Bhan, Prateek Chandra; Vornberger, Judith; Wen, Jinglin

Working Paper

Reflection and mental health: Experimental evidence from Germany

Working Paper Series, No. 38

Provided in Cooperation with:
University of Konstanz, Cluster of Excellence "The Politics of Inequality. Perceptions, Participation and Policies"

Suggested Citation: Bhan, Prateek Chandra; Vornberger, Judith; Wen, Jinglin (2024) : Reflection and mental health: Experimental evidence from Germany, Working Paper Series, No. 38, University of Konstanz, Cluster of Excellence "The Politics of Inequality", Konstanz, https://doi.org/10.48787/kops/352-2-1f340fv8qyhhu1

Working Paper Series
Nº 38 | August 2024

Universität Konstanz

Reflection and Mental Health: Experimental Evidence from Germany

Prateek Chandra Bhan, University of Konstanz, prateek.bhan@uni-konstanz.de
Judith Vornberger, University of Konstanz, judith.vornberger@uni-konstanz.de
Jinglin Wen, University of York, jinglin.wen@york.ac.uk

Working Paper Series of the Cluster "The Politics of Inequality":
→ http://inequality.uni.kn/working-papers

Cluster of Excellence
The Politics of Inequality

Konstanzer Online-Publikations-System (KOPS)
DOI: https://doi.org/10.48787/kops/352-2-1f340fv8qyhhu1

About the authors

Prateek Chandra Bhan is a postdoctoral researcher at the University of Konstanz, where he is a member of both the Cluster of Excellence “The Politics of Inequality" and the Department of Economics. He received his doctorate from the University of Glasgow and formerly studied in the University of Bristol and Delhi. His research focuses on the role of mental well-being, agency and hope in bettering the lives of children and youth, and its implications on development and inequality.

Judith Vornberger is a PhD student at the University of Konstanz, where she is a member of the Department of Economics. Her research focus lies on causes of mental health and the role played by institutions and policy measures. As part of this, she is also involved in topics related to the economics of education and experimental economics.

Jinglin Wen is a research fellow at the Centre for Health Economics of the University of York. He received his PhD in Economics from the University of Glasgow. His research interests include violence, development, and health inequalities. The context of his research has been the UK, the US, India and Germany.

Please cite as: Bhan, Prateek Chandra, Judith Vornberger and Jinglin Wen 2024, Reflection and mental health: Experimental evidence from Germany. No. 38, Cluster of Excellence “The Politics of Inequality”, University of Konstanz.

Reflection and Mental Health: Experimental Evidence from Germany

Prateek Chandra Bhan[1], Judith Vornberger[2], Jinglin Wen[3]

Abstract:[4] Despite increasing mental health problems and an existing care gap among university students, cost-effective solutions to bridge this gap are still lacking. Using a reflection intervention, we conduct a randomized controlled trial with undergraduate students in Germany. As part of a thought experiment, the treatment group reflected for ten minutes on questions related to stressors and their remedies. Combining survey and administrative data we find a significant improvement in students' mindful behavior, mental health and well-being as well as perseverance in performance. Our results show the self-empowering potential of a low-cost soft-touch intervention in students to aid their mindful behavior, mental health and well-being as well as performance and thus demonstrate one way universities as institutions can provide support.

[1] University of Konstanz, Konstanz, Germany
[2] University of Konstanz, Konstanz, Germany
[3] Centre for Health Economics, University of York, England
[4] We are grateful to the participants and the University for their support. We express our sincere gratitude to Christina Felfe, Maike Schlosser, and Theodore Koutmeridis for their invaluable support to this project since its inception. Special acknowledgment is due to Sabrina Gado, Erwin Winkler, and Celina Högn for their feedback and the participants at the 15th IHEA World Congress on Health Economics, at the 2nd Workshop on Field Experiments in Economics, and University of Konstanz for their useful comments during the seminar or conference presentations. We are incredibly thankful to Christina Felfe and the Graduate School of the Social and Behavioural Sciences (GSBS), Germany for funding support. IRB approval for the study was received from the University of Glasgow on 10-10-2022 with the approval number: 400220036. We received GDPR approval with the help of the data protection officer of the University of Würzburg. The study was registered as AEARCTR-0010277 (Bhan et al. 2022).

1 Introduction

Globally, one in eight, which amounts to 970 million people, suffer from mental health disorders (Institute of Health Metrics and Evaluation 2022; The Gallup Organization Ltd. 2021). Lifelong mental illness primarily sets in before, or at the age of entering college, or during the early years (Auerbach et al. 2016; Kessler et al. 2005). University students represent an especially vulnerable group with increasing mental health issues and simultaneously a persistent care-seeking gap (Auerbach et al. 2016; Ebert et al. 2019; Eisenberg et al. 2007; Hendrickx et al. 2020; Kessler et al. 2005; Storrie et al. 2010; The Oxford Student 2019; Watkins et al. 2012). This situation has been further exacerbated by the COVID-19 pandemic (Kaparounaki et al. 2020; Yonemoto and Kawashima 2023) calling urgent attention not only on the grounds of health and well-being but also due to its close association with human capital development and academic performance (Eisenberg et al. 2007; Hysenbegasi et al. 2005).

Poor mental health, particularly depressive symptoms, is associated with lower cognitive performance (Cornaglia et al. 2015; Eisenberg et al. 2009; Nafilyan et al. 2021). Growing evidence shows that interventions during the sensitive period of onset can assist in lowering the duration and severity of problems (McGorry et al. 2011). At the same time, insufficient provision, stigma and other factors cause a care gap in mental health service. This gap exists for almost 85% of students screened positive for mental health disorders (Ebert et al. 2019; Eisenberg et al. 2007; Hendrickx et al. 2020; Watkins et al. 2012). Addressing these problems requires us to increase the take-up, and likewise the supply, of mental health services.

Existing studies explore the potential of interventions to reduce stress levels and to improve mental health in university students (Cassar et al. 2022; Khoury et al. 2015; Ma et al. 2019). However, these interventions are costly (time and money), pose barriers to implementation or usage (need of specialist/therapist, user-friendliness, perils of stigma, burden of secrecy or shame), and mainly neglect the spillover effects on performance. Accordingly, an open call remains demanding action from universities to introduce sufficient services that aid student mental health and offer a conducive environment for self-help (Duffy et al. 2019; Eisenberg et al. 2007; Ochnik et al. 2021; The Oxford Student 2019). To cope with this challenge, in terms of provision and uptake barriers, we propose self-reflection as one option with bare-minimum resource costs to empower students to cope with these

issues. To the best of our knowledge, our study is the first using a very short soft-touch intervention that is easy to integrate in daily university life to answer this open call.

We conducted a randomized controlled trial (RCT) with undergraduate students at a German University. This involved introducing a one-shot reflection treatment in the form of a thought experiment. The treatment group was encouraged to reflect on aspects of stressors and how to manage them over the course of university life to improve student well-being. A placebo group received priming-free questions on their University architecture and how to improve these aspects. 185 students participated in this exercise as part of a general introductory "getting to know our students" online survey session. As part of this, we collected data on mindfulness as well as mental health and well-being and combined them with administrative data from the University. This provides us with a great and rather rare data set on mental well-being combined with performance data.

Through our intervention, we guide students to reflect. Reflection is the process, which helps individuals to review their beliefs and foster the creation of new knowledge. Rather than being spontaneous, self-reflection requires serious contemplation and information processing that can be attained via the deliberative use of prompts or questions to guide and elevate the focus of the reflective process (Boud et al. 2013; Coulson and Harvey 2013; Dyment and O'Connell 2011; Radović et al. 2023; Ryan 2013; Trede and Jackson 2021). As per Bandura (1993), people construct and rehearse anticipatory scenarios on the basis of their believes on themselves. Through reflection, directed action can be garnered utilizing such self-efficacy beliefs that act as an antecedent to reflective thinking (Bandura 1993; Czyżowska and Gurba 2021; Phan 2014).

In this study, we explore the effects of reflection on mindful behavior, mental health and well being as well as performance. For the former, we find that a 10-minutes reflection intervention in the beginning of the semester significantly increases mindfulness by 0.32 SD half-way through the term. This is driven by changes in student behavior with students reporting more frequent participation in sports, scheduling their exercises and exploring new ways to bring selfcare into life. This significant effect on more frequent sports activity remains until the end of the semester.

In terms of mental health and well-being, we observe no overall midterm effect on depressive symptomss, but the intervention decreases hopelessness and low energy levels. At the end of the semester, we find no effects on overall perceived stress within students.

Our intervention does not lead to better performance in terms of higher test scores of in-course tests or exams. Nonetheless, students in the treatment group show greater perseverance than their placebo counterparts. Conditional on having failed the main exam, students in the treatment group tend to pass the re-sit exam more often. This perseverance might be a result of higher hope and mindfulness in the treated students.

10-minutes of reflection during a lecture at the start of the semester results in a higher number of students passing the course after succumbing to failure with no detectable effect on their overall scores. This suggests to be more likely a result of psychological and behavioral changes rather than changes in cognitive skills, especially in the span of 6-months. This also hints towards the lasting effect of the intervention. All of these results withstand several robust checks such as variations in the estimation specification (e.g., an ANCOVA) and three different multiple hypothesis tests.

Current research provide evidence suggesting interventions to be beneficial for stress reduction and mental health improvements of university students. They examine interventions such as literacy interventions (Acampora et al. 2022) or mindfulness-based interventions (Khoury et al. 2015; Ma et al. 2019) that are e.g. premised on meditation (Cassar et al. 2022) or measures based majorly on technology or mobile applications (Lee and Jung 2018) like chat bots acting as cognitive-behavioral therapists (Fitzpatrick et al. 2017). One of these studies additionally examines the effects on performance concluding that mindfulness meditation can be valuable even though it takes time (Cassar et al. 2022). The literature on reflection indicates that (self-)reflection interventions build awareness, stabilize perceived stress, increase well-being, prevent the onset of mental illness and improve self-efficacy (Czyżowska and Gurba 2021; Falon et al. 2021; McCrindle and Christensen 1995).

Nevertheless, existing literature has its limitations. Mindfulness and reflection interventions, as in the case of Falon et al. (2021), Cassar et al. (2022) Czyżowska and Gurba (2021), and McCrindle and Christensen (1995) are intensive in terms of time commitment. Moreover, the assessment interval in Czyżowska and Gurba (2021) and McCrindle and Christensen (1995) has been rather short-term limiting the attention to immediate impacts and these evaluations rely on the treatment being repeated at several periodic occasions. Accordingly, these interventions may offer opportunities for addressing the care-seeking gap on the supply side. However, the take-up may increase only moderately due to persistent barriers.

Our contribution to current research is multi-dimensional. First, we introduce an easy to- integrate and administer, reflection intervention that can be simply adopted in universities and schools worldwide. While most studies (e.g. Cassar et al. 2022; Khoury et al. 2015; Ma et al. 2019) focus on mindfulness interventions, we take a step back and treat it as an outcome in itself of a reflective process. By doing so, we offer institutions the opportunity to respond to the constantly growing mental health problems among students with no need of specialists or therapists, increasing the supply of mental health support. We also provide a low-barrier approach potentially increasing the take-up of support. Combining these two aspects may translate into an overall narrowing of the care seeking gap. The simplicity of integrating the intervention further allows numerous ways in which the treatment can be replicated, modified, reproduced, scaled and evaluated independently or in combination with other intervention strategies.[5]

Second, we contribute to the narrow literature strand that links mindfulness and mental health data with administrative performance data. By doing so, we not only create a unique dataset, we also extend the existing literature by examining the impact of reflection on mental health and performance. Rather than analyzing the channels, we aim to propose reflection as an easy-to-implement strategy to improve students' mental health and well-being and in turn their performance.

Third, we expand existing research by considering a special cohort of university students. We conducted this intervention in the first lecture of an Economics course in the winter semester 2022. By timing the study at the beginning of the first semester back in full in-person teaching after remote mode during the COVID-19 pandemic, we strive to provide an opportunity for universities to support students after this period with increasing mental health issues (Kim et al. 2022; Li et al. 2021a; Li et al. 2021b). Against the backdrop of growing attention to mental health and well-being since the COVID-19 pandemic, we offer a warm glow for university administrators. It is a warm glow because they are providing a simple service creating positive changes and students may recognize that their University is making an effort.

Fourth, the stressors described by students in our thought experiment may reveal everyday problems in university life and thus offer a possible starting point for universities to react. Likewise, the remedies to reduce stress provided by the students offer potential new solutions for students and universities and therefore ideas for future interventions.

[5] A study already underway, for example, is the project "Student Agency, Institutional Fluency, and Diversity" (SAID) at the University of Konstanz.

Last, we introduce a highly cost-effective soft-touch intervention that can stimulate behavioral change. By its nature, such an intervention requires introspection and serious thought, which besides bearing no monetary costs are channels of self-empowerment. There could be multiplier effects from repeated acts of reflection, which would require future investigations. Through reflection, students find their own means to accept and navigate the course of university life. This agency is key to their empowerment and reflection plays a critical role in identifying and realizing it.

The remainder of this paper is structured as follows. In Section 2, we present our research design including the set-up and the interventio. Next, we introduce the data and methodology (Section 3). We further present the results in Section 4, discuss them in Section 5, and finally conclude with Section 6.

2 Research design

2.1 Set-up

The study was conducted in a cohort of predominantly second year undergraduate students in a public University in Germany. Following the free education format, there is no specific course fee but a minimal semester fee. The University follows the German academic calendar with the winter semester running from October to February including the main examinations in February and re-sit exams in April.[6] We conducted the study in the winter term 2022/23.

The study was introduced to all students enrolled in one Economics course, offered in the third semester[7] to students across all disciplines in Social Sciences and administered in English. The sample comprised of 185 students who offered their informed consent. The majority of these students come from Germany with a few international students from different parts of Europe, who attend their first in-person course since remote teaching during the COVID-19 pandemic.[8] Students enrolled in the course follow the standard protocols as per the University's guidelines with regards to attendance in the course and exams. Attendance in lectures or tutorials is non-binding with students having the freedom to take the course examination as many times as desired with prior registration. In case of

[6] The summer semester runs from April to July with examinations in August and September. However, the periods may vary slightly depending on the year.

[7] The study plan is not compulsory and students are allowed to repeat the course. Accordingly, 72.9% of the sample are in the third semester, but the remaining students are spread from the first to the ninth semester.

[8] The University regulations allowed return to campuses in summer 2022, followed primarily by in-person attendance in courses from the subsequent winter term.

failing the main exam in February, students are entitled to register for the similarly designed re-sit exam in April or take the exam one year later.

We incorporated the study design into the course structure (see Figure 2) and collected data in three waves each at the beginning of the lecture. All data collection rounds followed the same protocol (Bhan and Vornberger 2022). By entering the lecture hall as usual, students were assigned to a seat and a row alternating free. The seats were marked with two QR codes and web links for students without a QR code scanner. By scanning one of the QR codes or opening the link, students were given the choice between the English or German version of the questionnaire and were taken directly to the survey. The first pages informed them about the study and participation (Zizzo 2010). Beyond that, the research team announced the study using a predefined script. These information gave all necessary instructions to the respondents without revealing the actual intent of the study (Zizzo 2010). By giving consent, the students were guided to the questionnaire.

The course, and therefore also the data collection, took place in two lecture halls (one in-person and one live-broadcasted room). The principal investigator (in the in-person room) and the lead research assistant (in the live-broadcasted room) communicated via text messages during data collection for parallel execution. Several trained research assistants supported in both rooms, avoiding any interaction such as communication or mutual distraction.[9] Apart from ensuring compliance, these assistants did not help with individual queries. These were directed to the principal investigator/lead research assistant or were self-addressed as part of the instructions by the research team and within the surveys. The lecture started immediately afterwards further limiting student interactions.

The study was rolled out in the first lecture (in mid-October) with a baseline survey lasting approximately 20-25 minutes. This contained a battery of questions on students' background, mindful behavior as well as mental health and well-being (for details on the collected data see Table 3). Towards its end, the questionnaire included a reflection intervention described in more detail in Section 2.2.

Collecting data during this period is advantageous for gathering information on the state of mental health and well-being of freshly returning students at the start of the academic year. Moreover, these students are considered a special cohort who had only remote access to their first year of tertiary education. Bearing this in mind, the treatment was sensitively simplified to a reflective exercise of

[9] To ensure correct execution of the data collection, we provide the data collection protocol as a manual to the research assistants (see Bhan and Vornberger 2022).

10 minutes and the entire study was conducted as part of the regular university life to not expose students to any elements that may seem out of the ordinary.

After the baseline, a follow-up survey was conducted mid-way through the course in week 6 and a final follow-up in the last lecture in week 10. After this, lecture 11 was a doubt-clearing and an information session about the exam format. No additional content relevant for the exam was taught, making week 10 the final real lecture. We surveyed data on mindful behavior along with mental health and well-being in both follow-ups within approximately 10 minutes (for more details see Table 3).

The course included four voluntary in-course tests distributed evenly across the entire semester and performed in weeks 4, 6, 8, and 10. At the beginning of each respective lecture, students had 15 minutes time to respond to each of these tests consisting of five multiple-choice questions. Each test offered a score out of 25, with a cumulative score of 100. These tests in multiple-choice format captured student attendance and performance across the semester, besides offering a grade point jump of 0.3 for the passed exam for those attending all tests and achieving at least 50% of the total attainable scores. Similar to the data collection, a seating plan with empty seats and rows was implemented during the tests by placing QR code notes. Additional monitoring by trained research assistants avoided student interaction during the tests.[10]

A moderate incentivization scheme offered every participant, irrespective of their treatment assignment or any personal characteristics, a 10 (5) euro voucher[11] in the baseline (in the first follow-up) redeemable at a drugstore chain, while the vouchers were distributed using a lottery in the last follow-up. Doing so, we reward students for the time and effort they spend in completing the surveys.

2.2 Intervention

The intervention consists of a 10-minute reflection task, strategically and seamlessly introduced to students as part of a thought experiment. This thought experiment is similar for both groups in terms of script, instruction, administration, and response fields, while the subsequent questions are similar in wording but differ in topics. The design as well as its similarities and differences for both groups, described hereafter, are visualized and listed verbatim in Figure 1.

The fictitious scenario of the thought experiment, encourages the participant to imagine being invited to spend a year at the imaginary "Thomas Mann Excellence University". This scenario further

[10] More details are available in the data collection protocol (Bhan and Vornberger 2022).
[11] This amount is calculated and adjusted for the hourly minimum wage in the region.

suggests that the participant writes small articles for local newspapers or university journals as a hobby. In this made-up scenario, s/he is requested by the editor of the local press/student magazine to write on a certain topic. For this purpose, the respondent is asked to make notes on two questions. Therefore, each of these questions is accompanied by an open field. These carefully worded questions introduce the treatment and differ by group assignment. To randomize the students at an individual level into two groups (treatment and placebo), we use a function of the software Qualtrics that conducts a real-time randomization on individual level. The experiment is double-blinded, as neither the research team nor the students were aware of the group assignment (Bhide et al. 2018; Kendall 2003).[12]

The treatment group is asked two questions on the subject of stress. The first question ("Based on your experience as a student, what are the main stressors (factors that can cause stress) among college students?") concerns the main stressors among college students, while the second question ("Please carefully write down 7-8 ways (points) to deal with these stressors and improve student well-being.")[13] asks students how to deal with these stressors and improve student wellbeing. The first question invites the students to reflect on stressors arising from their own experience, whereas the second question implies gathering new information or giving advice based on individual experiences.[14]

In contrast to the treatment group, the placebo group receives, priming-free questions on architecture. Along with the introductory text, the first part of each question is identical to the treatment group (see Figure 1). The first question ("Based on your experience as a student, what are the important architectural aspects of a university building?") inquiries about important architectural aspects of the University building based on the respondent's experience. The second question ("Please carefully write down 7-8 ways (points) to improve these aspects.") invites them to write about ways that could improve these aspects.

Similar to the treatment group, we encourage the placebo group to reflect. However, the topic of reflection differs. We chose architecture of their University building, among many other possible topics, as the placebo questions to address an everyday topic. We thereby stimulate reflection but

[12] The experiment remains double-blinded until the end of data collection. We only merge the data after the last data collection by using a pseudo-anonymized student ID (created using the matriculation number).

[13] For consistency within this thesis, we write "well-being" here, whereas it was referred to as "wellbeing" in the intervention.

[14] We later record if the students sought some information from Google or any other source. Only 11 students (12%) of the treatment group used Google to answer their questions. Of these, 63.6% based less than half of their answers on a Google search. Thus, our intervention is primarily a reflection intervention and not a literacy intervention. Moreover, the usage of Google to answer the questions is balanced between the two groups (Difference: -0.025; $p = 0.584$)

avoid the risk of introducing a topic that can act as a primer. Naturally, some students may like the University building or its architecture more than others, or are more interested in architecture. This is taken into account by a balanced sample. Overall, we intentionally refer to the University building, to maintain a commonplace connection between the participant and the question, and to keep the question priming-free.

After completing these two questions, the students are directed to the last questions and subsequently to the final survey page including a thank you note for participating. Since the research team is able to monitor survey progress and participation in real-time, the lecturer is notified once the surveys are complete. At this point the lecture commences and students leave the room only at the lecture end.

We choose reflection as a treatment for various reasons. First, we expect reflection to have positive effects on mental health and performance. In psychology, self-reflection is considered a cognitive process that assists the reflector in sorting and evaluating experiences and facilitating behavioral changes (Anseel et al. 2015, 2009; Ellis et al. 2014; Falon et al. 2021; Yang et al. 2018). This can lead to positive changes in mental well-being and performance (Czyżowska and Gurba 2021; Ellis et al. 2014; Falon et al. 2021). Moreover, our treatment gives us insight into stressors and coping strategies of the students through their answers, allowing us to gain insights for future interventions. As a further aspect, our intervention has lower barriers for take-up than other interventions and is easier to integrate into everyday life. One factor here is that, in contrast to other mindfulness-based (e.g., Cassar et al. 2022) or reflection (e.g., Czyżowska and Gurba 2021; Falon et al. 2021) interventions, our intervention has lower costs, in terms of time and money, and requires neither specialists nor therapists. In addition, we integrate the intervention into the students' everyday life (during the lecture) reducing the costs and barriers to accept help through an everyday feeling. We further deliberately choose the third person narration and do only slightly target the reflector themselves. By doing so, we intend to encourage the reflector to give advise (Eskreis-Winkler et al. 2018) and to reduce the barrier of talking about problems.

With the timing of the intervention at the beginning of the semester, we decide not to consider the participants in an heightened stress phase (in contrast to e.g., Falon et al. (2021)). In addition, we decide against exposing the participants to certain stressors before the reflection (in contrast to e.g., Lepore et al. (2004)) and to only softly guide the reflection. All of these three aspects allow easy

integration into everyday life, reduce the risk of backfiring effects and measure the effects in daily life instead of in extraordinary situations. In this way, and because self-reflection is a personal choice, we intend to leave room for thoughts that participants may wish not to disclose, but can reflect on during and after the intervention.

There are further reasons why the design of our intervention keeps the risk of backfiring effects at a negligible level. We let the students reflect in the form of a hypothetical scenario with a fictitious thought experiment as active, impartial spectators, thereby further triggering the positive effects of counseling (Eskreis-Winkler et al. 2019). We assume that the risk of them remembering stressful events when reflecting on and answering the two questions is minor. We assume that in particular the measured increased hope and the positive, future-oriented questioning of giving advise how to deal with stressors act as a buffer for possible negative effects on mental health due to repeated rumination (Eskreis-Winkler et al. 2019; Geiger and Kwon 2010).[15]

Overall, we opted for this intervention as we conside reflection as a beneficial tool that is easy to integrate into everyday life.

[15] Interventions aimed at behavioral changes run the risk of backfiring (Stibe and Cugelman 2016). Reflection on stressors can help to sort thoughts (Clark 1993; Meichenbaum and Fitzpatrick 1993) and improve mental health (Smyth 1998). Psychologists also encourage reflection in therapies by asking questions (Tomm 1988). In the context of reflection about stressors, backfiring may be triggered by not feeling understood by the interlocutor (Lepore et al. 2004; Wortman and Lehman 1985) or by repeated rumination (Geiger and Kwon 2010). The latter can lead to increased depressive symptoms, especially in people already suffering depressive symptoms (Starr 2015). We expect low to no backfiring effects. First, we let the students reflect in the form of a hypothetical scenario with a fictitious thought experiment as active, impartial spectators, thereby further triggering the positive effects of giving advice (Eskreis-Winkler et al. 2019). Second there is further no risk of negative effects from an interlocutor due to the absence of an interlocutor within our intervention. Moreover, high hope acts as a buffer for the negative effect of rumination on depressive symptoms (Geiger and Kwon 2010). We consciously focus our reflection on how to deal with stressors and observe increased hope in our results.

Figure 1: Intervention

Thought experiment

You can take 10 minutes to respond to this task. Feel free to reflect and search online for any relevant information that you may need to complete it. Read the passage belowand answer the two questions that follow.

"The Thomas Mann Excellence University is a public research university. It is one of the oldest institutions of higher learning, having been founded in 1400s. The university initially had a brief run as a small institution and extended widely in 1581 on the initiative of some of the finest researchers at that time. Today, the university is renowned for high educational standards, strict entrance requirements and a strong international research reputation. This has made the university a competitive destination for students worldwide. It is a World Top 100 university that constantly strives for excellence. You were invited to visit were invited to visit the university and spend a year there. You have a hobby to write small pieces in the local newspaper or the university magazine. Upon arrival to The Thomas Mann Excellence University, you were invited by the editor of the local press/student magazine to write on a specific topic. The topic required you to offer some notes on the following two questions."

Question 1: Based on your experience as a student, what are the...

Treatment → ... main stressors (factors that can cause stress) among college students?

Placebo → ... important architectural aspects of a university building?

Question 2: Please carefully write down 7-8 ways (points) to...

Treatment → ... deal with these stressors and improve student well-being.

Placebo → ... improve these aspects.

Notes: In this figure, we zoom in to the first lecture and present the structure and verbatim text of the intervention for both groups. The thought experiment and the respective two questions appear on the screen consecutively by clicking the "next" button.

3 Data and methodology

3.1 Data

Our data are derived from two different sources: Survey data and administrative data. In the former, we collect data from three categories (i) socio-economic characteristics, (ii) mindful behavior (mindfulness and conscious activities) as well as (iii) mental health and well-being (depressive symptoms, perceived stress, life satisfaction, and risky behavior). These data are captured within three waves consisting of baseline and two follow-ups. The questionnaires are provided in English or German depending on the student's choice. The administrative data include information on attendance, performance, and perserverance measured during and at the end of the semester (we refer to Table 3 for a summary of the collected variables and its time of collection).

In the baseline, we ask questions about socio-economic characteristics such as gender, age, working situation, and the number of friends in the course. In each survey wave, we further collect the matriculation number, which serves as an student identifier (anonymized matriculation number) to link all data sets.

Mindful behavior comprises information on mindfulness and conscious activities. To collect both, we use a series of questions on the frequency of healthy and mindful behavior (e.g., sports, scheduling exercise, and selfcare) during the past week based on the validated Mindful Self-Care Scale (MSCS) (Cook-Cottone and Guyker 2018; Hotchkiss and Cook-Cottone 2019). We measure mindfulness using 12 items[16] in the first follow-up, which is our first main outcome variable.[17] Limited by time constraints, we shortened this scale in the baseline and the second follow-up to capture at least a subset of the information. We refer to this secondary variable with conscious activities.

Mental health and well-being include information on depressive symptoms, perceived stress, life satisfaction, and risky behavior. To record this information, we rely on established scales. We screen depressive symptoms via the validated Patient-Health-Questionnaire (PHQ-8) (Gräfe et al. 2004; Kroenke et al. 2009; Löwe et al. 2002). Students' perceived stress is measured based on the

[16] We shortened the 33-item MSCS and adapted it to the university context bearing in mind the limited time for surveying the students. We refer to the data collection protocol for more detailed explanations (Bhan and Vornberger 2022).

[17] We expand the outcome variables mentioned in the pre-analysis plan (Bhan et al. 2022) by the mindfulness scale, as it is of great interest, especially as a possible mediator for all other outcomes. Since we cannot disentangle the role as a mediator, we keep it as an outcome.

validated Perceived Stress Scale (PSS) (Cohen et al. 1994; Schneider et al. 2020). Beyond that, we study life satisfaction using the SOEP (Socio-Economic Panel) 11pt question for general life satisfaction (Berlin, DIW and others 2022; Richter et al. 2017)[18] and pose two single questions of Busch et al. (2014) on smoking and drinking to gauge risky behavior. Data on depressive symptoms are collected in all waves, while the remaining four variables are recorded only in the baseline and the second follow-up. We use depressive symptoms and perceived stress as the two main outcome variables within the mental health and well-being category, as these are particularly prevalent among university students (American College Health Association 2022; Ibrahim et al. 2013). We classify the remaining single-item variables as secondary.

By merging the first data set with the administrative data, we expand our data with attendance, performance, and perserverance. We measure the first two variables in four in-course tests across the semester and the main and re-sit exam at the end of the semester.[19] Using the performance data of both exams, we measure students' perseverance. This covers passing the second attempt (re-sit exam) after failing the main exam (for details on this variable, see Equation (2)). Focusing on performance, we consider the performance data (performance and perseverance) and add attendance as a supplement.

We are aware that we encounter a selectivity problem regarding performance and attendance in the exam and test data. The composition of the re-sit exam sample depends on the outcome of the main exam. This can either be strategic (e.g. to gain more time for preparation)[20] or non-strategic (failing the first exam). The situation is similar with the test data. These data are successively interdependent, as the grade bonus and thus the motivation for the voluntary tests depends on participation in all tests and achieving at least 50% of the total score. The respective samples are subject to selection after the first stage (main exam or first test). Accordingly, we refrain from examining the data of the re-sit exam (or the second to fourth test) in isolation and only analyze them in combination with the main exam (or first test of the series).

Table 1 presents statistical balance across the two groups for baseline data including baseline descriptive statistics. The overall sample consists of 185 students. After individual randomization,

[18] We follow Gesis 2023 to recode the SOEP 11pt question.

[19] An overview of the collected indicators can be found in Table 3. Further information is available in the data collection protocol (Bhan and Vornberger 2022)

[20] This option exists as students in this degree program at this University can attempt the exams as many times as they wish until they pass the exam.

the treatment and placebo groups contain 91 and 94 students, respectively. The majority of the participants are native (German) students aged on average almost 22 years. More than half of the sample are male and engaged in part-time employment whilst studying. Most students are enrolled in one Bachelor program and more than 60% have a religious affiliation and parents with high degree of education. Almost no student reports suffering from a bad financial situation, with every student having at least one close friend as part of the same course. The sample is overall balanced and comparable across the groups for baseline data. However, the groups differ significantly from each other in two of 22 variables at a 10% level, which was to be expected as a statistical artifact. However, we consider these differences in our robustness checks (see Section 5).

185 students participated in the baseline, while the sample comprises 151 in the first and 136 in the second follow-up. Within the merged sample, 169 are registered for the main exam with 35 re-registered for the re-sit exam after failling. 170 of the 185 students participated in the first in-course test, while 160 took part in the second test with 154 remaining in the third and 152 in the last test. Our incentive scheme is designed to avoid possible attrition and Table 5 acknowledges low attrition rates along with remaining balanced samples.

We present initial levels of depressive symptoms (Panel A) and perceived stress (Panel B) in Figure 3 using the baseline data and cut-offs developed by Kroenke et al. (2009) (for depressive symptoms) and Department of Administrative Services (2023) (for perceived stress levels).[21] At the beginning of the semester, nearly 30% of the sample already suffered from moderate to severe depressive symptoms and only 25% of the students report no depressive symptoms. Along with this, only 22% of our sample report low levels of perceived stress compared to 78% experiencing moderate to high levels.

By categorizing and analyzing the thought experiment answers, we identify exam pressure, workload management or the lack of time, and problems with social contacts as the primary stressors during university (see Figure 4 Panel A). For instance, one of the students answered *"competitive environment ... forces to study and sacrifice free time. Often students prepare just for an exam to pass the subject, instead of studying the subject."* Others mentioned the pressure of *"Not living up to expectations"* or *"To be as Perfect as you could be"* as the main stressors.[22]

[21] The Perceived Stress Scale is no diagnostic instrument, therefore the creators did not set any cut-off values. They only evaluate the perceived stress by stating that higher values represent more stress (Cohen et al. 1994; Klein et al. 2016). We use the cut-off values developed by Department of Administrative Services (2023) to categorize and improve the evaluation of students' initial stress levels.

[22] Some other student responses included *"not feeling worthy enough based on certain grades"* or *"Exams"* or *"Bad grades after having studied a good amount, Family responsibilities, Health..."* as some of the key reasons for stress.

Table 1: Baseline balancing table

	Pooled	Placebo		Treatment			
	Mean	N	Mean	N	Mean	Diff	P-Value
Female	0.438	94	0.447	91	0.429	0.018	0.804
Age in years	21.834	94	21.738	91	21.933	-0.195	0.712
Native German	0.859	94	0.862	91	0.857	0.005	0.929
Father educated in Germay	1.195	94	1.160	91	1.231	-0.071	0.254
Mother educated in Germany	1.270	94	1.255	91	1.286	-0.030	0.644
High parental education	0.692	94	0.734	91	0.648	0.086	0.209
Religious affiliation	0.741	94	0.777	91	0.703	0.073	0.258
Practice religion actively	0.234	73	0.192	64	0.281	-0.089	0.220
More than one Bachelor	0.119	94	0.149	91	0.088	0.061	0.202
Working while studying	0.573	94	0.596	91	0.549	0.046	0.527
Working hours per week	11.830	56	11.661	50	12.020	-0.359	0.799
(Close) friends in this course	1.848	94	1.957	90	1.733	0.224	0.420
Bad financial situation	0.092	94	0.096	91	0.088	0.008	0.855
Duration to fill out survey (in min.)	24.423	94	24.640	91	24.200	0.439	0.761
Using Google to answer	0.108	94	0.096	91	0.121	-0.025	0.584
Percentage of using Google	2.250	9	1.889	11	2.545	-0.657	0.371
Drinking alcohol	3.324	94	3.245	91	3.407	-0.162	0.789
Smoking cigarettes	3.784	94	2.862	91	4.736	-1.875	0.131
Conscious activities	3.486	94	3.402	91	3.574	-0.173*	0.060
Life satisfaction	6.730	94	6.542	91	6.933	-0.391*	0.062
Perceived stress (PSS)	18.638	94	18.777	91	18.495	0.282	0.751
Depressive symptoms (PHQ-8)	7.562	94	7.628	91	7.495	0.133	0.833
Overall number of observations	185						

Notes: The (baseline) descriptive statistics include number of observations (N) and means (Mean) for the pooled sample as well as treatment and placebo group separately. The mean differences (Diff) of both groups is listed in column (6). The p-value corresponds to p-values of the test under the null hypothesis of the equality of means between the groups. The overall sample includes 185 observations, with fewer observations for the variables practice religion actively (N=137), working hours per week (N=106), and percentage of using Google (N=20). Significance levels at the 10%, 5%, and 1% are indicated by *, **, and ***, respectively.

Beyond that, students mainly mention aspects they can implement themselves as remedies to cope with these stressors and improve well-being (see Figure 4, Panel B). They most frequently propose spending time with contact people, time management, food and rest, as well as better studying habits or doing sport as useful strategies. They wrote for example *"exercise and talking to friends"*, *"talk to others, take time to do something fun, go to the gym/ do sport....thinking about the reasons to keep going, listening to music.."* and *"enough sleep, meditation, balanced diet, sufficient vitamin intake"* to be relievers of stress.[23] However, a few suggestions are directed towards the university as an institution,

[23] Other responses included *"Time management, Goal setting"* and *"Study in groups, plan the week, take notes"*, among several others.

providing students with a supportive environment. Here, the most often addressed proposal category is restructuring teaching and reorganizing the university. Beyond that, they even explicitly desire support, especially for new students.

3.2 Methodology

The design of our experiment is in accordance with the RCT standards. The experiment is double-blind and we randomize at an individual level in real-time. The randomization is successfully processed with Qualtrics generating a balanced sample of the two groups (see Table 1). By avoiding interactions, accomplished by a seating plan and research assistants walking around, we ensure compliance. By doing so, we also reduce spillover effects during implementation. However, we cannot control interactions that occur in the phases between data collections outside the course room. Referring to Duflo et al. (2006), we therefore incorporate the number of close friends within the course in our main specification (Equation (1)) as a measure of connectivity to account for potential spillover effects.

Given the design of our study, we can use the following OLS regression to identify the effect of treatment on several outcome variables:

$$Y_i = \alpha_0 + \alpha_1 Treatment_i + \alpha_2 X_i^{'} + \epsilon_i \tag{1}$$

where, Y_i is the outcome of interest, $Treatment_i$ represents the dummy for individual treatment (=1 if student i belongs to the treatment group), and α_1 is the coefficient of interest capturing the effect of treatment. We further add a vector of covariates $X_i^{'}$ collected in the baseline for student i. These contain age, number of friends as well as dummies for gender, native (German), high parental education, bad financial situation, enrolled in more than one Bachelor's degree, working while studying, and religious affiliation.[24],[25] We include robust standard errors by ϵ_i (Abadie et al. 2023; White 1980) to obtain heteroscedasticity-consistent estimators.[26] We standardize all outcome variables except dummy variables to mean 0 and standard deviation (SD) 1 to simplify effect comparability.

[24] We rename $X_{io}^{'}$ of the pre-analysis plan (Bhan et al. 2022) in $X_i^{'}$ for consistency within this paper.

[25] In spite of having such information, we are limited in our capacity to conduct any heterogeneity analysis due to the sample size. We hereby deviate from the pre-analysis plan (Bhan et al. 2022).

[26] We hereby take a standard approach to deal with the common problem of homoscedasticity (Cunningham 2021). Robust standard errors are conservative. Clustered standard errors are not applicable to solve the issue since randomization is done on unit level (Abadie et al. 2023).

Beyond that, we use the data of both exams to measure *perseverance*. To do so, we interact the dummy variables *Apr* and *Feb*, which are coded inversely. *Feb* measures failing the main exam (= 1, if student failed the main exam in February) whereas *Apr* indicates passing the re-sit exam (=1, if student passed the re-sit exam in April). This creates the variable *perseverance* listed in Equation (2), which is 1 if the student managed to pass in the second exam conditional on having failed the first, otherwise it is zero.

$$Perseverance = Feb * Apr = \begin{cases} 1 & \text{if } Feb = 1 \text{ and } Apr = 1 \\ 0 & \text{if } Feb = 0 \text{ and/or } Apr = 0 \end{cases} \tag{2}$$

4 Results

Using Equation (1), we investigate the effects of our reflection intervention mid-way through the term or at the end of the semester. Table 2 presents our main results of mindful behavior (Panel A), mental health and well-being (Panel B) as well as performance (Panel C). Since we examine the effect of an extremely short soft-touch intervention, we expand the analysis of the scales by considering their individual items or subscales.

First, we detect a 0.32 SD midterm increase in mindfulness significant at a 5% level, which is driven mainly by engaging in sports, scheduling exercises, and exploring new ways of self-care (see Table 2, Panel A). The effect on doing sports (0.33 SD) lasts until the end of the semester, while we did not collect data on the other two items at the end of the term (Table 4, Panel A).[27]

Considering mental health and well-being, we do not identify an overall treatment effect on depressive symptoms mid-way through the term, but a significant reduction in hopelessness by 0.26 SD and low energy by 0.27 SD (Table 2, Panel B.i). We further find no effects on perceived stress or its two main sub-scales[28] at the end of the term (Table 2, Panel B.ii). Likewise, the reflection intervention does not significantly affect depressive symptoms, life satisfaction, or engagement in risky behaviors at the end of the semester (Table 4, Panel B).

[27] We collected certain aspects of lifestyle to create the mindfulness scale. However, constrained with time, we limited ourselves to only four measures of the mindfulness scale in the final follow-up. The effects on these items are depicted in Table 4.

[28] We created both sub-scales following Taylor (2015).

Overall, we do not observe improvements in students' performance. First, we do not detect any effects on cumulative test or exam scores (Table 2, Panel C.i and C.ii), or going back one step, also no effect on effort in terms of attendance of in-course tests and exams (Table 4, Panel C). The same applies to attendance and performance in the respective first stage and thus the first test and the main examination (Table 2, Panel C.i and C.ii, Table 4, Panel C). However, the sign of the performance coefficients in the first stage and the aggregated outcomes indicate as in Cassar et al. (2022) a negative but not yet significant effect on performance 4-6 months after the intervention. Nevertheless, perseverance significantly increases by 6 percentage points for the treated group. In other words, conditional on having initially failed more students pass the re-sit exam. This is suggestive of a higher degree of acceptance of failure and persevering to try again among the treated students.

To interpret our results, we summarize that we measure an effect on mindfulness that is accompanied by increments in components of depressive symptoms as well as decrements in student exhaustion. Treated students reported scheduling and engaging in sports and physical exercise more than the placebo group. Such a behavior has further potential to translate into improved general and mental health (Chekroud et al. 2018; Yao et al. 2022). Alongside, we detect an increase in student performance in terms of passing the re-sit exams after having been unsuccessful in the first attempt. Having failed, the treated students not only tried again but they succeeded more often. In our study, we cannot isolate the effect of increased mindful behavior as well as improved mental health and well-being on performance, given that all three components are affected by the intervention. However, based on literature, we attribute the psychological and behavioral implications of reflection to have contributed as a mechanism to such an improvement in performance (Cassar et al. 2022; Vorontsova-Wenger et al. 2021).[29]

Our results indicate that reflection can aid mental health symptoms and perseverance and foster mindful behavior that improves it. These results are in line with existing literature regarding mental health and well-being. In a similar sample of German undergraduate students, Cassar et al. (2022) find that mindfulness training[30] improves students' mental health and non-cognitive skills. Likewise, further mindfulness-based interventions (Khoury et al. 2015; Ma et al. 2019) and reflection interven-

[29] Despite our goal formulated in the pre-analysis plan (Bhan et al. 2022) to assess the effect of mental health on performance, it is only possible to measure the treatment effects individually and it is only possible to assume a relationship based on existing literature.

[30] They introduce a mindfulness course that comprise 8 weekly 60-minute group sessions and at the same time all the treatment group participants receive an audio recording and a hand-out and they were encouraged to do specific exercises once a day.

tions (Czyżowska and Gurba 2021; Falon et al. 2021), even if not explicitly investigating students, confirm positive effects on mental health and well-being. In terms of stress, the literature is controversial. Falon et al. (2021) suggests that reflection can stabilize stress levels. According to Khoury et al. (2015), mindfulness-based interventions lead to stress reduction, whereas Lee and Jung (2018) found no effects. Similarly, we find no effects on perceived stress.

Alongside negative association between mindfulness and depressive symptoms, Vorontsova-Wenger et al. (2021) investigate a positive association between the former and exam performance in the past. In terms of performance, McCrindle and Christensen (1995) identify reflection about learning behavior as beneficial. Alan et al. (2019) present a teacher-training program on grit (i.e. persistence in working towards a goal, and conscientiousness) underlying topics. This intervention leads to more persistence even after failing, setting higher goals, and higher success rates to achieve them even in the long-run. Cassar et al. (2022) report negative short-term effects on performance (grades). Only in the long term, they identify positive effects on performance (grades). Similarly, our effects on performance (score) are negative (although not yet significant) in the short term. Accordingly, our intervention coupled with the increased perseverance (as in the study of Alan et al. (2019)) might lead to positive effects on performance (scores) in the long term.

Table 2: Main results

Panel A: Mindful behavior (midterm)													
	Mind-fulness	Nut-rition	Sport	Sched. exerc.	Liked things	Listen to me	Selecting thoughts	Accept failure	Purpose	Special people	Rest	Relax	Selfcare
Treatment	**0.321****	-0.079	0.452***	0.385**	0.023	0.230	0.128	0.032	0.208	0.165	-0.055	0.053	0.336**
	(0.154)	(0.164)	(0.151)	(0.158)	(0.159)	(0.161)	(0.163)	(0.162)	(0.158)	(0.168)	(0.168)	(0.164)	(0.163)
N	**151**	151	151	151	151	151	151	151	151	151	151	151	151
Panel B: Mental health and well-being													
Panel B.i: Depressive symptoms (midterm)													
	PHQ-8	Little pleasure	Hopeless	Sleeping problems	Little energy	Poor/over-eating	Feel failure	Low concentr.	Speaking issues				
Treatment	**-0.178**	-0.159	-0.260*	-0.085	-0.271*	-0.130	-0.211	0.179	0.078				
	(0.153)	(0.162)	(0.156)	(0.163)	(0.160)	(0.159)	(0.153)	(0.169)	(0.162)				
N	**151**	151	151	151	151	151	151	151	151				
Panel B.ii: Perceived stress (end of term)													
	PSS		Perceived helplessness			Lack of self-efficacy							
Treatment	**-0.117**		-0.241			0.140							
	(0.173)		(0.176)			(0.171)							
N	**136**		136			136							

Continued on next page

Table 2 (contd.): Main results

Panel C: Performance			
Panel C.i: Performance in-course tests (during the term)			
	First in-course test	All in-course tests	
Treatment	-0.102	-0.148	
	(0.167)	(0.159)	
N	142	159	
Panel C.ii: Performance exams (end of term)			
	Main exam	Both exams	Perseverance
Treatment	-0.230	-0.223	0.056*
	(0.168)	(0.159)	(0.032)
N	142	156	142

Notes: This table reports the effect of treatment on mindful behavior (Panel A), mental health and well-being (Panel B) as well as performance (Panel C). All estimations follow Equation (1) with control variables (such as gender and age). Panel A shows the estimates for overall mindfulness (sum of all single items and divided by amount of items) in column (1) and for its 12 single items. Here, higher values signify more mindful behavior. Panel B contains the treatment effect on depressive symptoms (PHQ-8, Panel B.i) in midterm (first follow-up) and perceived stress (PSS, Panel B.ii) at the end of the term (second follow-up). Panel B.i includes the estimates for the PHQ-8 scale in column (1) and for its eight single items with higher values indicating more problems. In Panel B.ii, we present the effect of treatment on perceived stress (column (1)) and its two sub-scales perceived helplessness and lack of self-efficacy. Here, higher values signify more problems. We developed both sub-scales following Taylor (2015). We further display the estimates for performance in in-course tests (Panel C.i) and exams (Panel C.ii) in Panel C. The score of all in-course tests is calculated by the sum of all tests divided by the number of attended tests. The combined score of both exams contains the last reached score (the main exam score is replaced by the res-it score, in case of attendance in the re-sit exam). We standardize all outcomes except the dummy variable perseverance to mean 0 and SD 1. Standard errors are reported in parenthesis. We list the number of observations (N) for each panel in the last row. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

5 Discussion

Our study faces both spillover effects and external validity as challenges. We discuss these below and provide robustness checks for our results.

Post treatment communication after the first lecture between friends makes treatment spillover possible. Treated and placebo group students might interact after the first lecture including the intervention and between the data collection times. While we monitor any possible compliance issue, we cannot control the environment outside the lecture halls enabling spillovers due to interactions. We already capture a portion of the spillover effects by including the size of the friend network in the estimation (for the estimation without the friends network see estimation without controls in Table 6, column (4)). However, this might not fully solve the problem of possible spillovers.

Two scenarios might play a role in interpreting our results. First, treated students may give advice to the placebo group students, which can positively affect the mental health of placebo students. A student that did not reflect can benefit from the moral support and counsel of a student who reflected on stressors and its remedies due to social effects (Duflo and Saez 2003[31]) or just the act of support (Colella et al. 2018[32]). Moreover, the placebo group might be stimulated by the treatment group and thereby start reflecting itself. These aspects would intimate an argument that our findings are the lower bounds of the actual effect on the treatment group.[33]

Secondly, the treated students can benefit further (Eskreis-Winkler et al. 2018, 2019), from offering this moral support to the placebo group students through either repeated reflecting on the topic in the process of offering support or the warm glow effect of helping out a peer. In this case, the treated

[31] Duflo and Saez (2003) find significantly higher registration of employees in a tax deferred account (TDA) in a group wherein some employees received an encouragement and information about TDAs in comparison to a group where nobody received it.

[32] Colella et al. (2018) analyzed the role of moral support by taking the example of the Government of Argentina's decision to ban visiting team's supporters in the stadium during league matches. They find that excluding supporters increases the likelihood to lose by 20% and the score difference between the two teams.

[33] Considering a special post-COVID cohort does not play a role in this context, as both groups were equally exposed to the COVID-19 pandemic and its measures.

students would further benefit from giving advice.[34] In such an instance, our findings are upper bounds of the actual effect or biased upwards.

Both of these situations strengthen the cost-efficiency of our intervention strategy. Within our set-up we cannot ascertain either the prevalence of or test if one overshadows the other. However, from a welfare perspective, the relative merits of both reinforce the need for future experiments with multiple treatment arms.

Overall, there are several other reasons to assume a lower bound estimation of the measured effects. First, the students participating in the RCT are at least in their third semester and have thus likely already developed coping strategies to deal with stress during their first year of university. Unlike first year students, they have already taken several exams (5-6 exams are taken per semester) and are thus more experienced when the intervention takes place. The intervention could therefore be less potent for these experienced students compared to those in their first semester. This hypothesis requires future testing. Second, the placebo group may feel important and valued just by being asked for their opinion about the University building and to give advice for improvements (on the basis of Eskreis-Winkler et al. (2018)). Accordingly, the outcomes of the placebo and treatment groups may converge and the effect may be underestimated.

To ensure that our results are robust we adopt two main strategies. In our first strategy, we test the results using four variations of our main specification (see Table 6). These variations first include interacting our treatment with a dummy indicating if the student answered the treatment seriously (see Table 6, column (2)). To do so, we classify the answer as serious if it fits the question in a meaningful way and can therefore be assigned to one of our categories (see Figure 4, Panel A and B). Since almost all students responded seriously to the questions and the students may be inspired to reflect by the question itself, we refrain from dropping the observations from the main sample, but check for robustness. This test indicates that all coefficients except hopelessness (p= 0.105) stay significant and of similar effect size.

[34] Eskreis-Winkler et al. (2019) conducted an experiment on high school students. The treatment group is asked to fill multiple-choice questions related to optimal study location and strategies, write a motivational letter to an anonymous younger student, and answer a battery of self-reported questions and fulfill a behavior task. On the other hand, the control group was not asked to give advice. This resulted in treated students scoring better in math and their target subject. Eskreis-Winkler et al. (2018) conducted an experiment on middle school children. The treatment group was asked to read and reply (give advice) to a letter (once per week for three weeks) written by a younger student. The intervention increased the time spend on studying.

Second, we replace the OLS main specification with an ANCOVA following McKenzie (2012) (see Table 6, column (3)). This is only possible for outcomes with baseline data. All coefficients, with the exception of sport at the end of the semester, remain either significant or close to significance. Here all signs are consistent, whereby the effect sizes tend to decrease.

As a third robustness check, we exclude the control variables (see Table 6, column (4)). By doing so, all coefficients (except hopelessness (p=0.106)) retain their significance and the effect sizes remain stable.

Last, we add baseline life satisfaction and conscious activities to the main specification (see Table 6, column (5)) to control for the two variables that are imabalanced in the baseline(see Table 1). The effects on sport, scheduling exercise, and perseverance stay significant, with selfcare (p=0.143) and sport at the end of the term (p=0.167) remaining close to significant. All coefficients retain their sign, although the effect sizes tend to slightly decrease. For mindfulness, the coefficient is reduced by half, which corresponds to the imbalance of conscious activities in the baseline.

As our second main strategy to check robustness, we perform several multiple hypothesis tests. We consider multiple hypothesis testing for the single items of the mindfulness and depressive symptoms scale and thus reduce the possibility of random false-positive results. By doing so, we follow three established strategies. First, we perform a Romano-Wolf multiple hypothesis correction for the single items of each scale separately (Clarke et al. 2020). Second, we use two different strategies with aggregated measures from both scales to adjust for multiple testing by reducing the number of performed tests in both approaches. For this, we follow Ashraf et al. (2020) and create an aggregated index of the 20 items (12 items of the mindfulness scale and 8 items of the PHQ-8 scale) by averaging the individual standardized variables. We further proceed as in Mani et al. (2020), and follow Anderson (2008) by computing the weighted mean of the standardized individual variables as an aggregated index.[35]

The Romano-Wolf tests for the mindfulness items in Table 7 and for the depressive symptoms items in Table 8 reveal that only the midterm effect on doing sport persists after the correction. In this respect, selfcare, scheduling exercise, low energy, and hopelessness lose significance. In contrast, the multiple hypothesis tests conducted using the two aggregations show significant improvements in Table 9. Accordingly, our results bear up to most tests.

[35] Since the baseline data are required for the calculation (Mani et al. 2020), we are only able to include 13 variables in total in this weighted index, as not all 12 items of the mindfulness scale were collected in the baseline.

6 Conclusion

Students' mental health plays a key role in how they perceive circumstances and respond to them inside and outside the classroom, especially under situations involving stress or overburdening (Chen et al. 2020; Huppert 2009; Kroher et al. 2023). We identify reflection to be instrumental for students to draw lessons from the past and envision concerted actions for the future, which can both influence individual mental health and behavior in the present and performance in future. These are exhibited in our results, wherein reflecting about past *"experiences with stress"* and directed actions in the future *"to deal with them to improve student well-being"* resulted in mindful behavior, heightened mental health and well-being, and higher perseverance in performance.

In spite of their merits, our results are subject to a number of limitations. First, we adapted the validated mindful self-care scale and translated this scale and the questions on risky behavior into German. The translation was done by the research team, who are fluent native German speakers, due to a lack of translation. Moreover, the findings on performace should be treated with caution due to the possible selection problem in the second stage. Another limitation lies in the limited sample size, which does not allow for investigation of heterogeneities and leaves scope for future research. Considering a special post-lockdown cohort during the first semester back in full in-person teaching offers a further limitation and at the same time a great advantage. On the one hand, this provides the possibility of supporting students after this difficult time and curbing the increasing mental health problems. At the same time, it imposes a limitation in terms of the external validity of our results.

Our soft-touch intervention serves as a great possibility to address the care gap in mental health services and likewise as a starting point for future research. Small adjustments can strengthen our results. For instance, increasing the sample size by conducting and randomizing the study at a university or even country level, might further limit potential spillovers and enable the analysis of heterogeneities. This remain subject to future research, as it requires further time and especially financing.

We developed an approach with low barriers for uptake and supply at universities and possibly other institutions that may address the care gap in mental health services in the long term. Accordingly, we responded to the open call issued by the research community and the remedies mentioned by the students in our experiment to expand services at universities to support students' mental health.

We hope that the simple nature of our intervention will allow it to be integrated into everyday university life with potential further adjustments that can both spark and inform future research. Overall, our findings may influence the design of university life globally and thereby significantly impact wider society.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge (2023). "When should you adjust standard errors for clustering?" In: *The Quarterly Journal of Economics* 138.1, pp. 1–35.

Acampora, Michelle, Francesco Capozza, Vahid Moghani, et al. (2022). *Mental Health Literacy, Beliefs and Demand for Mental Health Support among University Students*. Tech. rep. Tinbergen Institute.

Alan, Sule, Teodora Boneva, and Seda Ertac (2019). "Ever failed, try again, succeed better: Results from a randomized educational intervention on grit." In: *The Quarterly Journal of Economics* 134.3, pp. 1121–1162.

American College Health Association (2022). "Undergraduate student Reference Group, Executive Summary Spring 2022." In: *Silver Spring, MD: American College Health Association*.

Anderson, Michael L (2008). "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." In: *Journal of the American statistical Association* 103.484, pp. 1481–1495.

Anseel, Frederik, Adam S Beatty, Winny Shen, Filip Lievens, and Paul R Sackett (2015). "How are we doing after 30 years? A meta-analytic review of the antecedents and outcomes of feedback-seeking behavior." In: *Journal of management* 41.1, pp. 318–348.

Anseel, Frederik, Filip Lievens, and Eveline Schollaert (2009). "Reflection as a strategy to enhance task performance after feedback." In: *Organizational Behavior and Human Decision Processes* 110.1, pp. 23–35.

Ashraf, Nava, Natalie Bau, Corinne Low, and Kathleen McGinn (2020). "Negotiating a better future: How interpersonal skills facilitate intergenerational investment." In: *The Quarterly Journal of Economics* 135.2, pp. 1095–1151.

Auerbach, Randy P, Jordi Alonso, William G Axinn, Pim Cuijpers, David D Ebert, Jennifer G Green, Irving Hwang, Ronald C Kessler, Howard Liu, Philippe Mortier, et al. (2016). "Mental disorders among college students in the World Health Organization world mental health surveys." In: *Psychological medicine* 46.14, pp. 2955–2970.

Bandura, Albert (1993). "Perceived self-efficacy in cognitive development and functioning." In: *Educational psychologist* 28.2, pp. 117–148.

Berlin, DIW and others (2022). *SOEP-IS 2020 - Fragebogen für die SOEP-Innovations-Stichprobe*. Tech. rep. SOEP Survey Papers.

Bhan, Prateek C, Christina Felfe, Theodore Koutmeridis, Maike Schlosser, Judith Vornberger, and Jinglin Wen (2022). *Students and their thoughts: reflecting for a good semester. AEA RCT Registry*. https://doi.org/10.1257/rct.10277-1.0. Registry on: 2022-10-25.

Bhan, Prateek C and Judith Vornberger (2022). *Data Collection Protocol. Students and Their Thoughts*. https://www.dropbox.com/scl/fi/wjx581l7bwc0oger71rfj/Protocol_2.pdf?rlkey=pk0qcadsurlfpzi1tgjs3amzk&dl=0.

Bhide, Amar, Prakesh S Shah, and Ganesh Acharya (2018). "A simplified guide to randomized controlled trials." In: *Acta obstetricia et gynecologica Scandinavica* 97.4, pp. 380–387.

Boud, David, Rosemary Keogh, and David Walker (2013). "Promoting reflection in learning a model." In: *Boundaries of adult learning*. Routledge, pp. 32–56.

Busch, Susan H, Ezra Golberstein, and Ellen Meara (2014). "The fda and abcs unintended consequences of antidepressant warnings on human capital." In: *Journal of Human Resources* 49.3, pp. 540–571.

Cassar, Lea, Mira Fischer, and Vanessa Valero (2022). *Keep calm and carry on: The short-vs. long-run effects of mindfulness meditation on (academic) performance*. Tech. rep. IZA Discussion Papers.

Chekroud, Sammi R, Ralitza Gueorguieva, Amanda B Zheutlin, Martin Paulus, Harlan M Krumholz, John H Krystal, and Adam M Chekroud (2018). "Association between physical exercise and mental health in $1 \cdot 2$ million individuals in the USA between 2011 and 2015: a cross-sectional study." In: *The lancet psychiatry* 5.9, pp. 739–746.

Chen, Shu-Ping, Wen-Pin Chang, and Heather Stuart (2020). "Self-reflection and screening mental health on Canadian campuses: validation of the mental health continuum model." In: *BMC psychology* 8.1, pp. 1–8.

Clark, Leslie F (1993). "Stress and the cognitive-conversational benefits of social interaction." In: *Journal of social and clinical psychology* 12.1, pp. 25–55.

Clarke, Damian, Joseph P Romano, and Michael Wolf (2020). "The Romano–Wolf multiple-hypothesis correction in Stata." In: *The Stata Journal* 20.4, pp. 812–843.

Cohen, Sheldon, Tom Kamarck, Robin Mermelstein, et al. (1994). "Perceived stress scale." In: *Measuring stress: A guide for health and social scientists* 10.2, pp. 1–2.

Colella, Fabrizio, Patricio Dalton, and Giovanni Giusti (2018). "You'll never walk alone: the effect of moral support on performance." In.

Cook-Cottone, Catherine P and Wendy M Guyker (2018). "The development and validation of the Mindful Self-Care Scale (MSCS): An assessment of practices that support positive embodiment." In: *Mindfulness* 9, pp. 161–175.

Cornaglia, Francesca, Elena Crivellaro, and Sandra McNally (2015). "Mental health and education decisions." In: *Labour Economics* 33, pp. 1–12.

Coulson, Debra and Marina Harvey (2013). "Scaffolding student reflection for experience-based learning: A framework." In: *Teaching in Higher Education* 18.4, pp. 401–413.

Cunningham, Scott (2021). *Causal inference: The mixtape*. Yale university press.

Czyżowska, Natalia and Ewa Gurba (2021). "Does reflection on everyday events enhance meaning in life and well-being among emerging adults? Self-efficacy as mediator between meaning in life and well-being." In: *International Journal of Environmental Research and Public Health* 18.18, p. 9714.

Department of Administrative Services (2023). *Perceived Stress Scale*. https://www.das.nh.gov/wellness/docs/percievedstressscale.pdf. Accessed: 2022-08-24.

Duffy, Anne, Kate EA Saunders, Gin S Malhi, Scott Patten, Andrea Cipriani, Stephen H McNevin, Ellie MacDonald, and John Geddes (2019). "Mental health care for university students: a way forward?" In: *The Lancet Psychiatry* 6.11, pp. 885–887.

Duflo, Esther, Michael Kremer, and Jonathan Robinson (2006). "Understanding technology adoption: Fertilizer in western Kenya, preliminary results from field experiments." In: *Mimeo*.

Duflo, Esther and Emmanuel Saez (2003). "The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment." In: *The Quarterly journal of economics* 118.3, pp. 815–842.

Dyment, Janet E and Timothy S O'Connell (2011). "Assessing the quality of reflection in student journals: A review of the research." In: *Teaching in Higher Education* 16.1, pp. 81–97.

Ebert, David Daniel, Philippe Mortier, Fanny Kaehlke, Ronny Bruffaerts, Harald Baumeister, Randy P Auerbach, Jordi Alonso, Gemma Vilagut, Kalina U Martínez, Christine Lochner, et al. (2019). "Barriers of mental health treatment utilization among first-year college students: First cross-national results from the WHO World Mental Health International College Student Initiative." In: *International journal of methods in psychiatric research* 28.2, e1782.

Eisenberg, Daniel, Ezra Golberstein, and Sarah E Gollust (2007). "Help-seeking and access to mental health care in a university student population." In: *Medical care*, pp. 594–601.

Eisenberg, Daniel, Ezra Golberstein, and Justin B Hunt (2009). "Mental health and academic success in college." In: *The BE Journal of Economic Analysis & Policy* 9.1.

Ellis, Shmuel, Bernd Carette, Frederik Anseel, and Filip Lievens (2014). "Systematic reflection: Implications for learning from failures and successes." In: *Current Directions in Psychological Science* 23.1, pp. 67–72.

Eskreis-Winkler, Lauren, Ayelet Fishbach, and Angela L Duckworth (2018). "Dear Abby: Should I give advice or receive it?" In: *Psychological Science* 29.11, pp. 1797–1806.

Eskreis-Winkler, Lauren, Katherine L Milkman, Dena M Gromet, and Angela L Duckworth (2019). "A large-scale field experiment shows giving advice improves academic outcomes for the advisor." In: *Proceedings of the national academy of sciences* 116.30, pp. 14808–14810.

Falon, Samantha L, Eyal Karin, Danny Boga, Daniel F Gucciardi, Barbara Griffin, and Monique F Crane (2021). "A clustered-randomized controlled trial of a self-reflection resilience-strengthening intervention and novel mediators." In: *Journal of occupational health psychology* 26.1, p. 1.

Fitzpatrick, Kathleen Kara, Alison Darcy, and Molly Vierhile (2017). "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial." In: *JMIR mental health* 4.2, e7785.

Geiger, Katherine A and Paul Kwon (2010). "Rumination and depressive symptoms: Evidence for the moderating role of hope." In: *Personality and Individual Differences* 49.5, pp. 391–395.

Gesis, Leibniz Institut für Sozialwissenschaften (2023). *General Life Satisfaction*. https://www.gesis.org/en/services/processing-and-analyzing-data/data-harmonization/question-link/general-life-satisfaction. Accessed: 2022-08-24.

Gräfe, Kerstin, Stephan Zipfel, Wolfgang Herzog, and Bernd Löwe (2004). "Screening psychischer Störungen mit dem "Gesundheitsfragebogen für Patienten (PHQ-D)"." In: *Diagnostica* 50.4, pp. 171–181.

Hendrickx, Gaelle, Veronique De Roeck, Athanasios Maras, Gwen Dieleman, Suzanne Gerritsen, Diane Purper-Ouakil, Frédérick Russet, Renate Schepker, Giulia Signorini, Swaran Preet Singh, et al. (2020). "Challenges during the transition from child and adolescent mental health services to adult mental health services." In: *BJPsych bulletin* 44.4, pp. 163–168.

Hotchkiss, Jason T and Catherine P Cook-Cottone (2019). "Validation of the Mindful Self-Care Scale (MSCS) and development of the Brief-MSCS among hospice and healthcare professionals: A confirmatory factor analysis approach to validation." In: *Palliative & supportive care* 17.6, pp. 628–636.

Huppert, Felicia A (2009). "Psychological well-being: Evidence regarding its causes and consequences." In: *Applied psychology: health and well-being* 1.2, pp. 137–164.

Hysenbegasi, Alketa, Steven L Hass, and Clayton R Rowland (2005). "The impact of depression on the academic productivity of university students." In: *Journal of mental health policy and economics* 8.3, p. 145.

Ibrahim, Ahmed K, Shona J Kelly, Clive E Adams, and Cris Glazebrook (2013). "A systematic review of studies of depression prevalence in university students." In: *Journal of psychiatric research* 47.3, pp. 391–400.

Institute of Health Metrics and Evaluation (2022). *Global Health Data Exchange (GHDx). GBD Results*. https://vizhub.healthdata.org/gbd-results/. Accessed: 2022-08-22.

Kaparounaki, Chrysi K, Mikaella E Patsali, Danai-Priskila V Mousa, Eleni VK Papadopoulou, Konstantina KK Papadopoulou, and Konstantinos N Fountoulakis (2020). "University students' mental health amidst the COVID-19 quarantine in Greece." In: *Psychiatry research* 290, p. 113111.

Kendall, JM (2003). "Designing a research project: randomised controlled trials and their principles." In: *Emergency Medicine Journal* 20.2, pp. 164–168.

Kessler, Ronald C, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters (2005). "Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication." In: *Archives of general psychiatry* 62.6, pp. 593–602.

Khoury, Bassam, Manoj Sharma, Sarah E Rush, and Claude Fournier (2015). "Mindfulness-based stress reduction for healthy individuals: A meta-analysis." In: *Journal of psychosomatic research* 78.6, pp. 519–528.

Kim, Hanjoo, Gavin N Rackoff, Ellen E Fitzsimmons-Craft, Ki Eun Shin, Nur Hani Zainal, Jeremy T Schwob, Daniel Eisenberg, Denise E Wilfley, C Barr Taylor, and Michelle G Newman (2022). "College mental health before and during the COVID-19 pandemic: Results from a nationwide survey." In: *Cognitive therapy and research* 46.1, pp. 1–10.

Klein, Eva M, Elmar Brähler, Michael Dreier, Leonard Reinecke, Kai W Müller, Gabriele Schmutzer, Klaus Wölfling, and Manfred E Beutel (2016). "The German version of the Perceived Stress Scale–psychometric characteristics in a representative German community sample." In: *BMC psychiatry* 16, pp. 1–10.

Kroenke, Kurt, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad (2009). "The PHQ-8 as a measure of current depression in the general population." In: *Journal of affective disorders* 114.1-3, pp. 163–173.

Kroher, Martina, Mareike Beuße, Sören Isleib, Karsten Becker, Marie-Christin Ehrhardt, Frederike Gerdes, Jonas Koopmann, Theresa Schommer, Ulrike Schwabe, Julia Steinkühler, Daniel Völk, Frauke Peter, and Sandra Buchholz (2023). "The student survey in Germany: 22nd Social Survey. The Economic and Social Situation of Students in Germany 2021."

Lee, Rebecca Anne and Mary Elizabeth Jung (2018). "Evaluation of an mhealth app (destressify) on university students' mental health: pilot trial." In: *JMIR mental health* 5.1, e8324.

Lepore, Stephen J, Pablo Fernandez-Berrocal, Jennifer Ragan, and Natalia Ramos (2004). "It's not that bad: Social challenges to emotional disclosure enhance adjustment to stress." In: *Anxiety, Stress & Coping* 17.4, pp. 341–361.

Li, Yang, Aiwen Wang, Yalin Wu, Nana Han, and Huiming Huang (2021a). "Impact of the COVID-19 pandemic on the mental health of college students: a systematic review and meta-analysis." In: *Frontiers in psychology* 12, p. 669119.

Li, Yuanyuan, Jingbo Zhao, Zijuan Ma, Larkin S McReynolds, Dihuan Lin, Zihao Chen, Tong Wang, Dongfang Wang, Yifan Zhang, Jinfang Zhang, et al. (2021b). "Mental health among college students during the COVID-19 pandemic in China: a 2-wave longitudinal survey." In: *Journal of affective disorders* 281, pp. 597–604.

Löwe, B, RL Spitzer, S Zipfel, W Herzog, et al. (2002). "Manual: Komplettversion und Kurzform Autorisierte Deutsche Version Des "Prime MD Patient Health Questionnaire (PHQ)"." In: *Nervenarzt* 6, pp. 2–11.

Ma, Liang, Yingnan Zhang, and Zeshi Cui (2019). "Mindfulness-based interventions for prevention of depressive symptoms in university students: a meta-analytic review." In: *Mindfulness* 10, pp. 2209–2224.

Mani, A, S Ghosal, S Jana, and S Roy (2020). "Sex workers, stigma and self-image: evidence from Kolkata brothels." In.

McCrindle, Andrea R and Carol A Christensen (1995). "The impact of learning journals on metacognitive and cognitive processes and learning performance." In: *Learning and instruction* 5.2, pp. 167–185.

McGorry, Patrick D, Rosemary Purcell, Sherilyn Goldstone, and G Paul Amminger (2011). "Age of onset and timing of treatment for mental and substance use disorders: implications for preventive intervention strategies and models of care." In: *Current opinion in psychiatry* 24.4, pp. 301–306.

McKenzie, David (2012). "Beyond baseline and follow-up: The case for more T in experiments." In: *Journal of development Economics* 99.2, pp. 210–221.

Meichenbaum, Donald and Deborah Fitzpatrick (1993). "A constructivist narrative perspective on stress and coping: Stress inoculation applications." In.

Nafilyan, Vahé, Mauricio Avendano, and Augustin de Coulon (2021). "The causal impact of depression on cognitive functioning: evidence from Europe." In.

Ochnik, Dominika, Aleksandra M Rogowska, Cezary Kuśnierz, Monika Jakubiak, Astrid Schütz, Marco J Held, Ana Arzenšek, Joy Benatov, Rony Berger, Elena V Korchagina, et al. (2021). "Mental health prevalence and predictors among university students in nine countries during the COVID-19 pandemic: a cross-national study." In: *Scientific reports* 11.1, p. 18644.

Phan, Huy P (2014). "Self-efficacy, reflection, and achievement: A short-term longitudinal examination." In: *The Journal of Educational Research* 107.2, pp. 90–102.

Radović, Slaviša, Olga Firssova, Hans GK Hummel, and Marjan Vermeulen (2023). "Improving academic performance: Strengthening the relation between theory and practice through prompted reflection." In: *Active Learning in Higher Education* 24.2, pp. 139–154.

Richter, David, Julia Rohrer, Maria Metzing, Wiebke Nestler, Michael Weinhardt, and Jürgen Schupp (2017). *SOEP scales manual (updated for SOEP-Core v32. 1)*. Tech. rep. SOEP Survey Papers.

Ryan, Mary (2013). "The pedagogical balancing act: Teaching reflection in higher education." In: *Teaching in Higher Education* 18.2, pp. 144–155.

Schneider, Eva Elisa, Sandra Schönfelder, Mila Domke-Wolf, and Michèle Wessa (2020). "Measuring stress in clinical and nonclinical subjects using a German adaptation of the Perceived Stress Scale." In: *International Journal of Clinical and Health Psychology* 20.2, pp. 173–181.

Smyth, Joshua M (1998). "Written emotional expression: effect sizes, outcome types, and moderating variables." In: *Journal of consulting and clinical psychology* 66.1, p. 174.

Starr, Lisa R (2015). "When support seeking backfires: Co-rumination, excessive reassurance seeking, and depressed mood in the daily lives of young adults." In: *Journal of social and clinical psychology* 34.5, pp. 436–457.

Stibe, Agnis and Brian Cugelman (2016). "Persuasive backfiring: When behavior change interventions trigger unintended negative outcomes." In: *Persuasive Technology: 11th International Conference, PERSUASIVE 2016, Salzburg, Austria, April 5-7, 2016, Proceedings 11*. Springer, pp. 65–77.

Storrie, Kim, Kathy Ahern, and Anthony Tuckett (2010). "A systematic review: students with mental health problems—a growing problem." In: *International journal of nursing practice* 16.1, pp. 1–6.

Taylor, John M (2015). "Psychometric analysis of the ten-item perceived stress scale." In: *Psychological assessment* 27.1, p. 90.

The Gallup Organization Ltd. (2021). *Wellcome Global Monitor*. https://wellcome.org/reports/wellcome-global-monitor-mental-health/2020. Accessed: 2023-04-18.

The Oxford Student (2019). "JCR Presidents sign open letter calling for action on mental health." In: *Oxford University's Student Newspaper*.

Tomm, Karl (1988). "Interventive interviewing: Part III. Intending to ask lineal, circular, strategic, or reflexive questions?" In: *Family process* 27.1, pp. 1–15.

Trede, Franziska and Denise Jackson (2021). "Educating the deliberate professional and enhancing professional agency through peer reflection of work-integrated learning." In: *Active Learning in Higher Education* 22.3, pp. 171–187.

Vorontsova-Wenger, Olga, Paolo Ghisletta, Valentin Ababkov, and Koviljka Barisnikov (2021). "Relationship between mindfulness, psychopathological symptoms, and academic performance in university students." In: *Psychological reports* 124.2, pp. 459–478.

Watkins, Daphne C, Justin B Hunt, and Daniel Eisenberg (2012). "Increased demand for mental health services on college campuses: Perspectives from administrators." In: *Qualitative Social Work* 11.3, pp. 319–337.

White, Halbert (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." In: *Econometrica: journal of the Econometric Society*, pp. 817–838.

Wortman, Camille B and Darrin R Lehman (1985). "Reactions to victims of life crises: Support attempts that fail." In: *Social support: Theory, research and applications*. Springer, pp. 463–489.

Yang, Miles M, Yucheng Zhang, and Feifei Yang (2018). "How a reflection intervention improves the effect of learning goals on performance outcomes in a complex decision-making task." In: *Journal of Business and Psychology* 33, pp. 579–593.

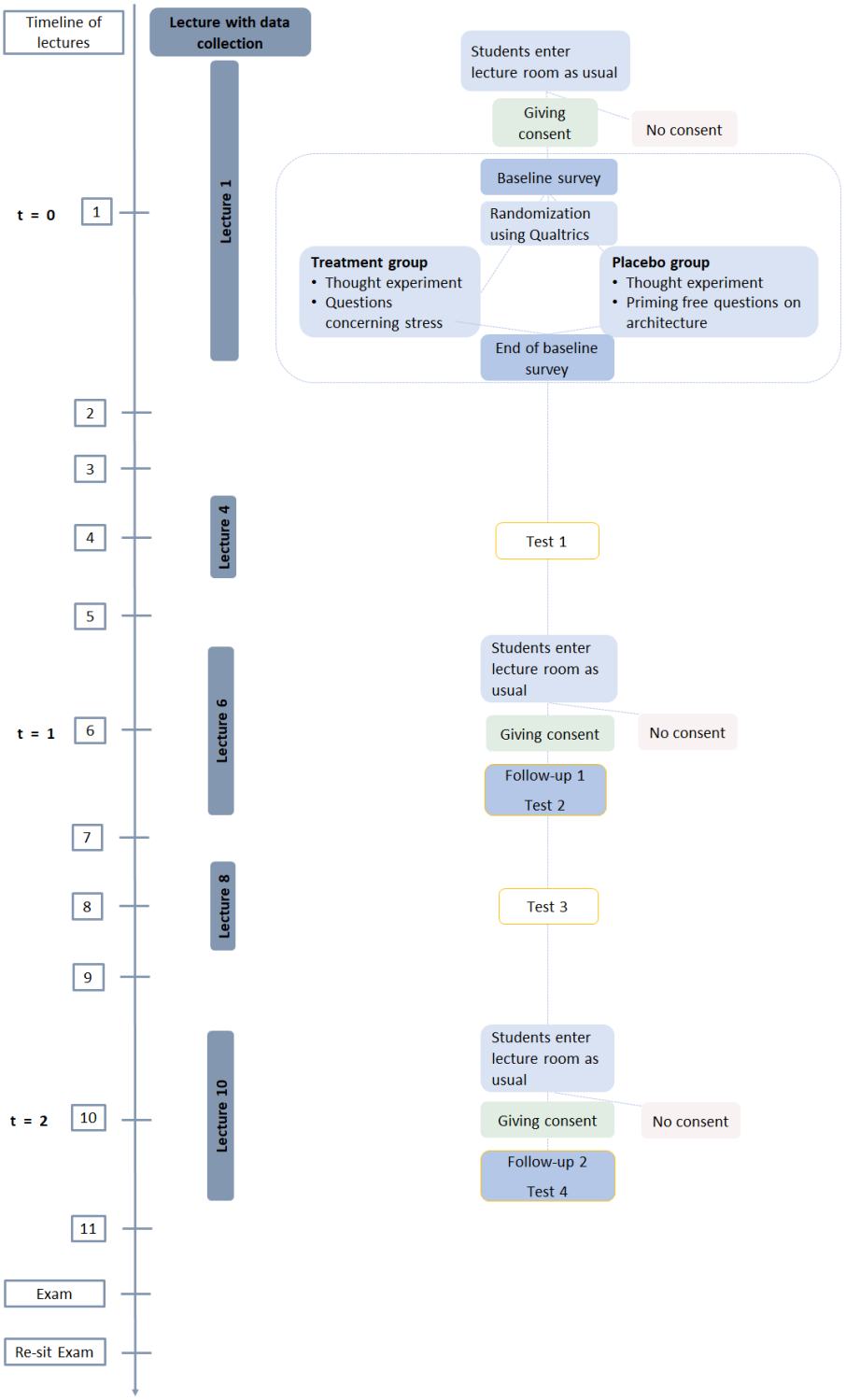
Yao, Yingying, Jianqiao Chen, Dan Dong, Yi Feng, and Zhihong Qiao (2022). "The relationship between exercise and mental health outcomes during the COVID-19 pandemic: from the perspective of hope." In: *International Journal of Environmental Research and Public Health* 19.7, p. 4090.

Yonemoto, Naohiro and Yoshitaka Kawashima (2023). "Help-seeking behaviors for mental health problems during the COVID-19 pandemic: A systematic review." In: *Journal of Affective Disorders* 323, pp. 85–100.

Zizzo, Daniel John (2010). "Experimenter demand effects in economic experiments." In: *Experimental Economics* 13, pp. 75–98.

7 Appendix

Figure 2: Flowchart of timeline



Notes (*continued on next page*): This figure represents the timeline and structure of data collection procedure using a flowchart.

Notes (Figure 2 contd.): We collected data within one academic semester from October 2022 (lecture 1) to April 2023 (re-sit exam), whereby the course instructor provides us with exam data and data of in-course tests. The chart displays the timeline of the course including all lectures held as a vertical blue arrow. The lectures and exams are highlighted with corresponding markers on the left side. The time intervals between the lectures symbolize one semester week (not equivalent to one calendar week, depending on vacations). The lectures marked in dark blue in the second column (lecture 1, 4, 6, 8, 10) represent the time points in which data were collected and in-course tests were conducted. The adjacent column in the middle of the figure illustrates the data collection processes in more detail. Here, consent is marked in green and non-consent in red, whereas the surveys are colored in a medium shade of blue and the remaining information in light blue. Each of the four multiple-choice tests took place in one of the lectures framed in orange (lectures 4, 6, 8, 10). We illustrate the chronological sequence with the dashed connecting lines, whereas the dashed box encloses the elements that occurred within the basic questionnaire.

Table 3: Summary on data collection and administrative data

Category	Indicator	Scale/Items	Data collection
Panel A: Survey data			
Merging variable	Matriculation number	Single survey question	B, F1, F2, IT, ME, RE
Demographic characteristics	Demographic characteristics	Single survey questions	B, F2
Mental health and well-being	Depressive symptoms	Patient-Health-Questionnaire (PHQ-8)	B, F1, F2
	Perceived stress	Perceived Stress Scale (PSS)	B, F2
	Life satisfaction	SOEP 11pt question	B, F2
	Risky behavior	Two single questions	B, F2
Mindful behavior	Mindful behavior	Battery of questions with Mindful Self-Care Scale (MSCS) used as a basis	B, F1, F2
Panel B: Administrative data			
Attendance and performance	Attendance	Attendance	B, F1, F2, IT, ME, RE
	Performance	Test scores	IT
	Performance	Exam scores, grades	ME, RE
	Drop-out	Attendance, cross-outs	ME, RE

Notes: This table provides a summary of the collected and administrative data with their categories, as well as information on indicators, the scales used for measurement and data collection time. We collect data at or use data from various points in time: baseline (B), follow-up 1 (F1), follow-up 2 (F2), four in-course tests (IT), main exam (ME), re-sit exam (RE).

Table 4: Treatment effects at the end of the term

	(1)	(2)	(3)	(4)
Panel A: Mindful behavior	Sport	Things I like	Listener	Relaxing
Treatment	0.330*	0.065	0.135	-0.182
	(0.191)	(0.182)	(0.172)	(0.191)
Controls	YES	YES	YES	YES
N	136	136	136	136
			Risky behaviour	
Panel B: Mental health and well-being	PHQ-8	Life satisfaction	Smoking cigarettes	Drinking alcohol
Treatment	-0.217	0.255	0.779	-0.103
	(0.173)	(0.174)	(1.111)	(0.352)
Controls	YES	YES	YES	YES
N	136	136	136	136
Panel C: Attendance	First in-course test	All in-course tests	Main exam	Both exams
Treatment	0.063	0.034	0.009	-0.038
	(0.060)	(0.074)	(0.060)	(0.051)
Controls	YES	YES	YES	YES
N	184	184	184	184

Notes: This table reports estimates of the effect of treatment on items of mindfulness (Panel A), mental health and well-being (Panel B) as well as attendance (Panel C). All estimations follow Equation (1) with control variables (such as gender and age). The listed outcome variables differ from Table 2, as other variables are collected in the second follow-up (for details see Table 3). In Panel A, higher values signify more mindful behavior. In Panel B in all columns except column (2), higher values signify more problems. In Panel C all attendance variables are dummy variables. In column (1) and (3) the dummy is 1 if the student attended the respective test/exam, in column (2) the dummy is 1 if the student attended all tests, while in column (4) if the student attended at least one of both exams. All outcomes of Panel A and B are standardized (mean, 0; SD, 1). Standard errors are reported in parenthesis. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Table 5: Attrition tests

	(1)	(2)	(3)	(4)
Panel A: T-test of equality	Placebo group	Treatment group	P-Value	
Baseline	94	91		
Attrition rate follow-up 1	0.160	0.209	0.390	
Attrition rate follow-up 2	0.213	0.308	0.142	
Attrition rate main exam	0.223	0.242	0.769	
Attrition rate re-sit exam	0.883	0.890	0.879	
Attrition rate test 1	0.255	0.209	0.457	
Attrition rate test 2	0.277	0.352	0.274	
Attrition rate test 3	0.372	0.385	0.864	
Attrition rate test 4	0.383	0.407	0.744	
Panel B: Regression	Attrition follow-up 1	Attrition follow-up 2	Attrition main exam	Attrition re-sit exam
Treatment	0.049	0.095	0.018	0.007
	(0.057)	(0.065)	(0.062)	(0.047)
Controls	NO	NO	NO	NO
N	185	185	185	185
Panel C: Regression	Attrition test 1	Attrition test 2	Attrition test 3	Attrition test 4
Treatment	-0.047	0.075	0.012	0.024
	(0.062)	(0.068)	(0.072)	(0.072)
Controls	NO	NO	NO	NO
N	185	185	185	185

Notes: This table presents results for attrition tests. Panel A examines the attrition rate for survey, exam, and in-course test participation by groups. Column (3) shows the t-test of equality of the means across the two groups for attrition. We further show the attrition rate for survey participation (Panel B, column (1) and (2)), exam participation (Panel B, column (3) and (4)) and test participation for each test (Panel C) by groups using a linear regression without controls, where the outcome variables are indicators for the respective attendance. Standard errors are reported in parenthesis. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Table 6: Robustness checks with different specifications

	(1)	(2)	(3)	(4)	(5)
Panel A: Mindfulness					
Treatment	0.321**	0.330**		0.368**	0.153
	(0.154)	(0.154)		(0.161)	(0.126)
Controls	YES	YES		NO	YES
N	151	151		151	151
Panel B: Sport					
Treatment	0.452***	0.431***	0.231**	0.496***	0.363***
	(0.151)	(0.152)	(0.112)	(0.159)	(0.138)
Controls	YES	YES	YES	NO	YES
N	151	151	151	151	151
Panel C: Sched. exerc.					
Treatment	0.385**	0.363**		0.428***	0.320**
	(0.158)	(0.158)		(0.159)	(0.152)
Controls	YES	YES		NO	YES
N	151	151		151	151
Panel D: Selfcare					
Treatment	0.336**	0.285*	0.178	0.355**	0.226
	(0.163)	(0.165)	(0.157)	(0.163)	(0.153)
Controls	YES	YES	YES	NO	YES
N	151	151	151	151	151
Panel E: Hopeless					
Treatment	-0.260*	-0.254	-0.216	-0.261	-0.097
	(0.156)	(0.156)	(0.132)	(0.160)	(0.145)
Controls	YES	YES	YES	NO	YES
N	151	151	151	151	151
Panel F: Little energy					
Treatment	-0.271*	-0.286*	-0.198	-0.287*	-0.193
	(0.160)	(0.159)	(0.135)	(0.162)	(0.168)
Controls	YES	YES	YES	NO	YES
N	151	151	151	151	151
Panel G: Sport (end of term)					
Treatment	0.330*	0.335*	0.070	0.364*	0.251
	(0.191)	(0.190)	(0.159)	(0.198)	(0.181)
Controls	YES	YES	YES	NO	YES
N	136	136	136	136	136
Panel H: Perseverance					
Treatment	0.056*	0.058*		0.059*	0.064*
	(0.032)	(0.033)		(0.034)	(0.035)
Controls	YES	YES		NO	YES
N	142	142		142	142

Notes: In this table we report the coefficients of the main specification (see Equation (1)) in midterm in column (1). For column (2), we use the main specification, but interact the treatment dummy with the dummy of seriously answering the treatment questions (=1 if first and second answer seriously answered). The coefficients of the ANCOVA estimation is presented in column (3). For this we follow McKenzie (2012) and include respective outcome measured in the baseline as an additional control variable in Equation (1). ANCOVA analysis is only possible to perform for outcome variables also measured in the baseline. The main specification without control variables (Equation (1) without X_i') is depicted in column (4). We expand the control variables of the main specification by baseline life satisfaction and conscious activities in column (5). We do so to control for the significant differences between the two groups in these two variables (see Table 1). All outcomes (except Panel H) are standardized (mean, 0; SD, 1). Standard errors are reported in parenthesis. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Table 7: Multiple hypothesis tests of mindfulness items

	(1)	(2)	(3)	(4)
	Nutrition	Sport	Sched. exerc.	Liked things
Treatment	-0.079	0.452***	0.385**	0.023
	(0.164)	(0.151)	(0.158)	(0.159)
	[1.000]	[0.059]	[0.255]	[1.000]
N	151	151	151	151
	Listen to me	Selecting thought	Accept failure	Purpose
Treatment	0.230	0.128	0.032	0.208
	(0.161)	(0.163)	(0.162)	(0.158)
	[0.745]	[0.961]	[1.000]	[0.765]
N	151	151	151	151
	Special people	Rest	Relax	Selfcare
Treatment	0.165	-0.055	0.053	0.336**
	(0.168)	(0.168)	(0.164)	(0.163)
	[0.863]	[1.000]	[1.000]	[0.373]
N	151	151	151	151

Notes: This table reports estimates of the effect of treatment on items of mindfulness in the first follow-up. All variables are standardized (mean, 0; SD, 1). Standard errors are reported in parenthesis. Significance of main estimation using Equation (1) at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively. Romano-Wolf p-value, that include a multiple hypothesis correction, are reported in square brackets (Clarke et al. 2020).

Table 8: Multiple hypothesis tests of PHQ-8 items

	(1)	(2)	(3)	(4)
	Little pleasure	Hopeless	Sleeping problems	Little energy
Treatment	-0.159	-0.260*	-0.085	-0.271*
	(0.162)	(0.156)	(0.163)	(0.160)
	[0.765]	[0.431]	[0.804]	[0.412]
N	151	151	151	151
	Poor/ overeating	Feel failure	Low concentr.	Speaking issues
Treatment	-0.130	-0.211	0.179	0.078
	(0.159)	(0.153)	(0.169)	(0.162)
	[0.784]	[0.569]	[0.745]	[0.804]
N	151	151	151	151

Notes: This table reports estimates of the effect of treatment on items of PHQ-8 in the first follow-up. All variables are standardized (mean, 0; SD, 1). Standard errors are reported in parenthesis. Significance of main estimation using Equation (1) at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively. Romano-Wolf p-value, that include a multiple hypothesis correction, are reported in square brackets (Clarke et al. 2020).

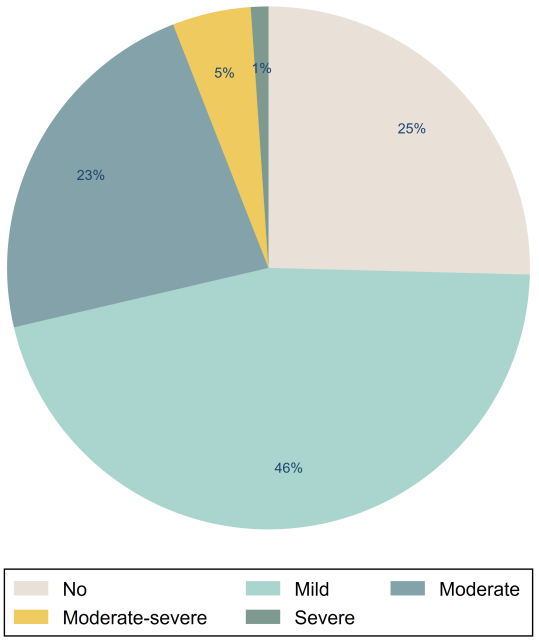
Table 9: Multiple hypotheses tests with aggregated indexes

	(1)	(2)
	Unweighted index	Weighted index
Treatment	0.137*	0.130*
	(0.070)	(0.074)
N	151	151

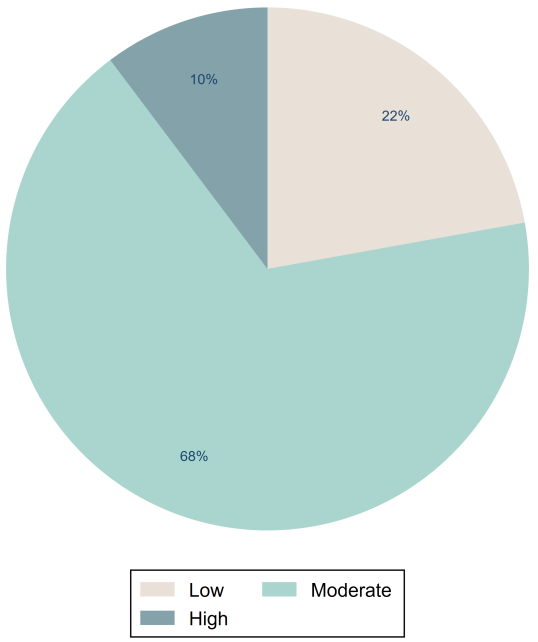
Notes: This table presents two adjustments for multiple hypotheses testing for midterm mindfulness and PHQ-8 items. We reverse-code the items of the PHQ-8 scale for both aggregates. Thus for both indexes, higher values signify an improvement. We follow Ashraf et al. (2020) in column (1) and create an aggregated index of the 20 items (12 items of the mindfulness scale and 8 items of the PHQ-8 scale) by averaging the individual standardized variables. In column (2), we proceed as in Mani et al. (2020), and follow Anderson (2008) by computing the weighted mean of the standardized individual variables as an aggregated index. Since the baseline data are required for the calculation (Mani et al. 2020), we are only able to include 13 variables in this weighted index, as not all 12 items of the mindfulness scale were collected in the baseline. Standard errors are reported in parenthesis. Significance at the 10%, 5%, and 1% levels is indicated by *, **, and ***, respectively.

Figure 3: Baseline levels of mental health

Panel A: Level of depressive symptoms



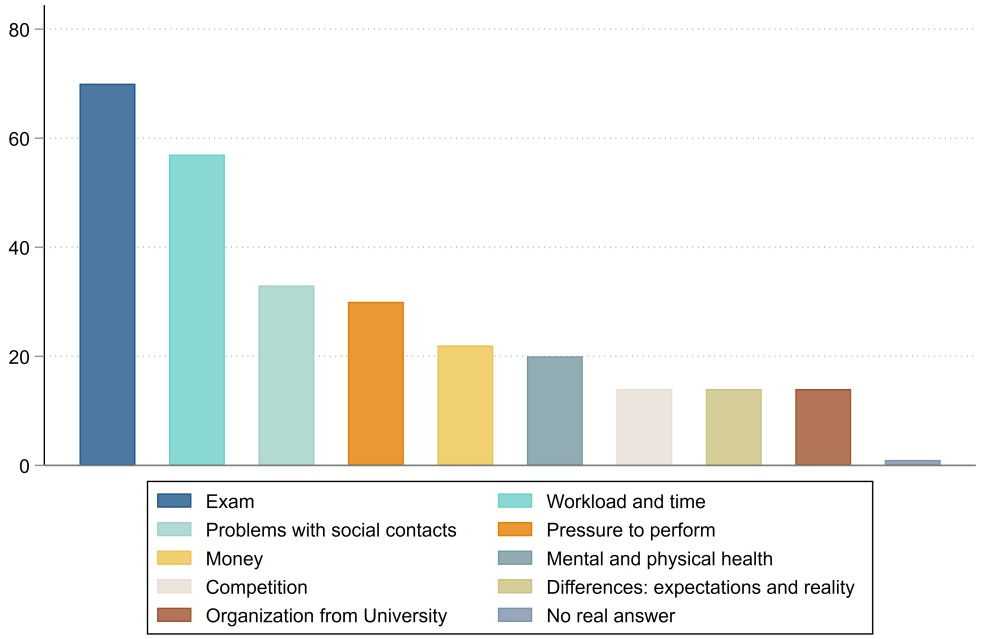
Panel B: Level of perceived stress



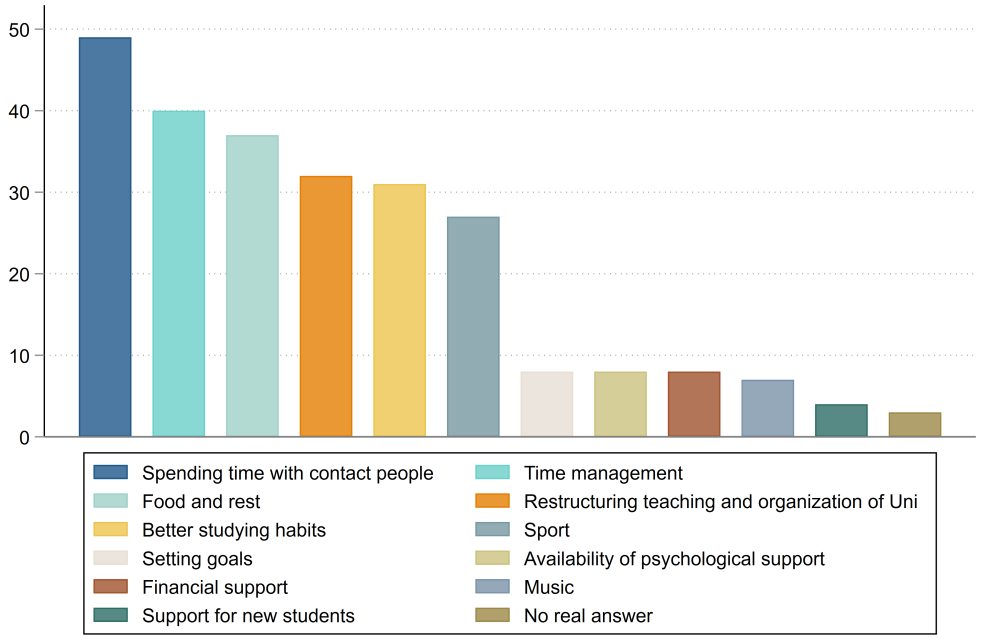
Notes: We present baseline levels of depressive symptoms (PHQ-8, Panel A) and perceived stress (PSS, Panel B) in this graph. The respective categorizations and their cut-off values are defined according to the guidelines of Kroenke et al. (2009) (for PHQ-8) and Department of Administrative Services (2023) (for PSS). (The Perceived-Stress-Scale is no diagnostic instrument, therefore the creators did not set any cut-off values. They only evaluate the perceived stress by stating that higher values represent more stress (Cohen et al. 1994, Klein et al. 2016). We use the cut-off values developed by Department of Administrative Services (2023) to categorize and improve the evaluation of students' initial stress levels.)

Figure 4: Answers on treatment questions

Panel A: Main stressors among college students



Panel B: Remedies to deal with stressors



Notes: This graph illustrates the frequency of certain response categories to the two treatment questions: Panel A: Based on your experience as a student, what are the main stressors (factors that can cause stress) among college students? Panel B: Please carefully write down 7-8 ways (points) to deal with these stressors and improve student well-being. The students wrote their answers in a blank box. Then, we created and mapped the above categories according to the answers. A student might have given an answer that belonged to several categories. This is then assigned to multiple categories.