *48*(9), 580–589. https://doi.org/10.3102/0013189X19891955

Kraemer, H. C., Wilson, G. T., Fairburn, C. G., & Agras, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry*, *59*(10), 877–883. https://doi.org/10.1001/archpsyc.59.10.877

Lister, M. (2016). Gamification: The effect on student motivation and performance at the post-secondary level. *Issues and Trends in Educational Technology*, *3*(2). https://doi.org/10.2458/azu_itet_v3i2_lister

Lösel, F., & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behavior: A systematic review of randomized evaluations. *Annals of the American Academy of Political and Social Science*, *587*, 84–109. https://doi.org/10.1177/0002716202250793

Mandler, G., & Sarason, S. B. (1952). A study of anxiety and learning. *The Journal of Abnormal and Social Psychology*, *47*(2), 166–173. https://doi.org/10.1037/h0062855

Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, *88*(2), 203–214. https://doi.org/10.1037/0022-0663.88.2.203

Montecel, M. R., Cortez, J. D., & Cortez, A. (2004). Dropout-Prevention Programs: Right Intent, Wrong Focus, and Some Suggestions on Where to Go from Here. *Education and Urban Society*, *36*(2), 169–188. https://doi.org/10.1177/0013124503261327

Oreopoulos, P., & Petronijevic, U. (2019). *The Remarkable Unresponsiveness of College Students to Nudging And What We Can Learn from It*. https://doi.org/10.3386/w26059

Pekrun, R. (2001). Test Anxiety and Academic Achievement. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 15610–15614). Elsevier. https://doi.org/10.1016/B0-08-043076-7/02451-7

Piantadosi, S. (2005). Clinical Trials. In W. A. Shewhart & S. S. Wilks (Eds.), *Clinical Trials: A Methodologic Perspective* (Second). John Wiley & Sons, Inc. https://doi.org/10.1002/0471740136

Pritchard, M. E., & Wilson, G. S. (Gregory S. (2003). Using Emotional and Social Factors to

Predict Student Success. *Journal of College Student Development*, *44*(1), 18–28. https://doi.org/10.1353/csd.2003.0008

Ringeisen, T., Lichtenfeld, S., Becker, S., Minkley, N., Ringeisen, T., Lichtenfeld, S., Becker, S., Minkley, N., Ringeisen, T., Lichtenfeld, S., & Becker, S. (2018). Stress experience and performance during an oral exam : the role of self-efficacy , threat appraisals , anxiety , and cortisol. *Anxiety, Stress, & Coping*, *0*(0), 1–17. https://doi.org/10.1080/10615806.2018.1528528

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*, *3*(1), 16–20.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28. https://doi.org/10.1037/h0092976

Rotter, J. B. (1992). Some comments on the "Cognates of personal control." *Applied and Preventive Psychology*, *1*(2), 127–129. https://doi.org/10.1016/S0962-1849(05)80154-4

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sarason, I. G. (1980). *Test Anxiety: Theory, Research, and Applications*. Lawrence Erlbaum Associates. http://www.mrc.stlmath.com/pdf/anxiety/scale.pdf

Shrauger, J. S., & Schohn, M. (1995). Self-Confidence in College Students: Conceptualization, Measurement, and Behavioral Implications. *Assessment*, *2*(3), 255–278. https://doi.org/10.1177/1073191195002003006

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, *85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis*, *31*(4), 500–506. https://doi.org/10.3102/0162373709352369

Stinebrickner, T. R., & Stinebrickner, R. (2012). Learning about academic ability and the college dropout decision. *Journal of Labor Economics*, *30*(4), 707–748. https://doi.org/10.1086/666525

Tenney, E. R., Logg, J. M., & Moore, D. A. (2015). (Too) optimistic about optimism: The belief that optimism improves performance. *Journal of Personality and Social Psychology*, *108*(3), 377–399. https://doi.org/10.1037/pspa0000018

Villase, P. (2014). *The different ways that teachers can influence the socio-emotional development of their students : A literature review* (Issue 2013). World Bank Publications. http://pubdocs.worldbank.org/en/285491571864192787/Villaseno-The-different-ways-that-teachers-can-influence-the-socio-emotional-dev-of-students.pdf

Weis, R., Osborne, K. J., & Dean, E. L. (2015). Effectiveness of a Universal, Interdependent Group Contingency Program on Children's Academic Achievement: A Countywide Evaluation. *Journal of Applied School Psychology*, *31*(3), 199–218. https://doi.org/10.1080/15377903.2015.1025322

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*(1), 49–78.

https://doi.org/10.1006/ceps.1999.1015

Wigfield, A., Eccles, J. S., Fredricks, J. A., Simpkins, S., Roeser, R. W., & Schiefele, U. (2015). Development of Achievement Motivation and Engagement. In *Handbook of Child Psychology and Developmental Science* (Issue March). https://doi.org/10.1002/9781118963418.childpsy316

Zenner, C., Herrnleben-Kurz, S., & Walach, H. (2014). Mindfulness-based interventions in schools-A systematic review and meta-analysis. *Frontiers in Psychology*, *5*(JUN). https://doi.org/10.3389/fpsyg.2014.00603

Tables

Table 1. Balance test. The difference in means between students allocated to Group A relative to Group B for each baseline covariate separately

	All students	Students with Endline Questionnaire	Students with Baseline Questionnaire
Female	-0.008	0.000	-0.038
Age	0.000	0.001	0.001
Students' ability	-0.002	-0.002	0.001
Students' ability is missing	-0.002	0.001	0.000
Full-time training	0.000	-0.006	0.038
State-financed training	0.005	0.003	-0.018
Bachelor level	-0.001	-0.002	-0.004
Master level	-0.003	0.001	-0.014
Undivided	0.003	0.005	0.022
Higher-level vocational training	0.001	-0.007	-0.048
First-year students	0.003	-0.005	0.009
Exam difficulty	0.001	0.015	**0.074**
Exam difficulty is missing	0.001	-0.003	0.005
Baseline test anxiety	0.002[$]	-0.005[$$]	0.002
Baseline self-confidence	-0.011[$]	-0.007[$$]	-0.011
External control	0.019[$]	0.003[$$]	0.019
Parental education (university degree	0.029[$]	0.037[$$]	0.029
N	15,539	7,026	-0.038

* The difference is significant at 5% level using a two-tailed t-test.
[$] N = 2,305; [$$] N = 1,612
Bold coefficients mark the mean differences that are larger than +/- 0.05.

Table 2: Treatment effect on students' endline exam grades, unstandardized regression coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T]	0.017	0.016	0.011	0.039	-0.006	0.018	0.003	0.048	0.049	0.052	0.051	0.031
(treated = 1)	(0.021)	(0.021)	(0.022)	(0.026)	(0.065)	(0.022)	(0.023)	(0.052)	(0.052)	(0.052)	(0.052)	(0.064)
β_2: Exam [E]	-0.075***	-0.075***	-0.075***	-0.075***	-0.075***	0.052*	-0.074***	-0.025	-0.023	-0.021	-0.022	-0.026
(second = 1)	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)	(0.022)	(0.021)	(0.053)	(0.053)	(0.053)	(0.053)	(0.053)
β_3: Carry-over [T×E]	-0.040	-0.040	-0.040	-0.040	-0.040	-0.038	-0.043	-0.079	-0.082	-0.085	-0.086	-0.078
	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.032)	(0.080)	(0.080)	(0.079)	(0.080)	(0.080)
β_6: Interaction[a]		0.033*	0.019	-0.039	0.024	-0.000	0.083		0.007	-0.009	0.017	0.031
(T×Main effet[Z])		(0.016)	(0.029)	(0.028)	(0.064)	(0.001)	(0.058)		(0.035)	(0.034)	(0.035)	(0.069)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	-0.080**	✓	✓	✓
									(0.027)			
Baseline self-confidence[b]								✓	✓	0.168***	✓	✓
										(0.027)		
Baseline external control[b]								✓	✓		-0.072**	✓
											(0.027)	
Parental education								✓	✓	✓	✓	0.009
												(0.055)
Students' ability[b]	✓	0.208***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		(0.015)										
First-year student	✓	✓	-0.102***	✓	✓	✓	✓	✓	✓	✓	✓	✓
			(0.025)									
Female	✓	✓	✓	0.177***	✓	✓	✓	✓	✓	✓	✓	✓
				(0.023)								
Has mobile phone					-0.005							
					(0.050)							
Day of message						-0.019***						
						(0.001)						
Exam difficulty	✓	✓	✓	✓	✓	✓	-1.659***	✓	✓	✓	✓	✓
							(0.047)					
Constant	3.730***	3.735***	3.732***	3.718***	3.735***	3.550***	3.733***	2.305*	2.323*	2.348*	2.372*	2.277*
	(0.354)	(0.354)	(0.354)	(0.354)	(0.357)	(0.348)	(0.354)	(1.081)	(1.079)	(1.070)	(1.080)	(1.082)

Observations	28,156	28,156	28,156	28,156	28,156	28,156	28,156	4,335	4,335	4,335	4,335	4,335
N of students	15,264	15,264	15,264	15,264	15,264	15,264	15,264	2,295	2,295	2,295	2,295	2,295
Cohen's *d* effect size	0.011	0.011	0.007	0.027	-0.004	0.012	0.002	0.034	0.035	0.037	0.037	0.022

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 3: Treatment effect on students' endline self-efficacy, unstandardized regression coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T] *(treated = 1)*	0.304*** (0.065)	0.306*** (0.065)	0.344*** (0.073)	0.316*** (0.087)	0.628** (0.219)	0.291*** (0.070)	0.233** (0.075)	0.285* (0.132)	0.289* (0.127)	0.329** (0.123)	0.277* (0.131)	0.231 (0.166)
β_2: Exam [E] *(second = 1)*	-0.193** (0.073)	-0.194** (0.073)	-0.194** (0.073)	-0.193** (0.073)	-0.192** (0.073)	0.000 (0.077)	-0.181* (0.073)	-0.089 (0.151)	-0.063 (0.146)	-0.009 (0.142)	-0.091 (0.150)	-0.085 (0.151)
β_3: Carry-over [T×E]	-0.027 (0.116)	-0.026 (0.116)	-0.023 (0.116)	-0.027 (0.116)	-0.027 (0.116)	-0.016 (0.121)	-0.048 (0.117)	-0.033 (0.246)	-0.013 (0.235)	-0.105 (0.226)	-0.024 (0.244)	-0.041 (0.246)
β_6: Interaction[a] (T×Main effet[Z])		-0.057 (0.056)	-0.125 (0.100)	-0.021 (0.096)	-0.341 (0.219)	0.001 (0.004)	0.385+ (0.200)		0.227* (0.089)	-0.171+ (0.088)	0.084 (0.091)	0.111 (0.182)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	-0.828*** (0.073)	✓	✓	✓
Baseline self-confidence[b]								✓	✓	1.085*** (0.070)	✓	✓
Baseline external control[b]								✓	✓	✓	-0.333*** (0.073)	✓
Parental education								✓	✓	✓	✓	-0.210 (0.152)
Students' ability[b]	✓	0.104* (0.050)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
First-year student	✓	✓	0.157+ (0.082)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Female	✓	✓	✓	-0.397*** (0.079)	✓	✓	✓	✓	✓	✓	✓	✓
Has mobile phone					0.169 (0.173)							
Day of message						-0.024*** (0.003)						
Exam difficulty	✓	✓	✓	✓	✓	✓	-1.621*** (0.156)	✓	✓	✓	✓	✓
Constant	7.612*** (1.401)	7.600*** (1.401)	7.597*** (1.401)	7.610*** (1.401)	7.430*** (1.409)	7.422*** (1.394)	7.641*** (1.400)	6.728* (2.977)	7.551** (2.848)	6.522* (2.738)	6.995* (2.957)	6.872* (2.979)
Observations	8,296	8,296	8,296	8,296	8,296	8,296	8,296	2,016	2,016	2,016	2,016	2,016

N of students	6,908	6,908	6,908	6,908	6,908	6,908	6,908	1,594	1,594	1,594	1,594	1,594
Cohen's *d* effect size	0.12	0.12	0.14	0.13	0.25	0.12	0.09	0.11	0.11	0.13	0.11	0.09

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 4: Treatment effect on students' endline motivation, unstandardized regression coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T] *(treated = 1)*	0.101* (0.041)	0.100* (0.041)	0.149*** (0.045)	0.100+ (0.054)	0.229+ (0.138)	0.100* (0.044)	0.075 (0.047)	0.052 (0.078)	0.052 (0.078)	0.057 (0.078)	0.051 (0.078)	0.170+ (0.098)
β_2: Exam [E] *(second = 1)*	-0.000 (0.045)	-0.000 (0.045)	-0.002 (0.045)	-0.000 (0.045)	0.000 (0.045)	0.046 (0.048)	0.004 (0.046)	0.028 (0.090)	0.027 (0.090)	0.037 (0.090)	0.028 (0.090)	0.029 (0.089)
β_3: Carry-over [T×E]	-0.121+ (0.072)	-0.121+ (0.072)	-0.116 (0.072)	-0.121+ (0.072)	-0.120+ (0.072)	-0.116 (0.076)	-0.128+ (0.073)	-0.121 (0.144)	-0.123 (0.144)	-0.130 (0.144)	-0.121 (0.144)	-0.129 (0.144)
β_6: Interaction[a] (T×Main effet[Z])		0.009 (0.035)	-0.151* (0.063)	0.001 (0.060)	-0.135 (0.139)	-0.000 (0.003)	0.138 (0.125)		0.040 (0.055)	-0.004 (0.055)	0.008 (0.055)	-0.203+ (0.109)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	0.005 (0.045)	✓	✓	✓
Baseline self-confidence[b]								✓	✓	0.110* (0.045)	✓	✓
Baseline external control[b]								✓	✓	✓	-0.053 (0.043)	✓
Parental education								✓	✓	✓	✓	-0.069 (0.090)
Students' ability[b]	✓	0.001 (0.031)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
First-year student	✓	✓	0.010 (0.051)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Female	✓	✓	✓	0.150** (0.049)	✓	✓	✓	✓	✓	✓	✓	✓
Has mobile phone	✓	✓	✓	✓	-0.044 (0.054)							
Day of message	✓	✓	✓	✓	✓	-0.006** (0.002)						
Exam difficulty	✓	✓	✓	✓	✓	✓	-0.177+ (0.098)	✓	✓	✓	✓	✓
Constant	9.206*** (0.973)	9.208*** (0.973)	9.170*** (0.973)	9.206*** (0.973)	9.155*** (0.978)	9.132*** (0.973)	9.213*** (0.973)	15.151*** (1.944)	15.265*** (1.948)	14.807*** (1.944)	15.190*** (1.944)	15.037*** (1.942)
Observations	8,301	8,301	8,301	8,301	8,301	8,301	8,301	2,016	2,016	2,016	2,016	2,016

N of students	6,916	6,916	6,916	6,916	6,916	6,916	6,916	1,592	1,592	1,592	1,592	1,592
Cohen's *d* effect size	0.066	0.066	0.098	0.065	0.15	0.065	0.049	0.035	0.035	0.038	0.034	0.114

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is s first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 5: Treatment effect on students' endline test anxiety, unstandardized regression coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T] *(treated = 1)*	-0.053 (0.075)	-0.053 (0.075)	0.003 (0.084)	-0.026 (0.100)	-0.387 (0.251)	0.014 (0.081)	-0.049 (0.087)	-0.161 (0.155)	-0.173 (0.142)	-0.202 (0.151)	-0.160 (0.155)	-0.320+ (0.194)
β_2: Exam [E] *(second = 1)*	-0.073 (0.084)	-0.073 (0.084)	-0.074 (0.084)	-0.072 (0.084)	-0.074 (0.084)	-0.171+ (0.089)	-0.073 (0.084)	-0.211 (0.177)	-0.279+ (0.164)	-0.291+ (0.173)	-0.209 (0.177)	-0.208 (0.177)
β_3: Carry-over [T×E]	0.092 (0.135)	0.093 (0.135)	0.099 (0.135)	0.092 (0.135)	0.092 (0.135)	0.175 (0.141)	0.094 (0.135)	0.352 (0.287)	0.334 (0.259)	0.425 (0.279)	0.349 (0.287)	0.349 (0.287)
β_6: Interaction[a] (T×Main effet[Z])		-0.013 (0.065)	-0.174 (0.114)	-0.045 (0.110)	0.351 (0.252)	-0.011* (0.005)	-0.023 (0.230)		-0.083 (0.103)	0.261* (0.106)	-0.076 (0.107)	0.293 (0.214)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	1.307*** (0.081)	✓	✓	✓
Baseline self-confidence[b]								✓	✓	-0.808*** (0.086)	✓	✓
Baseline external control[b]								✓	✓	✓	0.080 (0.086)	✓
Parental education								✓	✓	✓	✓	-0.223 (0.178)
Students' ability[b]	✓	-0.049 (0.057)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
First-year student	✓	✓	0.129 (0.094)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Female	✓	✓	✓	1.120*** (0.091)	✓	✓	✓	✓	✓	✓	✓	✓
Has mobile phone					-0.156 (0.200)							
Day of message						0.012** (0.004)						
Exam difficulty	✓	✓	✓	✓	✓	✓	1.080*** (0.181)	✓	✓	✓	✓	✓
Constant	7.854*** (1.611)	7.850*** (1.612)	7.824*** (1.611)	7.851*** (1.611)	8.028*** (1.621)	7.887*** (1.611)	7.852*** (1.611)	10.527** (3.501)	9.159** (3.172)	10.732** (3.404)	10.556** (3.503)	10.576** (3.502)
Observations	8,316	8,316	8,316	8,316	8,316	8,316	8,316	2,014	2,014	2,014	2,014	2,014

N of students	6,925	6,925	6,925	6,925	6,925	6,925	6,925	1,590	1,590	1,590	1,590	1,590
Cohen's *d* effect size	-0.018	-0.018	0.001	-0.009	-0.133	0.005	-0.017	-0.054	-0.058	-0.068	-0.054	-0.108

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table 6: Hypothesized treatment heterogeneity

Baseline variables [Z]	*The treatment effect is **higher** among students*	**Primary outcome**	**Secondary outcomes**		
		Grades	Self-efficacy	Motivation	Test anxiety
Test anxiety	*with high baseline test anxiety*	NS	**Supported**	NS	NS
Self-confidence	*with low self-confidence*	NS	**Supported**	NS	**Supported**
External control	*with external control*	NS	NS	NS	NS
Parental education	*whose parents do not have a university degree*	NS	NS	NS	NS
Students' ability	*with weaker baseline performance*	**The opposite is supported**	NS	NS	NS
First-year student	*among first-year students*	NS	NS	**The opposite is supported**	NS
Female	*among female students*	NS	NS	NS	NS
Has mobile phone	*receiving a text message on mobile phone*	NS	NS	NS	NS
Day of message	*who received the message later*	NS	NS	NS	**Supported**
Exam difficulty	*who take a difficult exam*	NS	NS	NS	NS

NS = Not significant

Appendix Tables

Table A1: Pairwise correlation between various psychological measures

	Baseline test anxiety	Endline test anxiety	Baseline self-confidence	Endline self-efficacy	Endline motivation
Endline test-anxiety	0.4769* (1,042)				
Baseline self-confidence	-0.4906* (2,305)	-0.2685* (1,042)			
Endline self-efficacy	-0.3505* (1,044)	-0.2248* (4,284)	0.4178* (1,044)		
Endline motivation	0.0148 (1,042)	0.2782* (4,283)	0.0794 (1,042)	0.2729* (4,289)	
External locus of control	0.1803* (2,305)	0.0781 (1,042)	-0.1796* (2,305)	-0.1542* (1,044)	-0.0584 (1,042)

Pairwise correlation coefficients, N of cases in parenthesis, * p<0.01
Baseline test anxiety is measured by using items from the Sarason (1980) test anxiety scale.
Baseline self-confidence is measured by using the academic subscale items from Shrauger and Schohn's (1995) Personal Evaluation Inventory.
External locus of control is measured by the four-item version of the test Rotter scale (Andrisani 1977; Goldsmith, Veum, and William 1996).
Endline measures are measured by single-item questions.

Table A2: Robustness check: main treatment effect deploying students' GPA in the last semester—instead of their admission scores—as baseline ability measure.

	Exam grades	Test anxiety	Self-efficacy	Motivation to do well on the exam
β_1: Treated [T] *(treated = 1)*	0.008 (0.020)	-0.047 (0.075)	0.295*** (0.065)	0.101* (0.041)
β_2: Exam [E] *(second = 1)*	-0.083*** (0.021)	-0.069 (0.084)	-0.190** (0.072)	0.002 (0.045)
β_3: Carry-over [T×E]	-0.024 (0.031)	0.085 (0.135)	-0.021 (0.116)	-0.123+ (0.072)
Constant	3.685*** (0.350)	7.644*** (1.611)	7.712*** (1.395)	9.163*** (0.973)
Observations Number of groups	28,156 15,264	8,316 6,925	8,296 6,908	8,301 6,916
Cohen's *d* effect size	0.01	-0.02	0.12	0.07

All models contain the following preregistered standard baseline control variables: student's gender, age, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.
Instead of the preregistered students' ability we control for student's GPA in the last semester. Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A3: Robustness check: main treatment effect with exam-subject fixed effects on the four endline outcome variables in the case of the first exam

	Exam grades	Test anxiety	Self-efficacy	Motivation to do well on the exam
Treated	0.003	-0.062	0.254**	0.088+
	(0.018)	(0.092)	(0.078)	(0.049)
Constant	3.866***	11.103*	3.578	9.802**
	(0.911)	(5.656)	(4.785)	(2.996)
N of students	14,673	5,325	5,310	5,318
Cohen's *d* effect size	0.00	-0.02	0.10	0.06

All models contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.
Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A4: Robustness check: main treatment effect with exam-subject fixed effects on the four endline outcome variables in the case of the second exam

	Exam grades	Test anxiety	Self-efficacy	Motivation to do well on the exam
Treated	-0.014	0.122	0.245*	0.005
	(0.020)	(0.136)	(0.116)	(0.077)
Constant	1.687	10.541***	11.373***	11.348***
	(1.711)	(2.632)	(2.267)	(1.537)
N of students	13,483	2,991	2,986	2,983
Cohen's *d* effect size	-0.01	0.04	0.09	0.00

All models contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.
Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A5: Robustness check: main treatment effect with student fixed effects on the four endline outcome variables

	Exam grades	Test anxiety	Self-efficacy	Motivation to do well on the exam
Treated	-0.010	0.001	0.227***	0.010
	(0.014)	(0.078)	(0.068)	(0.044)
Exam (second =1)	-0.010	-0.113	-0.030	-0.074
	(0.019)	(0.113)	(0.098)	(0.063)
Constant	104.030***	-112.185	151.131*	-19.269
	(13.934)	(74.519)	(65.040)	(41.851)
N of students	28,156	8,316	8,296	8,301
Cohen's *d* effect size	-0.01	0.00	0.09	0.01

All models contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.
Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A6.: Treatment effect on students' endline exam grades, among those who answered the endline questionnaire

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T]	0.063+	0.060+	0.056	0.077	0.095	0.060	0.014	0.070	0.067	0.076	0.068	-0.010
(treated = 1)	(0.034)	(0.034)	(0.039)	(0.047)	(0.126)	(0.037)	(0.040)	(0.069)	(0.068)	(0.068)	(0.069)	(0.090)
β_2: Exam [E]												
(second = 1)		0.077*	0.021	-0.024	-0.033	-0.000	0.267*		0.072	-0.027	0.031	0.143
		(0.032)	(0.057)	(0.054)	(0.127)	(0.002)	(0.111)		(0.053)	(0.053)	(0.053)	(0.106)
β_3: Carry-over [T×E]	-0.141***	-0.140***	-0.141***	-0.141***	-0.141***	-0.001	-0.134***	-0.107	-0.105	-0.091	-0.109	-0.105
	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)	(0.041)	(0.039)	(0.080)	(0.080)	(0.080)	(0.080)	(0.080)
β_6: Interaction[a]	-0.083	-0.084	-0.084	-0.083	-0.084	-0.070	-0.096	-0.023	-0.016	-0.038	-0.019	-0.022
(T×Main effet[Z])	(0.059)	(0.059)	(0.059)	(0.059)	(0.059)	(0.062)	(0.059)	(0.121)	(0.121)	(0.120)	(0.121)	(0.121)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	-0.155***	✓	✓	✓
									(0.039)			
Baseline self-confidence[b]								✓	✓	0.170***	✓	✓
										(0.039)		
Baseline external control[b]								✓	✓	✓	-0.054	✓
											(0.039)	
Parental education								✓	✓	✓	✓	-0.043
												(0.080)
Students' ability[b]	✓	0.222***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		(0.026)										
First-year student	✓	✓	-0.085*	✓	✓	✓	✓	✓	✓	✓	✓	✓
			(0.043)									
Female	✓	✓	✓	0.142***	✓	✓	✓	✓	✓	✓	✓	✓
				(0.042)								
Has mobile phone					0.058							
					(0.092)							
Day of message						-0.018***						
						(0.002)						
Exam difficulty	✓	✓	✓	✓	✓	✓	-1.904***	✓	✓	✓	✓	✓
							(0.086)					
Constant	3.140***	3.161***	3.142***	3.138***	3.090***	3.018***	3.166***	2.651+	2.771+	2.627+	2.666+	2.626+
	(0.724)	(0.724)	(0.724)	(0.724)	(0.729)	(0.717)	(0.724)	(1.474)	(1.466)	(1.461)	(1.474)	(1.474)
Observations	8,158	8,158	8,158	8,158	8,158	8,158	8,158	1,981	1,981	1,981	1,981	1,981
N of students	6,809	6,809	6,809	6,809	6,809	6,809	6,809	1,565	1,565	1,565	1,565	1,565

Cohen's *d* effect size	0.03	0.02	0.02	0.03	0.07	0.02	0.01	0.03	0.03	0.03	0.03	-0.00

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A7.: Treatment effect on students' endline exam grades, among students who have two endline exam-grades

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T]	0.005	0.005	0.000	0.027	-0.015	0.010	-0.010	0.043	0.045	0.047	0.046	0.020
(treated = 1)	(0.021)	(0.021)	(0.024)	(0.027)	(0.066)	(0.022)	(0.024)	(0.054)	(0.054)	(0.053)	(0.054)	(0.066)
β_2: Exam [E]												
(second = 1)		0.032+	0.016	-0.037	0.021	-0.001	0.088		0.001	-0.010	0.016	0.042
		(0.016)	(0.030)	(0.028)	(0.065)	(0.001)	(0.060)		(0.035)	(0.035)	(0.035)	(0.070)
β_3: Carry-over [T×E]												
	-0.079***	-0.079***	-0.079***	-0.078***	-0.079***	0.070**	-0.077***	-0.024	-0.022	-0.019	-0.020	-0.024
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.023)	(0.022)	(0.054)	(0.054)	(0.053)	(0.054)	(0.054)
β_6: Interaction[a]	-0.026	-0.026	-0.026	-0.026	-0.026	-0.018	-0.029	-0.073	-0.077	-0.079	-0.079	-0.071
(T×Main effet[Z])	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)	(0.033)	(0.081)	(0.081)	(0.080)	(0.081)	(0.081)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	-0.070*	✓	✓	✓
									(0.028)			
Baseline self-confidence[b]								✓	✓	0.167***	✓	✓
										(0.027)		
Baseline external control[b]								✓	✓	✓	-0.069*	✓
											(0.027)	
Parental education								✓	✓	✓	✓	0.019
												(0.056)
Students' ability[b]	✓	0.214***	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		(0.015)										
First-year student	✓	✓	-0.094***	✓	✓	✓	✓	✓	✓	✓	✓	✓
			(0.025)									
Female	✓	✓	✓	0.178***	✓	✓	✓	✓	✓	✓	✓	✓
				(0.024)								
Has mobile phone					-0.011							
					(0.051)							
Day of message						-0.020***						
						(0.001)						
Exam difficulty	✓	✓	✓	✓	✓	✓	-1.652***	✓	✓	✓	✓	✓
							(0.048)					
Constant	3.743***	3.748***	3.745***	3.732***	3.753***	3.564***	3.746***	2.244*	2.259*	2.292*	2.304*	2.197*
	(0.356)	(0.356)	(0.356)	(0.356)	(0.359)	(0.349)	(0.356)	(1.086)	(1.085)	(1.076)	(1.085)	(1.087)
Observations	26,981	26,981	26,981	26,981	26,981	26,981	26,981	4,218	4,218	4,218	4,218	4,218
N of students	14,089	14,089	14,089	14,089	14,089	14,089	14,089	2,178	2,178	2,178	2,178	2,178

Cohen's *d* effect size	0.00	0.00	0.00	0.01	-0.01	0.00	-0.00	0.02	0.02	0.02	0.02	0.01

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A8.: Treatment effect on students' endline test anxiety among students who answered the endline survey twice

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T]	-0.045	-0.044	0.017	-0.011	-0.325	0.022	-0.049	-0.130	-0.150	-0.181	-0.131	-0.273
(treated = 1)	(0.078)	(0.078)	(0.087)	(0.103)	(0.254)	(0.083)	(0.089)	(0.158)	(0.144)	(0.154)	(0.158)	(0.198)
β_2: Exam [E]												
(second = 1)		-0.009	-0.186	-0.055	0.295	-0.013*	0.021		-0.123	0.275*	-0.124	0.262
		(0.065)	(0.116)	(0.112)	(0.254)	(0.005)	(0.234)		(0.104)	(0.108)	(0.108)	(0.216)
β_3: Carry-over [T×E]												
	-0.057	-0.058	-0.058	-0.057	-0.059	-0.149	-0.057	-0.179	-0.254	-0.272	-0.177	-0.176
	(0.085)	(0.085)	(0.085)	(0.085)	(0.085)	(0.093)	(0.085)	(0.178)	(0.164)	(0.174)	(0.178)	(0.178)
β_6: Interaction[a]	0.086	0.086	0.090	0.085	0.085	0.203	0.085	0.312	0.320	0.406	0.310	0.310
(T×Main effet[Z])	(0.137)	(0.137)	(0.137)	(0.137)	(0.137)	(0.146)	(0.138)	(0.290)	(0.260)	(0.282)	(0.290)	(0.290)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	1.355***	✓	✓	✓
									(0.082)			
Baseline self-confidence[b]								✓	✓	-0.821***	✓	✓
										(0.087)		
Baseline external control[b]								✓	✓	✓	0.075	✓
											(0.087)	
Parental education								✓	✓	✓	✓	-0.211
												(0.181)
Students' ability[b]	✓	-0.052	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		(0.058)										
First-year student	✓	✓	0.171+	✓	✓	✓	✓	✓	✓	✓	✓	✓
			(0.096)									
Female	✓	✓	✓	1.151***	✓	✓	✓	✓	✓	✓	✓	✓
				(0.093)								
Has mobile phone					-0.164							
					(0.202)							
Day of message						0.010*						
						(0.004)						
Exam difficulty	✓	✓	✓	✓	✓	✓	1.070***	✓	✓	✓	✓	✓
							(0.185)					
Constant	7.842***	7.840***	7.809***	7.840***	8.014***	7.848***	7.844***	10.477**	9.097**	10.652**	10.585**	10.531**
	(1.611)	(1.611)	(1.611)	(1.611)	(1.621)	(1.610)	(1.611)	(3.497)	(3.154)	(3.396)	(3.499)	(3.499)
Observations	7,922	7,922	7,922	7,922	7,922	7,922	7,922	1,947	1,947	1,947	1,947	1,947
N of students	6,531	6,531	6,531	6,531	6,531	6,531	6,531	1,523	1,523	1,523	1,523	1,523

Cohen's *d* effect size	-0.02	-0.02	0.01	-0.00	-0.11	0.01	-0.02	-0.05	-0.06	-0.07	-0.05	-0.11

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A9.: Treatment effect on students' endline self-efficacy among students who answered the endline survey twice

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T] *(treated = 1)*	0.273*** (0.068)	0.275*** (0.068)	0.313*** (0.075)	0.260** (0.089)	0.558* (0.221)	0.255*** (0.072)	0.197* (0.077)	0.288* (0.136)	0.299* (0.130)	0.342** (0.126)	0.277* (0.135)	0.221 (0.169)
β_2: Exam [E] *(second = 1)*		-0.046 (0.057)	-0.121 (0.101)	0.023 (0.097)	-0.300 (0.221)	0.003 (0.005)	0.408* (0.204)		0.227* (0.090)	-0.175* (0.089)	0.112 (0.092)	0.129 (0.184)
β_3: Carry-over [T×E]	-0.214** (0.074)	-0.214** (0.074)	-0.214** (0.074)	-0.214** (0.074)	-0.212** (0.074)	0.001 (0.080)	-0.201** (0.074)	-0.103 (0.153)	-0.069 (0.148)	-0.009 (0.143)	-0.107 (0.152)	-0.099 (0.153)
β_6: Interaction[a] (T×Main effet[Z])	0.013 (0.118)	0.014 (0.118)	0.016 (0.118)	0.013 (0.118)	0.014 (0.118)	0.005 (0.126)	-0.009 (0.119)	-0.027 (0.249)	-0.024 (0.238)	-0.121 (0.229)	-0.014 (0.247)	-0.032 (0.249)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	-0.849*** (0.074)	✓	✓	✓
Baseline self-confidence[b]								✓	✓	1.090*** (0.071)	✓	✓
Baseline external control[b]								✓	✓	✓	-0.333*** (0.074)	✓
Parental education								✓	✓	✓	✓	-0.197 (0.155)
Students' ability[b]	✓	0.094+ (0.050)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
First-year student	✓	✓	0.150+ (0.083)	✓	✓	✓	✓	✓	✓	✓	✓	✓
Female	✓	✓	✓	-0.454*** (0.081)	✓	✓	✓	✓	✓	✓	✓	✓
Has mobile phone					0.184 (0.175)							
Day of message						-0.025*** (0.004)						
Exam difficulty	✓	✓	✓	✓	✓	✓	-1.639*** (0.160)	✓	✓	✓	✓	✓
Constant	7.600*** (1.401)	7.592*** (1.401)	7.585*** (1.401)	7.602*** (1.401)	7.413*** (1.409)	7.449*** (1.394)	7.634*** (1.400)	6.891* (2.984)	7.741** (2.848)	6.735* (2.740)	7.078* (2.964)	7.021* (2.986)
Observations	7,906	7,906	7,906	7,906	7,906	7,906	7,906	1,950	1,950	1,950	1,950	1,950
N of students	6,518	6,518	6,518	6,518	6,518	6,518	6,518	1,528	1,528	1,528	1,528	1,528

Cohen's *d* effect size	0.11	0.11	0.12	0.10	0.22	0.10	0.08	0.11	0.12	0.13	0.11	0.09

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A10.: Treatment effect on students' endline motivation to do well on the exam among students who answered the endline survey twice

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
β_1: Treated [T]	0.082+	0.082+	0.136**	0.072	0.198	0.084+	0.057	0.055	0.056	0.061	0.053	0.168+
(treated = 1)	(0.042)	(0.042)	(0.047)	(0.056)	(0.140)	(0.045)	(0.048)	(0.080)	(0.080)	(0.080)	(0.080)	(0.101)
β_2: Exam [E]												
(second = 1)		0.011	-0.162*	0.017	-0.122	-0.001	0.135		0.038	-0.008	0.002	-0.194+
		(0.036)	(0.064)	(0.061)	(0.140)	(0.003)	(0.128)		(0.055)	(0.056)	(0.055)	(0.111)
β_3: Carry-over [T×E]	-0.009	-0.009	-0.010	-0.009	-0.010	-0.009	-0.005	0.035	0.034	0.045	0.034	0.035
	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)	(0.046)	(0.091)	(0.091)	(0.091)	(0.091)	(0.091)
β_6: Interaction[a]	-0.097	-0.097	-0.094	-0.097	-0.096	-0.097	-0.104	-0.121	-0.122	-0.132	-0.119	-0.127
(T×Main effet[Z])	(0.074)	(0.074)	(0.074)	(0.074)	(0.074)	(0.074)	(0.074)	(0.147)	(0.147)	(0.146)	(0.147)	(0.147)
β_5:Main effects[Z]												
Baseline test anxiety[b]								✓	-0.006	✓	✓	✓
									(0.046)			
Baseline self-confidence[b]								✓	✓	0.117**	✓	✓
										(0.045)		
Baseline external control[b]								✓	✓	✓	-0.060	✓
											(0.044)	
Parental education								✓	✓	✓	✓	-0.072
												(0.092)
Students' ability[b]	✓	-0.006	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
		(0.031)										
First-year student	✓	✓	0.010	✓	✓	✓	✓	✓	✓	✓	✓	✓
			(0.052)									
Female	✓	✓	✓	0.145**	✓	✓	✓	✓	✓	✓	✓	✓
				(0.050)								
Has mobile phone					0.046							
					(0.109)							
Day of message						-0.006**						
						(0.002)						
Exam difficulty	✓	✓	✓	✓	✓	✓	-0.183+	✓	✓	✓	✓	✓
							(0.100)					
Constant	9.334***	9.336***	9.294***	9.336***	9.280***	9.265***	9.341***	14.952***	15.033***	14.593***	14.990***	14.861***
	(0.974)	(0.974)	(0.974)	(0.974)	(0.979)	(0.973)	(0.974)	(1.958)	(1.961)	(1.957)	(1.956)	(1.956)
Observations	7,910	7,910	7,910	7,910	7,910	7,910	7,910	1,950	1,950	1,950	1,950	1,950
N of students	6,525	6,525	6,525	6,525	6,525	6,525	6,525	1,526	1,526	1,526	1,526	1,526

Cohen's *d* effect size	0.03	0.03	0.05	0.03	0.13	0.03	0.02	0.02	0.02	0.02	0.02	0.07

All models (Column 1-12) contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.

The table lists those variables that we preregistered as a variable to test treatment heterogeneity (Z). Some of the standard control variables are listed in the table as they appear among variables in Z. We marked these variables with the ✓ sign indicating that the given variable was included in the regression even though its estimated coefficient was not included in the table.

In addition to the standard baseline variables, columns 8-12 contain the following preregistered additional baseline variables from the baseline survey, and thus they are available for a subset of students: baseline test anxiety, baseline self-confidence, baseline external control, and parental education. Since all of the additionally used control variables were preregistered as a variable to test treatment heterogeneity (Z), all of them are listed in the table and therefore marked with the ✓ sign.

[a] To enhance readability, the *Interaction* (T×Z) refers to the product of the treatment variable (T) and a specific main effect (Z). The coefficient of the corresponding main effect is shown in the table. For example, in Column 2, the interaction refers to the product of T×Students' ability, and in Column 10, the *Interaction* refers to the product of T× Baseline self-confidence.

[b] z-standardized variable at 0 mean and 1 standard deviation

Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Table A11.: The mediation analysis

	(1)	(2)	(3)	(4)	(5)	(6)
Treated	0.069*	0.068*	0.053	0.021	0.061+	0.055
	(0.035)	(0.035)	(0.035)	(0.034)	(0.035)	(0.035)
Exam (second =1)	-0.148***	-0.150***	-0.148***	-0.125**	-0.141***	-0.139***
	(0.039)	(0.039)	(0.039)	(0.038)	(0.039)	(0.039)
Carry-over	-0.084	-0.083	-0.073	-0.073	-0.084	-0.078
	(0.059)	(0.059)	(0.060)	(0.058)	(0.060)	(0.059)
Endline test anxiety		-0.010*				
		(0.005)				
Endline self-efficacy				0.108***		
				(0.006)		
Endline motivation						0.068***
						(0.009)
Constant	3.134***	3.208***	3.172***	2.358***	3.317***	2.710**
	(0.724)	(0.725)	(0.726)	(0.712)	(0.823)	(0.824)
Observations	8,017	8,017	7,998	7,998	8,002	8,002
N of students	6,708	6,708	6,694	6,694	6,699	6,699

All models contain the following preregistered standard baseline control variables: student's gender, age, ability, student is a first-year student, the type of training, the financial form of training, the level of training, the difficulty of the exam, and study program fixed effects.
Standard errors in parentheses, *** p<0.001, ** p<0.01, * p<0.05, + p<0.1

Appendix A: Students' perceptions of the intervention

Five months after our encouragement campaign (in May 2020), we invited all the students in our analytic sample to participate in an online survey about their experiences with our campaign. Approximately 16% of our population answered the survey (N = 2,420)[22].

Although five months is a significant amount of time and students' memories might be attenuated, 79% of the respondents correctly recalled the content of the message, while 9.5% of students claimed not to remember. The rest of the respondents either did not answer the question (6.5%) or recalled incorrect content (5%). These figures indicate that students' memories about the intervention had not attenuated significantly by the time of the follow-up survey.

The retrospective online survey provides us with qualitative information; in particular on the reception of the campaign, on possible adverse effects, and on the magnitude of treatment contamination.

First, students on average positively evaluated the encouragement campaign. On a five-point Likert scale, 65% of students answered that they were "happy" or "very happy" when they received the encouragement, and 77% of students stated that they would like to receive similar encouragement messages in the future.

Second, students who had not received the encouragement message might have been discouraged. On a five-point Likert-scale, 17% of students indicated that they were "sad" or "very sad" when we had asked them the question: "How sad were you when you found out that your peers had received the encouragement message, but you had not?" The discouragement of untreated students may lead to adverse treatment effects. It moves our estimations into the anticonservative direction, as while the treatment boosts the outcomes of treated students, the absence of treatment can worsen control students' outcomes. However, our qualitative data suggest that the magnitude of the adverse treatment effect might be moderate.

Third, approximately one-third of students may have been informed about the treatment before receiving it. In particular, 37% of students had heard that fellow students had received encouragement messages at a time when the students themselves had not yet received it, and 33% of students shared the message with their peers at the university after they had received it. Consequently, our treatment may have lost its novelty over time. Therefore, students treated later might have experienced a smaller treatment effect. Nevertheless, this makes our estimation more conservative.

It is important to note that these two facts (first that some students were informed about the treatment before being treated, and second that some students shared the treatment with their peers) should not be interpreted as indications of spillover effects. It is rather a form of treatment contamination. Our treatment is if someone actually received the encouragement message and not when students have just heard about the existence of the treatment. In this regard, prior information about the existence of the encouragement messages is less relevant.

[22] Ideally, the online follow-up survey should be administered earlier, immediately after the treatment. It was not feasible, however, due to the closures and switch to online education caused by the COVID-19 pandemic. These changes challenged the university's online platform and required the full attention of the administrative staff who could administer the infrastructure of such an online survey.

Appendix B: Subsamples

Column 1 in Table 3 shows the composition of the full target population regarding students' baseline characteristics. We gathered data about students with two questionnaires. Students voluntarily answered these questionnaires; therefore, the variables collected via the survey do not have full coverage but are only available for a subsample of students.

As Column 2 in Table 3 shows, almost half of the full sample (45%) filled in the endline survey (N = 7,026). Compared to those who did not fill in the endline survey (Column 4), the subsample of students that filled in the survey is more advantaged. It contains younger and more able students who are more likely to be enrolled in full-time and state-financed education, and the proportion of female students is also over-represented among them. Most importantly, however, the share of students who were allocated to Group A is the same in the two groups. Despite these differences, the treatment effect on students' endline exam grades among those who answered the endline questionnaire is qualitatively similar to the results shown in the paper (Table A4).

Those students who filled in the endline questionnaire twice (Column 5) make up a younger and more able subsample of students than those who filled in the endline survey once (Column 7). Those allocated to Group A have a 3 percentage-point higher likelihood of answering the endline survey twice. This difference is statistically significant at the 5% significance level.

The treatment status significantly decreased students' willingness to answer the endline questionnaire, both before students' first and second exams by 3.6 and 5.2 percentage points, respectively. As the e-mail that the control students received prompted them to go directly to the lottery, control students received stronger incentives to participate in the survey and win, which might explain why control students were more likely to fill in the endline survey. This type of selection could undermine the results on the secondary outcomes. Nevertheless, as Appendix Tables A8 to A10 show, the estimations were qualitatively similar among those who answered the endline questionnaire twice, and thus filled in the questionnaire in the treated and also in the control condition.

A small subsample of students (15%) answered the baseline background questionnaire (Column 8). The subsample of these students is more advantaged than those who did not fill in the baseline background questionnaire (Column 10). However, there is no difference between the two groups in terms of the allocation of students to Group A[23.]

In sum, the composition of subsamples differs in terms of the baseline covariates. Across different subsamples, however, the share of students allocated to different treatment groups is similar, except for those who filled in the endline survey twice versus once. In this case, however, the difference is quite small (3 percentage points), and it is below the preregistered threshold.

[23] We preregistered to show the results calculated in the subsample of baseline survey data and endline survey data only in the case when these subsamples are a random sample of students. We deviate from this decision in the analysis since (as preregistered) all our estimations controlled for the baseline data, based on which we detected the significant differences between students in different subsamples.

Differences between the composition of various sub-samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	All students	Endline Questionnaire			Has filled in Endline Questionnaire			Baseline Questionnaire on psychological variables (see Appendix D)		
		Filled in the	Did not fill in	Diff.	Twice	Once	Diff	Filled in	Did not fill in	Diff
Group A	0.50	0.50	0.50	0.00	0.53	0.50	0.03*	0.50	0.50	0.00
Has two exams	0.92	0.94	0.90	0.05*	1.00	0.93	0.07*	0.95	0.91	0.04*
Has two endline survey data	0.09	0.20	n.a.	n.a.	1.00	0.00		0.19	0.08	0.11*
Female	0.57	0.59	0.56	0.03*	0.60	0.59	0.01	0.64	0.56	0.07*
Age	24.12	23.94	24.26	-0.32*	24.06	23.91	0.15	23.64	24.20	-0.56*
Students' ability[a]	0.00	0.03	-0.02	0.05*	0.07	0.02	0.06*	0.02	0.00	0.02
Students' ability is missing	0.32	0.30	0.33	-0.02*	0.28	0.31	-0.03*	0.29	0.32	-0.03*
Full-time training	0.83	0.84	0.81	0.02*	0.84	0.84	-0.01	0.87	0.82	0.05*
State-financed training	0.75	0.77	0.73	0.04*	0.78	0.77	0.01	0.81	0.74	0.07*
Bachelor level	0.59	0.59	0.59	0.01	0.62	0.59	0.03*	0.56	0.59	-0.03*
Master level	0.09	0.09	0.10	-0.01*	0.08	0.09	-0.01	0.09	0.09	0.00
Undivided	0.24	0.25	0.24	0.01	0.23	0.25	-0.03*	0.29	0.24	0.06*
Higher-level vocational training	0.08	0.07	0.08	-0.01	0.08	0.07	0.01	0.05	0.08	-0.03*
First-year student	0.31	0.33	0.30	0.03*	0.36	0.32	0.04*	0.33	0.30	0.03*
Exam difficulty	0.19	0.20	0.19	0.01*	0.22	0.19	0.02*	0.18	0.19	-0.01*
Exam difficulty is missing	0.19	0.17	0.20	-0.03*	0.16	0.17	-0.01	0.17	0.19	-0.02*
Baseline test anxiety[a]								0.00		
Baseline self-confidence[a]								0.00		
Baseline external control[a]								0.00		
Parental education (university degree								0.54		
N	15,539	7,026	8,513		1,433	5,593		2,305	13,234	
%	100.00%	45.22%	54.78%		9.22%	35.99%		14.83%	85.17%	

* The difference is significant at 5% level using a two-tailed t-test.

[a] z-standardized variable at 0 mean and 1 standard deviation

Appendix C: Descriptive statistics of the outcome variables in the whole sample and in the subsample of those who answered the endline questionnaire

The 15,539 students in the analytical sample have a total of 28,156 exam grades, out of which 14,673 grades are from the first exams, and 13,483 grades from the second exams. There are 866 students (5.57%) who did not take the first exam (for whatever reason, most likely because they became ill). 2,056 students (13.23%) did not take a second exam since they had only one exam, or because they became ill before the second exam[24]. The allocation of students to Group A has no effect on missing endline grades in the first (coef. = -0.004; p = 0.232) and second exams (coef. = -0.009; p = 0.09).

There are 7,026 students who filled in the endline survey; 20.4% of them answered the endline questionnaire twice before their first and second exams as well. Thus, there are 8,459 valid answers to the endline questionnaire. The majority of answers were filled in before students' first exams (64%, N = 5,414) and about one third were filled in before students' second exams (36%, N = 3,045). The allocation of students to Group A has no effect on students' participation in the survey (coef. = -0.002; p = 0.764).

Descriptive statistics of the outcome variables in the whole sample, and in the subsample of those who answered the endline questionnaire

	Exam grades	**Exam grades**	**Test anxiety**	**Self-efficacy**	**Motivation**
Panel A: Both exams	*Whole sample N=31,078*	*The subsample of those who answered the endline questionnaire N=8,459, from 7,026 students*			
Mean	3.559	3.545	6.367	5.899	9.011
SD	1.454	1.438	2.913	2.526	1.532
N of non-missing	28,156	8,158	8,316	8,296	8,301
N of missing	2,922	301	143	163	158
Panel B: First exam	*Whole sample N=15,539*	*The subsample of those who answered the endline questionnaire N=5,414*			
Mean	3.636	3.648	6.373	6.000	9.032
SD	1.425	1.405	2.889	2.477	1.494
N of non-missing	14,673	5,226	5,325	5,310	5,318
N of missing	866	188	89	104	96
Panel C: Second exam	*Whole sample N=15,539*	*The subsample of those who answered the endline questionnaire N=3,045*			
Mean	3.476	3.362	6.356	5.719	8.973
SD	1.480	1.478	2.957	2.602	1.595
N of non-missing	13,483	2,932	2,991	2,986	2,983
N of missing	2,056	113	54	59	62

[24] We preregistered to not analyze students who only had one exam in the semester, and thus could not be treated before their second exam. In the analysis, we deviated from this decision since it would decrease the sample size of the endline survey, which many students answered either before the first or before the second exam. In the appendix, however, we show results for students who have two endline grades (Table A7) and who answered the endline survey twice before both of their exams (Table A8-A10). The results of the tables in the appendix are qualitatively similar to the results shown in the paper.

Appendix D: Control variables

We preregistered to deploy the following baseline variables to increase the precision of the estimation. The source of the baseline variables was the university's administrative registry information about students.

1. Students' *gender* can be either male or female. Females are coded as 1.
2. Students' *age* is the difference between the date of the exam and their date of birth divided by 365.
3. Students' *ability* is measured by students' high-school final examination test scores, which also serves as the admission score to tertiary education[25]. Our measure relates to the students tests scores in the subjects that a particular study program expected them to take. Thus, our measure of ability is specific to the particular study program and field of study. The ability variable is z-standardized at 0 mean and 1 standard deviation.[26]
4. *Type of training* is a dummy variable indicating whether a student is engaged in full-time training (=1) or either correspondence training or distance learning (=0).
5. The *financial form of training* is coded as state-financed (=1) or self-financed (=0).
6. The *level of training* is coded as bachelor level (=1), master level (=2), undivided[27] (=3), and higher-level vocational training[28] (4).
7. A dummy variable *first-year student* indicates (=1) whether students are freshmen or more long-standing students at the university (=0).
8. The *difficulty of the exam* is defined as the proportion of those who failed the respective exam in the last semester.
9. We employ *study program* (for example, sociology) fixed effects. Study programs are assigned to university faculties such as the Faculty of Social Science.

We preregistered to replace missing values in control variables with 0. Separate dummy variables control for the missing status data in each variable so as not to lose observations.

In addition to the registry data, we sent out a baseline background questionnaire to all students in the population. We inquired into psychological variables and parental background information in the questionnaire. In particular, we deployed the following questions in the baseline questionnaire:

---

[25] All students sit the same standardized tests (exams) in the subjects that they have chosen. There are subjects like mathematics or Hungarian, for which taking the high-school final examination is mandatory. However, some study programs at the university might require students to take the high-school final examination in particular subjects. For example, not all students have to take high school final examinations in biology, but students applying to become medical doctors have to take this exam.

[26] We assume that students' admission score is a better proxy for their abilities than their prior GPA. Students' GPA in prior semesters is the outcome of the university's grading standard. It is hard to compare, therefore, students' GPAs between different faculties and disciplines. By contrast, the test scores of high school final examination are nationally standardized tests and thus serve as a universal standard of comparison. Furthermore, students' GPA incorporates their motivations and effort. Thus, their prior GPA is also a response by students to those endogenous shocks that they experienced at the university (like how much students like their subjects/university). Student admission scores are, however, by definition, not an effect of those endogenous shocks experienced at the university. Results are robust, however, for controlling for students' prior GPA instead of their admission scores (Table A2 in the Appendix).

[27] Undivided programs are the kind of program commonly offered for medical doctors or lawyers, where there is no division between BA and upper levels of the program.

[28] Higher-level vocational training involves two years of training taught in higher educational institutions and results in a qualification similar to a bachelor's degree, although the qualification itself is not a degree.

1. We asked students about their *baseline test anxiety* by using items from the Sarason (1980) test anxiety scale. This is a z-standardized variable at 0 mean and 1 standard deviation.
2. We collected data about *baseline self-confidence* by applying the academic subscale items from Shrauger and Schohn's (1995) Personal Evaluation Inventory. This is a z-standardized variable at 0 mean and 1 standard deviation.
3. Locus of control measures the sense of agency people feel over their lives. Locus of control is believed to be conceptually similar to self-efficacy (Rotter, 1992) and is conceptually connected to behavioral intention and control in Ajzen's theory of planned behavior (Ajzen, 2002). We measured the baseline external/internal *locus of control* (Rotter 1966) using the four-item version of the Rotter-scale test (Andrisani, 1977; Goldsmith et al., 1996). In the test, respondents choose between two sentences describing external and internal control conditions. People with an internal locus of control believe that their abilities and actions influence their life outcomes. By contrast, people with an external locus of control believe that random chance and environmental factors affect their life outcomes. Throughout the analysis, we used a scale of external control in which we calculated the sum of those answers in which respondents opted for the external control options. We have a z-standardized variable at 0 mean and 1 standard deviation.
4. We asked about parental education by deploying separate seven-grade questions for the mother's and father's highest educational level. Parental education is coded as 1 if either the father or mother has a university degree.

Further decisions for coding of the baseline variables are specified in the pre-analysis plan.