



Gutenberg School of Management and Economics
& Research Unit “Interdisciplinary Public Policy”

Discussion Paper Series

Effects of a Factorial Intervention on Emotional Skills

Wanda Schleder

9th Oktober 2025

Discussion paper number 2507

Johannes Gutenberg University Mainz
Gutenberg School of Management and Economics
Jakob-Welder-Weg 9
55128 Mainz
Germany
<https://wiwi.uni-mainz.de/>

Contact details

Wanda Schleder
Chair in Macroeconomics
Johannes-Gutenberg University
Jakob-Welder-Weg 4
55128 Mainz
Germany
waschled@uni-mainz.de

Effects of a Factorial Intervention on Emotional Skills*

Wanda Schleder[†]

October 9, 2025

Abstract

This paper examines the effectiveness of a factorial intervention aimed at improving emotional skills done at the Johannes Gutenberg University Mainz in the winter-semester 2023/2024. The four intervention components, mindfulness, emotional regulation, self-acceptance, and resource activation, were evaluated with respect to their impact on emotional skills and other well-being outcomes. Due to a high dropout rate, no definitive conclusions can be drawn about the optimal composition of training components. However, resource activation shows a significantly positive effect on stress and a depression-related score. The findings also indicate that the Big Five personality traits play a crucial role in determining outcome variables. In addition, several approaches to estimating treatment effects were compared. The results suggest that a regression approach that directly accounts for all intervention factors and baseline scores should be preferred over simpler effect size measures.

Keywords: non-cognitive skills, emotional skill intervention, factorial experiment, effect sizes

JEL Codes: C93, I10, I19, I31, J24

*The TESC course was a joint project with Klaus Wälde and Sarah Tran-Huu. I am deeply grateful to them for their organizational and technical guidance and for their contributions.

[†]Johannes Gutenberg University Mainz

1 Introduction

The motivation. The fact that not only cognitive skills, such as IQ, matter for the well-being, health, and income of an individual has been extensively discussed by economic research in the past years and decades (Heckman and Kautz 2012). Non-cognitive skills, often referred to as "soft skills", have become central to explaining a wide range of life outcomes. It has been shown that non-cognitive skills play a role for individuals' subjective well-being, health, wages, and more (Heckman 2006, Heckman and Kautz 2012). Additionally, an increasing return to noncognitive skills on wages has been shown, underscoring their continued, and perhaps even increased, importance (Deming 2017, Edin et al. 2022). Interventions for non-cognitive skills are widely popular and their effectiveness has been shown in various contexts (van Agteren et al. 2021). Elements of positive psychology, mindfulness, emotional regulation, and many more have been shown to effectively improve well-being (Recabarren et al. 2019, Heintzelman et al. 2020, Seppälä et al. 2020). However, while several studies have shown the effectiveness of individual interventions, there is a lack of direct comparisons between different types of interventions and their effectiveness, making it difficult to determine optimal structures for interventions.

The question. Given the lack of evidence on optimal intervention components, the main research question is which intervention contents targeting emotional skills, and which combination of them, show the highest effectiveness in improving emotional skills and other well-being outcomes. In addition, I examine the importance of the Big Five personality traits in determining intervention outcomes. Lastly, I want to determine how different types of effect size measurements affect the results.

The method. In this study, I am analyzing the effects of a course called TESC (Training of Emotional and Social Competencies), an intervention conducted at the Johannes Gutenberg University in Mainz with students in the winter semester 2023/2024. TESC was designed as a factorial intervention for emotional skills, following the approach suggested by Collins et al. 2014, to enable the analysis of the individual intervention components¹. I evaluate the effectiveness of the four emotional skill interventions done in the course: mindfulness, emotional regulation, self-acceptance, and resource activation. To evaluate their effectiveness, I employ a combination of methods. First, I apply standard effect sizes like simple regressions and Hedges' G. Second, I use ANOVA, which is a common approach for analyzing factorial experiments. Finally, I implement a regression-based approach. The use of the different methods serves the purpose of enabling a comparison of the results and allowing to draw conclusions as to which measurement is most suitable in this context. In the regression analysis, I additionally control for the Big Five personality traits that were included in the questionnaire.

The findings. The findings can be divided into two categories. The first category is methodological findings, resulting from the comparison of the different methods. First, it

¹The course was organized by Sarah Tran-Huu and Klaus Wälde at Johannes Gutenberg University Mainz.

is shown that the standard effect size measures Hedges' G and a standardized regression yield similar results, but have a number of limitations. Therefore, if one chooses to report these measures, it could be useful to supplement them in combination with other measures, such as the probability of superiority. Second, it is shown that, for factorial interventions, including all treatments simultaneously yields different results than the previous measurements. This is shown using factorial ANOVA. Third, the biggest difference in results occurs when switching from the previous measures to a regression approach, where the initial level of the outcome variables is directly accounted for in the regression. This shows that a side-by-side comparison of before and after effect sizes or ANOVA does not yield the same results as a regression that controls for the initial outcome levels directly. It is therefore sensible to conclude that a regression is the better option to evaluate treatment effects. The second category is content-specific takeaways. Unfortunately, TESC experienced a substantial drop-out rate, which limits the ability to draw conclusions about the effectiveness of individual components. However, two effects remained significant throughout all measures: the effect of the resource activation treatment on a depression-related score and on stress. Possible explanations for the lack of significant results for the other treatment components are discussed in Section 7.3. The analysis also shows that the Big Five personality traits, especially conscientiousness and neuroticism, play a significant role in determining outcome variables, even when controlling for initial skills, which suggests that personality affects the success of an intervention.

The contents. In section 2, I present the TESC study in detail, including the contents of the study, the collected data, the factorial design, and the participant allocation process. In section 3, I present some descriptive statistics concerning the demographics of the participants, the drop-outs over the duration of the course, as well as the average change in skills over the duration of the course. In section 4, I discuss the first results, namely the treatment effects as measured with effect sizes through a simple regression, a standardized regression, Hedges' G , and the probability of superiority, and discuss the similarities and differences in the results. In section 5, I analyze the effects of the course using ANOVA and compare to the previous results. In section 6, I analyze the effects of the course in a regression set-up and therefore extend the previous measures. I include covariates, such as the Big Five, apply group-clustered errors, and correct for drop-out self-selection with inverse probability weighting. Section 7 discusses the most important takeaways from this study, divided into methodological and content-related takeaways and mentions limitations. Section 8 concludes.

2 Study design and allocation of participants

In this section, the details of the study will be discussed. In section 2.1 the contents of the course and the questionnaire, as well as the factorial study design and the timing of the course will be presented. In section 2.2 the allocation of participants with stratified randomization is explained.

2.1 Study design

TESC was set up as a voluntary university course at Johannes Gutenberg University in Mainz. Students from all faculties and all stages of study were notified of the course via a university-wide e-mail and were then able to apply. The course was held during the winter semester 2023/2024 through weekly 90-minute sessions and three full-day block seminars on weekends. The total duration of the course was 14 weeks. In the following subsections, the content and structure of the course, as well as the allocation process and the measured outcome variables, will be discussed.

2.1.1 Contents, psychological model and questionnaire

The course consisted of four main intervention components: mindfulness (MI), emotional regulation (ER), self-acceptance (SA), and resource activation (RA)². In short, mindfulness means paying attention to the present moment without judging it (Kabat-Zinn 2015). Emotional regulation describes any attempt to influence the emotions one has, including how one experiences or expresses them (Gross 2015). Self-acceptance means acknowledging and accepting oneself as one is, including good and bad qualities (Ryff 1989). Lastly, resource activation describes the focus on goals, strengths, social relationships, values, and other resources (Gassmann and Grawe 2006, Tran-Huu 2023, Tran-Huu 2025).

Figure 1 describes the main assumed relationships between the intervention component and the outcome variables. The first column shows the four intervention components: MI, ER, SA, and RA. In the second column are the primary outcomes, i.e., the skills and competencies targeted by the intervention components. The arrows connect the intervention components directly with the targeted outcome measures, respectively. As visible in the figure, the outcomes related to mindfulness (such as mindfulness and self-compassion) are targeted by the mindfulness component. The outcome variable emotional regulation is targeted by the intervention component emotional regulation, and so forth. Naturally, the intervention components can also affect the other primary outcomes. The arrows merely connect the intervention components with their directly related outcomes.

Depicted in the third column are the secondary outcomes that are not assumed to be directly related to one specific component. They include mental health outcomes, such as depression, a general outcome variable for subjective well-being, and stress. Subjective well-being outcome variables are commonly used in meta-analyses to assess the effect of psychological interventions (van Agteren et al. 2021).

Figure 1 simultaneously serves as an overview of the measured outcome variables. The participating students were asked to fill out questionnaires at the beginning, in the middle, and at

²For further elaboration on the contents, see Tran-Huu 2023 and Tran-Huu 2025.

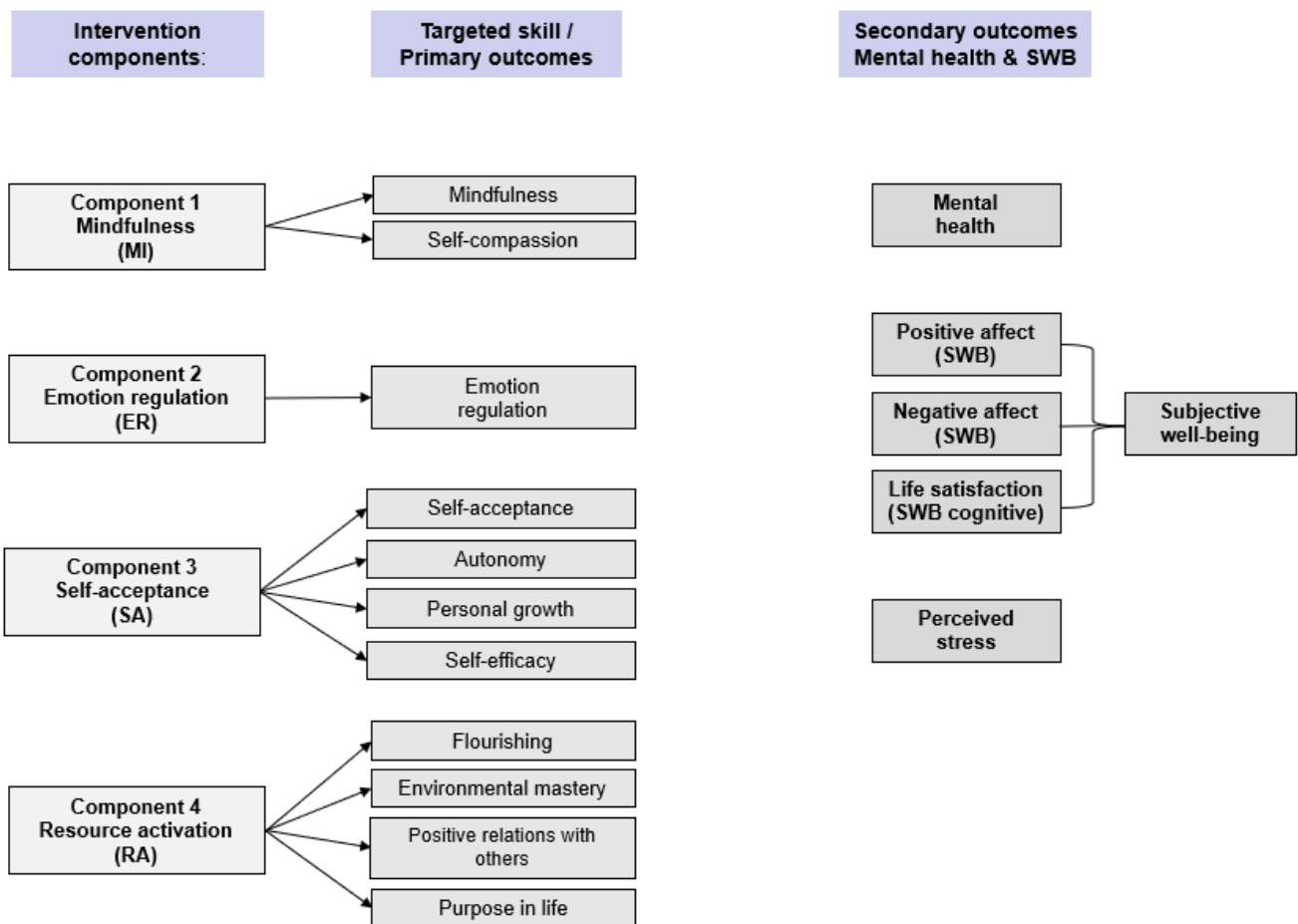


Figure 1: Intervention components and outcome variables

the end of the course. The outcome variables visible in Figure 1 were all part of the questionnaires.

Besides the outcome variables shown in Figure 1, an additional questionnaire included questions on demographic variables, the financial situation of the students, and most importantly, measures for the Big Five personality traits as known from Costa Jr and McCrae 1992³.

2.1.2 Timing and factorial design

The timeline of the study is depicted in figure 2. Students applied to the course via an application questionnaire and were allocated via stratified randomization in two waves, as will be explained in section 2.2. The questionnaires, entailing the questions on the outcome variables, were sent out in November 2023, January 2024, and February 2024, i.e., at the beginning, in the middle, and at the end of the course⁴. The additional questionnaire entailing the questions on the Big Five personality traits and additional demographic information was sent out in January 2024⁵.

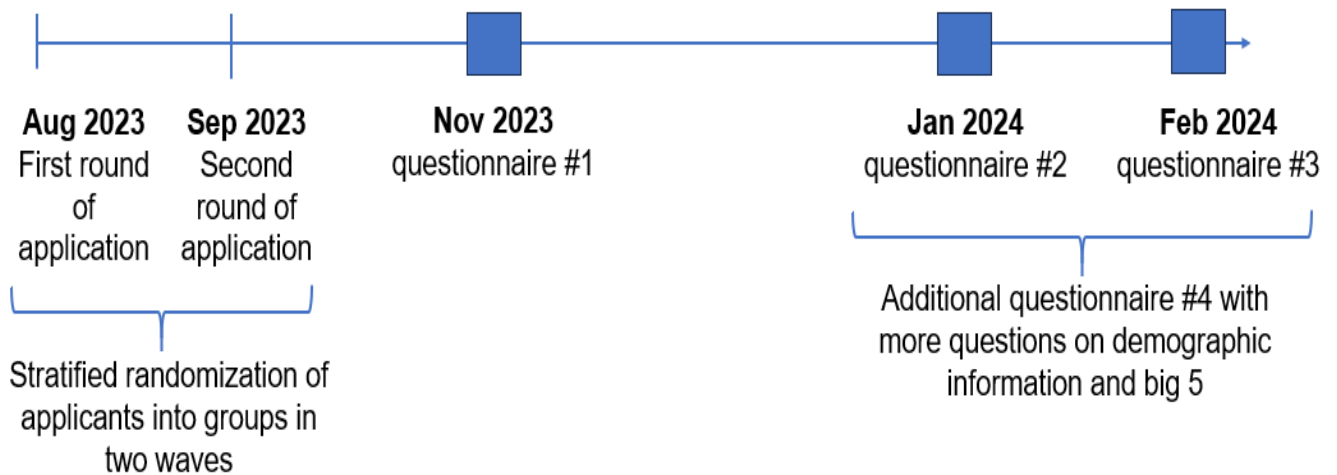


Figure 2: Timeline

Let me now focus on the factorial design and how the contents were implemented in this factorial set-up. Collins et al. 2014 suggested factorial designs as a way to measure the efficiency of individual intervention components for several reasons. First, the number of participants needed is comparatively low, which is practical for the implementation of the study. The reason for the reduced number of participants is that study participants serve as treatment and control groups simultaneously, depending on which treatment factor is being looked at. Secondly, it is possible to compare several intervention components to one another in a factorial set-up, which was one of the main goals of this project. This allows to study which types of intervention components are most effective in improving outcomes. Furthermore, there is no waiting group effect for the control group, as they are engaged in other activities. The factorial set-up of

³The Big Five are measured with the scales and questions from Schupp and Gerlitz 2014.

⁴Note that the first questionnaire was sent out in November, approximately 3-4 weeks after the course started.

⁵The timing of this questionnaire was not critical as the information asked in this questionnaire was assumed to be constant over the course of the semester.

TESC is described in figure 3.

Training of emotional and social skills

Content blocks:

Groups:

1	MI	ER	SA	RA	4 emotional skills components
2	MI	ER	SA	KO	3 emotional skills components
3	MI	ER	GR	RA	
4	MI	ST	SA	RA	
5	KL	ER	SA	RA	
6	MI	ER	GR	KO	2 emotional skills components
7	MI	ST	SA	GR	
8	MI	KL	ST	RA	
9	KL	ER	SA	GR	
10	KO	ER	ST	RA	
11	KO	KL	SA	RA	
12	MI	KO	KL	ST	1 emotional skills component
13	GR	ER	KL	ST	
14	GR	KO	SA	KL	
15	ST	GR	KO	RA	
16	ST	GR	KO	KL	0 emotional skills components

Figure 3: Factorial set-up

TESC included four factors (the four emotional skill intervention components described previously): mindfulness, emotional regulation, self-acceptance, and resource activation. The rows in figure 3 depict the 16 distinct groups that were taught. In each row/group, it is depicted what contents were covered for the duration of the course. Since there were four intervention components (MI, ER, SA, and RA) that each could have the status intervention or control, this gives $2 \times 2 \times 2 \times 2 = 16$ different combinations of the components. This translates into 16 different groups that did the TESC course. The rows therefore describe the contents that the respective group covered, i.e., for each factor, whether the group did intervention or control. The four intervention components/factors are depicted in one column each. The first column shows whether a group did mindfulness intervention or control, the second column shows whether a group did emotional regulation intervention or control, the third column shows whether a group did self-acceptance intervention or control, and the fourth column shows whether a group did resource activation intervention or control. A blue cell means the intervention component was implemented, and a white cell means a control component was implemented. The control components, the white cells in the graphic, were social skill contents, namely coping with stress (ST), conflict management (KL), communication (KO), and group dynamics (GR)⁶.

In general, the factorial set-up allows for comparing all those who did a specific intervention (the blue cells within a column) against those who did not do that specific intervention (the

⁶See Tran-Huu 2023 for more information on the control components.

white cells within a column). This is how the participants constitute their own control groups.

It is also important to note that the control components are not always the same in each column. For example, when comparing those who did the mindfulness intervention to those who did not, two control groups did conflict management, two did communication, two did coping with stress, and two did group dynamics. The same is true when comparing any of the other three intervention components⁷.

The columns also depict a timeline in the sense that the content blocks within each group were conducted from column 1 to column 4. At the beginning of the semester, either mindfulness or a control component was done, then either emotional regulation or a control component was done, then either self-acceptance or a control component was done, and lastly either resource activation or a control component was done.

2.2 Allocating participants with stratified randomization

The concept of stratified randomization is gaining more and more popularity in economic research. Duflo, Glennerster, and Kremer 2007, Bai 2022, Bruhn and McKenzie 2009, and Firpo, Foguel, and Jales 2020, for example, discuss the benefits and prevalence of stratified randomization in economic research. The idea is that researchers define strata (stratification characteristics) that could potentially influence the outcome variables. Then, researchers randomly assign participants to treatment and control groups within the strata. The goal is for the stratification characteristics to be balanced across treatment and control groups. Given that strata are defined based on which variables could potentially affect outcome measures, this achieves a pre-treatment balance in characteristics that is desirable for post-treatment analysis.

For the TESC course, students applied with a questionnaire that included all stratification characteristics, so that they could be taken into account for group allocation. As visible in figure 2, due to university structures, there were two rounds of applications and therefore two rounds of stratified randomization to sort the students into groups. Besides the stratification characteristics, students could also indicate time availability during the semester, which had to be taken into account for the allocation into groups, since there were 16 different groups with different schedules and the course had to fit into the semester schedule of the individual student⁸.

⁷This was done in order to avoid creating a situation where mindfulness intervention could only be compared against a specific control content.

⁸In particular, the students were allocated into groups within their 4 most preferred time-slots, in order to avoid allocation of a student who would have to cancel later on due to overlap with other classes.

The stratification characteristics were: gender, a broad categorization into faculties⁹, whether students had any mental health struggles¹⁰ and whether the students had previous therapy experience¹¹. The characteristics mentioned above could potentially influence outcome variables, which is why they were chosen as stratification characteristics. After each round of application as shown in figure 2, students were then randomly selected into groups, taking into account the stratification characteristics and the preferences over time availability. The 16 groups were filled up in a total of three rounds, once after each application deadline and a third, small round to fill up places of students who de-registered shortly before the beginning of the semester.

To check whether balancing the stratification characteristics was a success, I ran a linear regression from the treatment dummies on the stratification characteristics¹². The regression results are depicted in Table 1 and are, as expected, not significant. The lack of significance proves that the stratified randomization sufficiently balanced the characteristics of participants over the respective treatment and control groups.

Treatment	Gender	Faculty 2	Faculty 3	MINT faculties	MHI	Therapy
Mindfulness	0.0691	0.0243	-0.0075	-0.0695	-0.0427	0.0845
Emotional regulation	-0.0259	-0.0445	0.0720	0.0574	0.0120	0.0391
Self-acceptance	0.0120	0.0103	-0.0086	0.0816	-0.0142	0.0114
Resource activation	-0.0395	-0.0067	0.0782	0.0796	0.0541	0.0355

*Table 1: Balancing check after stratification. Linear regression of the treatment dummies on the stratification characteristics, including dummies for the faculties, a mental health dummy, and a therapy experience dummy. Significances are indicated as follows: * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$*

3 Descriptive statistics

In this section, descriptive statistics will be presented. In section 3.1, some demographic characteristics of all registered participants are shown. In section 3.2, the overall attendance and drop-out over the course are presented. Section 3.3 gives a first look at the average change in outcomes over the duration of the course.

⁹Faculties were divided into the two faculties with the most participants (faculties 2 and 3), MINT faculties, and other faculties. Using faculty as a stratification characteristic also made sense so that the probability of participants within one group knowing each other from their studies is reduced.

¹⁰such as issues with anxiety, depressive disorder, etc.

¹¹The reason for including this is that students with previous therapy experience might respond differently, as it is possible similar exercises were done in therapy.

¹²In detail, the regression looked as follows: $D_i = \beta_0 \text{Gender}_i + \beta_1 \text{Faculty}_2_i + \beta_3 \text{Faculty}_3_i + \beta_4 \text{Faculty_MINT}_i + \beta_5 \text{MHI}_i + \beta_6 \text{Therapy}_i$. It included the gender dummy, three faculty dummies for four faculty subdivisions, a mental health indicator dummy, and a previous therapy experience dummy. I ran four regressions (one for each treatment component), where the explanatory variable was an intervention/control dummy.

3.1 Demographic characteristics

To better understand the sample of participants, the following figures describe some demographic characteristics. The figures describe all the participants who registered to take the course¹³.

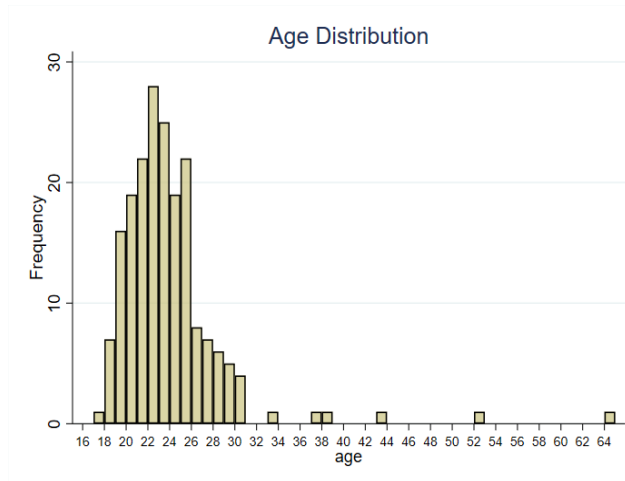


Figure 4: Age distribution of participants

Figure 4 shows the age distribution of participants. Since the intervention was conducted on a student population, the majority of individuals are between 19 and 27 years old. As expected, there are also some participants in their late 20s and early 30s, as well as some outliers up to age 64.

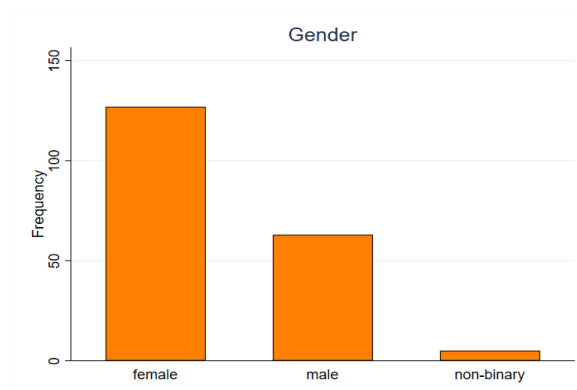


Figure 5: Gender distribution of participants

Figure 5 describes the gender distribution of the participants. As is common for voluntary courses such as TESC, there are more female registered students than male registered students, and only a few non-binary participants.

¹³i.e., the final allocation at the beginning of the semester

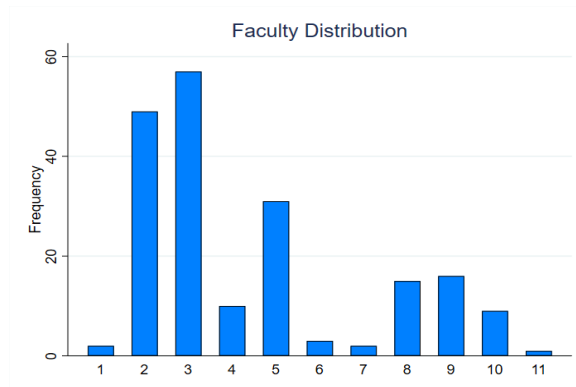


Figure 6: Faculty distribution of participants

Lastly, Figure 6 describes from which faculties the students are¹⁴. Most students come from faculty 2, 3, and 5, which can all be summarized under social sciences. This also includes students studying to become teachers (faculty 2). In comparison, fewer people came from MINT faculties (faculties 8,9, and 10)¹⁵. The rest of the students come from faculties with a generally low number of students.

3.2 Attendance and drop-outs over time

Next, I want to look at the attendance of the participants and the overall drop-out rate. Figure 7 shows the attendance over time in percent. Since there were 16 courses and each course had 17 course dates in total, there are 16 lines in the graph, each line representing a course. The thick black line shows the overall attendance. Overall attendance started with about 80% of registered students and slowly declined. In the last appointment, almost 60% of students attended on average. Since this was a voluntary course, this decrease in attendance is not unusual.

The drop-out in attendance is also visible in the questionnaires. The general questionnaire with outcome measures described in section 2.1.1, was sent out three times during the course. In November, after the course started, in January (in the middle of the course), and in February, directly after the course ended. Additionally, there was one questionnaire that included measures on several demographic characteristics and the Big Five personality traits that was sent out once (which is important as these measures serve as covariates in the regression later on). Since the only way to quantify the results is through the questionnaires, the drop-out rate of the questionnaires is of more relevance than attendance for the analyses. Therefore, figure 8 shows the drop-outs of the questionnaires. The figure, similar to the previous figure, shows a huge dropout rate over the course of the semester. The drop-out rate is even larger than the drop-out

¹⁴The faculties are as follows. Faculty 1: catholic and protestant theology. Faculty 2: Social sciences, media and sports (includes all students studying to become teachers). Faculty 3: law, economics, and business administration. Faculty 4: Medicine. Faculty 5: Philosophy and philology. Faculty 6: translation sciences, linguistics and cultural studies. Faculty 7: history and cultural studies. Faculty 8: Physics, mathematics, and computer science. Faculty 9: Chemistry, pharmacy, geography, and geoscience. Faculty 10: Biology. Faculty 11: Music academy.

¹⁵One has to take into account, however, that in general there are fewer students in the MINT faculties than in other faculties.

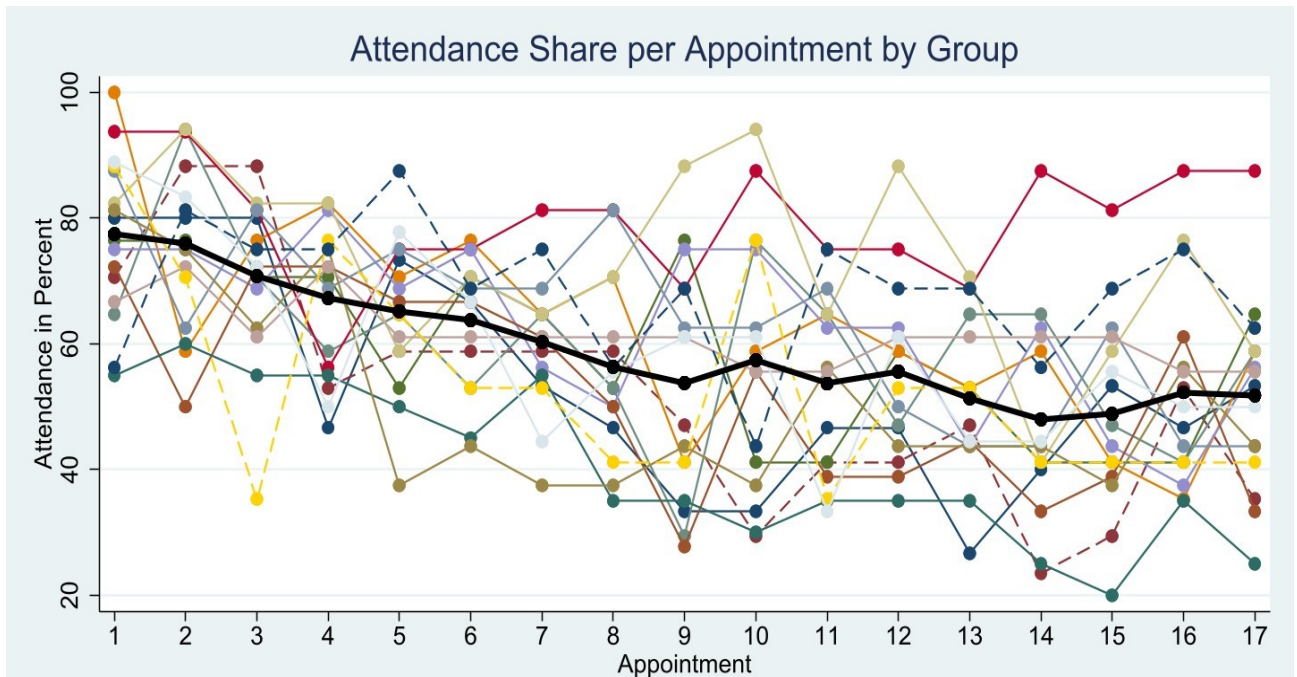


Figure 7: Attendance of participants

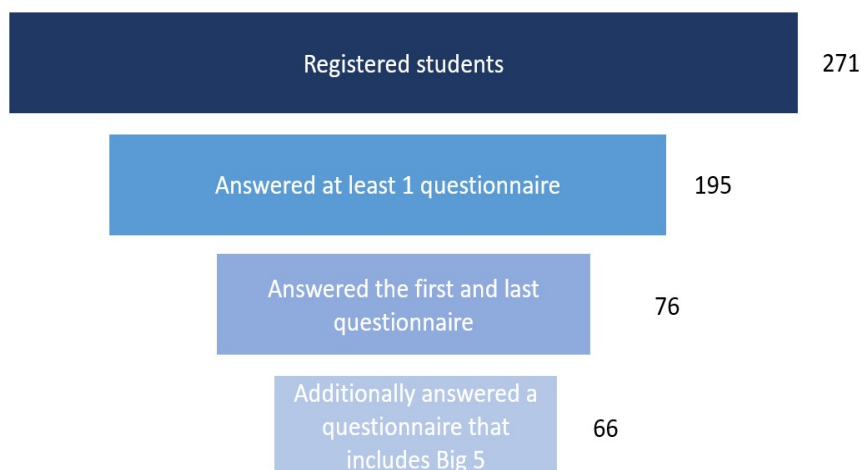


Figure 8: Drop-outs of the questionnaires

in attendance, as only 24% - 28% of all registered students answered all questionnaires¹⁶, while the attendance in the courses was still at almost 60%. This could depict a "questionnaire fatigue", as the questionnaires were quite long due to the number of outcome variables measured. The overall interest in remaining in the course was therefore higher than the questionnaire numbers suggest.

3.3 Average change in skills

Since the questionnaire involved various variables, I next want to look at the average changes in the outcome variables throughout the intervention to see whether there was any change during the duration of the intervention. For these descriptive statistics, they are not broken down by treatment but rather show the general trend. This is due to the 4 treatment components (and 16 combinations in total), which would make interpreting graphs broken down by treatment difficult to interpret. Further, the main focus here is to observe a general trend in all courses. The primary outcomes, as described in section 2.1.1 can be subcategorized into skills targeted by the mindfulness, emotional regulation, self-acceptance, and resource activation components. Secondary outcome measures are not related to any one component in particular. The mean changes are described in figures 9 - 12. Note that the variables were first standardized as described in section 4.1.2, therefore, the scale on the vertical axis is to be interpreted accordingly. The horizontal axis has only two points in time, where $time = 0$ stands for the measurement in November and $time = 1$ stands for the measurement in February, after the course ended. Figure 9 includes the average change in all mindfulness and emotional regulation measures, figure 11 includes the average change in all self-acceptance measures, figure 10 includes the average change in all resource activation measures and figure 12 includes the average change in all secondary outcome variables. The means at $time = 0$ and $time = 1$ include 95% confidence intervals.

The general trend that can be observed is that the mean changes in skills are predominantly positive. Note that for outcomes stress and depressive disorder, lower values are better. However, this should be interpreted with caution. The confidence intervals are often large compared to the changes in the variables themselves (i.e., the confidence intervals often overlap), which points to no significant change in means. In addition, the graphs show only the overall trend across all groups without taking into account one specific control group to compare the change to. The trend can be described as neutral to positive for almost all outcome variables except purpose and depressive disorder, where it is neutral to negative. Although it is not possible to make causal statements about the effect of the course, it allows us to see the general trends in outcome variables for the duration of the course. Looking at the figures, there seems to be a slight positive trend over time. Although it can not be measured or quantified in this study, the positive trend is there despite the fact that students often report higher stress and higher mental loads toward the end of the semester, which is where exams usually take place.

¹⁶depending on whether the additional questionnaire is included or not



Figure 9: Mean effects of MI and ER targeted skills

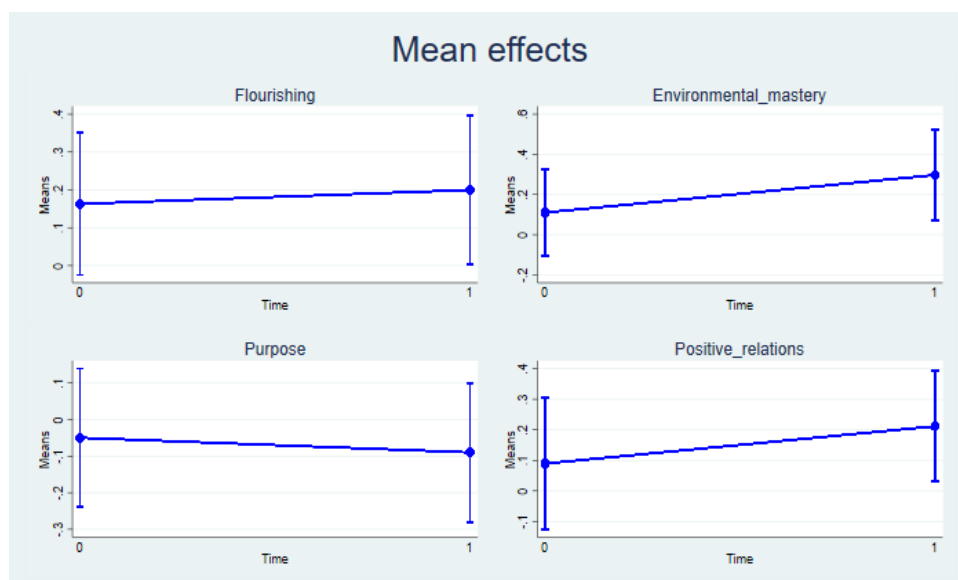


Figure 10: Mean effects of RA targeted skills

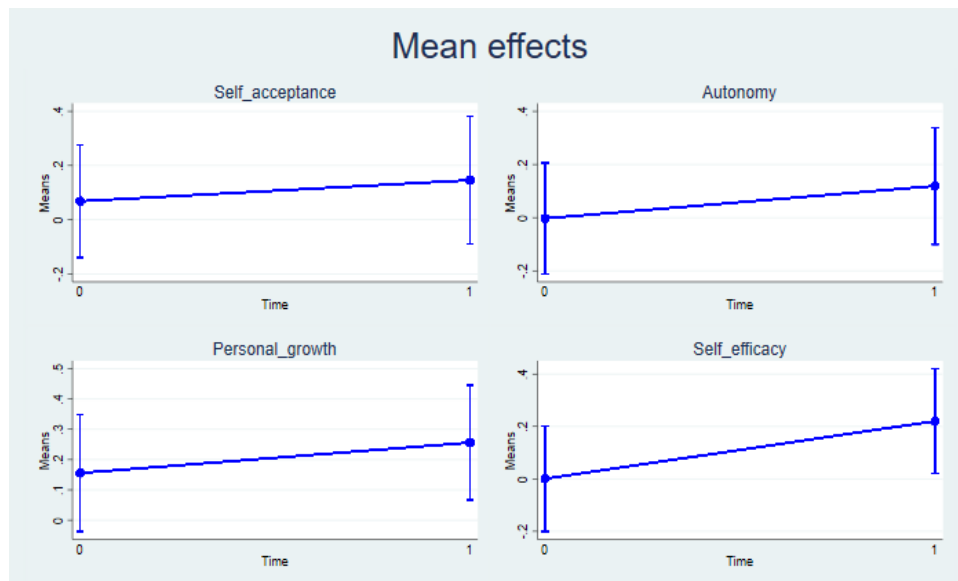


Figure 11: Mean effects of SA targeted skills

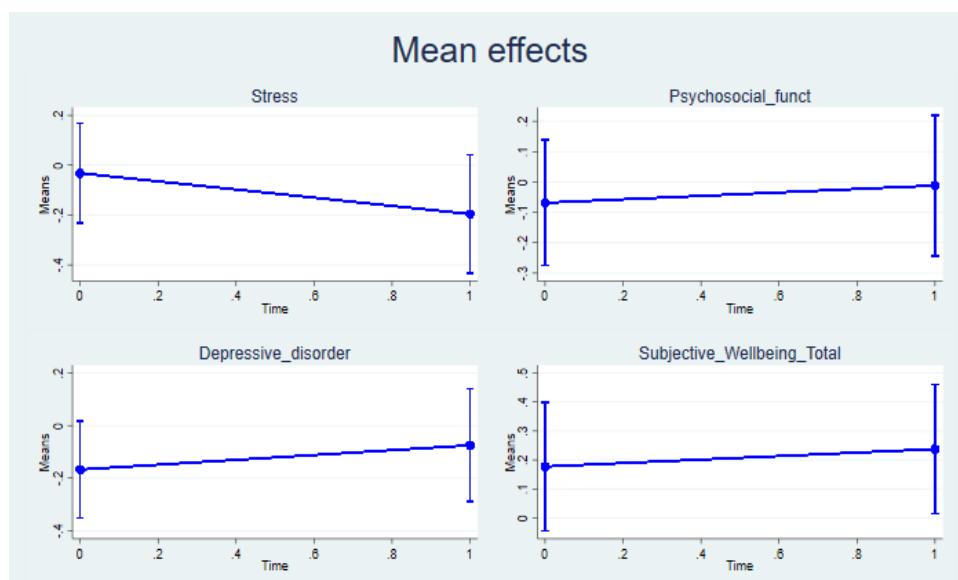


Figure 12: Mean effects of secondary outcome measures

4 Effect size measures compared

In this section, I compare effect size measures and their resulting measured treatment effects against each other. Namely, I compare a simple and a standardized regression with Hedges' G and the probability of superiority. In section 4.1, the choice of measurements is explained, and in section 4.2, the results are compared side by side.

4.1 Selection of measurements

Treatment effects can be measured in various ways. The standard economic approach to measure treatment effects is a regression, which is the first measurement. Here, I implemented one regression in the scales of the questionnaires (section 4.1.1) and one standardized regression (section 4.1.2) to facilitate interpretation of the coefficients. In order to be able to compare the results with other studies on emotional skill interventions, which are usually published in psychological journals, I additionally chose Hedges' G (section 4.1.3). Lastly, I chose another effect size measure that does not focus on the mean difference, the probability of superiority, presented in section 4.1.4. By not evaluating the mean difference, it circumvents the homogeneity assumption, as explained in section 4.1.3 and 4.1.4.

4.1.1 Measure 1: The simple regression

For the simple regression in levels, the analysis consists of a standard, robust dummy regression, where the dummy variable indicates whether a participant has received treatment. Since I want to compare the results to other effect size measures that can only handle one treatment at a time and look at outcomes at one point in time, this is also how the simple regression is set up:

$$Y_i^t = \beta_0 + \beta_1 D_i^j + \varepsilon_i \quad (1)$$

where Y_i^t stands for the score of an outcome variable of person i at time t (where t can be *nov* and *feb*, indicating measures of the outcome variable in the first and last questionnaire, respectively). D_j^i can be the MI, ER, SA, or RA treatment dummy. Therefore, there are eight regressions for each outcome variable, one for each treatment, and for two points in time, November and February.

4.1.2 Measure 2: The standardized regression

Standardizing the outcome variables makes sense because different outcome variables come from different questionnaires with different scales, which makes the comparison of treatment coefficients complicated. Standardizing makes side-by-side comparison easier. The standardization works as follows:

$$\tilde{Y}_i^t = \frac{Y_i^t - \mu_{nov}}{\sigma_{nov}} \quad (2)$$

Y_i^t is the outcome variable at time t , where t can be *nov* as well as *feb*. Also, μ_{nov} is the mean over participants of the outcome variable in November, and σ_{nov} is the standard deviation of the outcome variable in November. Therefore, the variables are all standardized with the mean and standard deviation of the same variable from the first questionnaire (the beginning of the intervention). The reason for using the November values to standardize variables of both points in time is not to "lose" average changes over all participants during the intervention. The general setup remains the same as in the previous measure, but the size of the coefficient is now measured in the unit of a standard deviation. The regression looks as follows:

$$\tilde{Y}_i^t = \beta_0 + \beta_1 D_i^j + \varepsilon_i . \quad (3)$$

where the only difference to the previous regression is the standardized outcome variable.

4.1.3 Measure 3: Hedges' G

Hedges' G and Cohen's D are effect size measures often used to measure the effectiveness of psychological interventions, especially in meta-analysis comparisons (see e.g. van Agteren et al. 2021, Hodzic et al. 2018 and Heintzelman et al. 2020). The two measures essentially measure a standardized mean difference between two groups (in this case, treatment and control groups). Hedges' G, formulated by Hedges 1981, in essence corrects Cohen's D for small sample sizes, as the latter tends to overestimate effects for small sample sizes. As the sample size grows, the two effect sizes grow increasingly closer to each other. For that reason I decided to only report Hedges' G in this analysis. The standardized measure facilitates comparison of results across studies, and since it also corrects for smaller sample sizes, it is favorable for the comparison of studies with varying sample sizes. Interpretation is fairly easy and is commonly classified as small ($G \approx 0.2$), moderate ($G \approx 0.5$), and large ($G \approx 0.8$) effects, as suggested by Cohen 1988¹⁷. Hedges' G is calculated as follows:

$$G = \frac{\bar{Y}_1 - \bar{Y}_0}{s^*} c(m) \quad (4)$$

where

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}} \quad (5)$$

$$c(m) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{\frac{m}{2}} \Gamma\left(\frac{m-1}{2}\right)} \quad (6)$$

¹⁷This division in effect sizes was originally made for behavioral/psychological studies, therefore, this classification seems appropriate for an emotional skill intervention. It should be noted that since Cohen 1988, the classification into effects has been criticized at times and some suggest even lower thresholds for the effect sizes (Gignac and Szodorai 2016).

and

$$m = n_0 + n_1 - 2 .$$

Note that the first part of the Hedges' G formula is the Cohen's D formula, and the $c(m)$ expression is the correction for sample size, as explained in Hedges 1981. Analogously to the previous two measures, I also calculated eight values for Hedges' G for each outcome variable, as the distinction in treatment ($\bar{Y}_1$) and control ($\bar{Y}_0$) depends on the definition of whether MI, ER, SA or RA is viewed as the treatment and can only be measured at one point in time (November and February), respectively.

Looking at the first part, s^* is simply the pooled standard deviation. Therefore, similar to the standardized regression, Hedges' G is interpreted in units of the standard deviation. It normalizes the difference between the intervention and the control groups and adjusts for small numbers of participants¹⁸. The disadvantage of effect size measures such as Cohen's D and Hedges' G is that they can not incorporate alterations or covariates like regressions can. One of the assumptions of Hedges' G, for example, is homogeneity, which is a strong assumption that often does not hold in practice. This homogeneity assumption holds for most effect sizes that center on the difference of means (Erceg-Hurn and Mirosevich 2008, Grissom and Kim 2001). For this reason, I also included another effect size, the probability of superiority.

4.1.4 Measure 4: Probability of superiority

The probability of superiority measure, as suggested by Grissom and Kim 2001, is also an effect size measure. But contrary to others, it does not center on the difference of means, but rather on a probability. It estimates the probability that, if I draw a random person from the intervention group, this person is better off (in terms of the outcome variable) than a randomly drawn person from the control group. It is calculated via pairwise comparison (see Erceg-Hurn and Mirosevich 2008):

$$PS = \frac{U}{mn} \quad (7)$$

where U is the Mann-Whitney statistic and m and n are the sample sizes. The Mann-Whitney statistic gives the number of times that the m subjects given treatment outrank the n subjects

¹⁸More specifically, Hedges' G first takes the difference between sample means and then normalizes the difference (and then corrects for small sample size). In the standardized regression in section 4.1.2, the outcome variables are first standardized, and then the group difference is assessed through the coefficient. So, in essence, Hedges' G takes the difference first and standardizes later, while the regression standardizes first and then takes the difference.

given control. The division by mn is the number of possible comparisons. The statistic looks as follows:

$$U = \sum_{i=1}^m \sum_{j=1}^n S(y_i, y_j) \quad (8)$$

where

$$S(y_i, y_j) == 1 \quad \text{if } y_i > y_j \quad (9)$$

$$S(y_i, y_j) == 0 \quad \text{otherwise.} \quad (10)$$

As the formulas show, it is calculated via a pairwise comparison of the individuals receiving treatment and those receiving control¹⁹. For example, a *PS* of 0.7 would indicate that 70% of comparisons resulted in better scores for the treated individuals. In this case, values above 0.5 generally speak for a positive intervention effect. Here, I again calculate eight measures for each outcome variable, depending on the definition of the treatment, and two times for each point in time. The downside is that the probability of superiority generally only allows one to make a statement on *whether* the intervention had an effect, not *how big* this effect is. But, due to the setup and the lack of variance in the calculation, the probability of superiority makes no assumptions about homogeneity²⁰. This may also be an argument that the two effect sizes, Hedges' *G* and the probability of superiority, are suitable to complement each other.

4.2 Effect size measure results

As explained in the previous section, the effect size measures can only look at one treatment at a time, at one point in time. Therefore, for the following tables, there are always four results per outcome variable for treatment versus control, namely the four treatments MI, ER, SA, and RA. The effect sizes for the primary outcomes in February can be found in Table 2 and for the secondary outcomes in February in Table 3. The respective tables for November can be found in appendix A.1.

Let me focus on the primary outcome measures first. Looking at table 2, the first column shows the outcome variable, and in the brackets, it shows which treatment is being looked at. The second column is the result from the simple regression and the third column is from the standardized regression. 10, 5, and 1% significance levels are labeled by one, two, or three stars, respectively. Since the standardized regression is simply a rescaling of the simple regression, there is no change in significances here, only in the coefficients themselves and the way to interpret them. The fourth column shows Hedges' *G*. Hedges' *G* does not come with a p-value

¹⁹Note that the exact calculation varies as sometimes a draw (i.e., $y_i = y_j$) is 0, as in my case, and sometimes it is 0.5.

²⁰The calculation of the confidence interval of the probability of superiority, however, sometimes assumes homogeneity due to the method of calculation. To circumvent this problem, I used a bootstrapping method to calculate the confidence intervals.

specifically, so if the 95% confidence interval was above zero (or below zero), the stars have been added to Hedges' G to facilitate interpretation. The last column gives the probability of superiority. This is the only measure where the coefficient is interpreted differently, namely as described in section 4.1.4.

Outcome variable (Treatment)	Reg.	Norm. Reg.	Hedges G	CI Low	CI High	Prob. Superiority
Mindfulness (MI)	-0.138	-0.186	-0.175	-0.618	0.269	0.456
Mindfulness (ER)	0.014	0.019	0.018	-0.429	0.465	0.497
Mindfulness (SA)	0.187	0.251	0.237	-0.208	0.681	0.559
Mindfulness (RA)	0.161	0.215	0.203	-0.244	0.649	0.560
Self-compassion (MI)	-0.041	-0.059	-0.059	-0.501	0.384	0.503
Self-compassion (ER)	0.370**	0.540**	0.555**	0.097	1.008	0.670**
Self-compassion (SA)	0.227	0.331	0.331	-0.115	0.776	0.582
Self-compassion (RA)	0.247	0.361	0.362	-0.088	0.810	0.612*
Emotional regulation (MI)	-0.037	-0.061	-0.060	-0.503	0.383	0.478
Emotional regulation (ER)	0.173	0.285	0.286	-0.164	0.734	0.583
Emotional regulation (SA)	0.239*	0.395*	0.400	-0.049	0.846	0.618*
Emotional regulation (RA)	0.164	0.271	0.272	-0.176	0.718	0.545
Self-acceptance (MI)	0.720	0.205	0.196	-0.249	0.639	0.513
Self-acceptance (ER)	1.568*	0.447*	0.434	-0.020	0.884	0.570
Self-acceptance (SA)	2.193***	0.626***	0.622**	0.166	1.073	0.611*
Self-acceptance (RA)	-0.239	-0.068	-0.065	-0.510	0.381	0.439
Autonomy (MI)	0.200	0.055	0.056	-0.386	0.499	0.480
Autonomy (ER)	0.773	0.214	0.219	-0.230	0.667	0.517
Autonomy (SA)	0.424	0.117	0.120	-0.323	0.563	0.491
Autonomy (RA)	0.499	0.138	0.141	-0.305	0.586	0.521
Personal growth (MI)	0.944*	0.336*	0.407	-0.042	0.853	0.567
Personal growth (ER)	0.750	0.267	0.321	-0.130	0.769	0.488
Personal growth (SA)	-0.007	-0.003	-0.003	-0.446	0.439	0.420
Personal growth (RA)	0.504	0.180	0.214	-0.233	0.660	0.473
Self-efficacy (MI)	-0.058	-0.011	-0.013	-0.455	0.430	0.440
Self-efficacy (ER)	1.136	0.224	0.252	-0.197	0.700	0.545
Self-efficacy (SA)	1.983**	0.391**	0.448**	-0.001	0.895	0.578
Self-efficacy (RA)	0.334	0.066	0.074	-0.372	0.519	0.488
Flourishing (MI)	-0.060	-0.008	-0.009	-0.452	0.433	0.468
Flourishing (ER)	3.848***	0.519***	0.621**	0.162	1.077	0.630**
Flourishing (SA)	3.234**	0.436**	0.515**	0.063	0.963	0.636**
Flourishing (RA)	0.709	0.096	0.109	-0.337	0.555	0.488
Environmental mastery (MI)	-0.463	-0.125	-0.123	-0.565	0.321	0.435
Environmental mastery (ER)	0.970	0.261	0.259	-0.190	0.707	0.537
Environmental mastery (SA)	2.336***	0.629***	0.653**	0.196	1.105	0.629**
Environmental mastery (RA)	-0.332	-0.090	-0.088	-0.533	0.358	0.437
Purpose (MI)	-0.192	-0.069	-0.081	-0.524	0.362	0.398
Purpose (ER)	-0.220	-0.079	-0.093	-0.540	0.354	0.446
Purpose (SA)	0.244	0.088	0.104	-0.340	0.546	0.480
Purpose (RA)	-0.365	-0.132	-0.155	-0.601	0.291	0.378*
Positive relations (MI)	-0.515	-0.135	-0.167	-0.610	0.277	0.397
Positive relations (ER)	0.970	0.254	0.318	-0.133	0.767	0.562
Positive relations (SA)	0.671	0.176	0.219	-0.226	0.662	0.520
Positive relations (RA)	0.142	0.037	0.046	-0.400	0.491	0.455

Table 2: Effect sizes of primary outcome variables for standard regression, standardized regression, Hedges G, and Probability of Superiority. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

The effect sizes in this table only compare the outcome variables in February for the treatment vs. the control group. To improve the informative value of these measures, they have to be compared to the same measures in November, depicted in Table A.10. For readability reasons, this table is available in appendix A.1. Comparing the table for February to the table for November allows to assess more accurately whether a significant difference between treatment

and control group in February was brought on during the intervention or was pre-existing before the course. Similarly, while Table 3 shows the February results for the secondary outcome variables, Table A.11 in appendix A.1 shows the November results for the secondary outcome variables.

A broad first look at tables 2 and 3 shows that the significances are often similar across the different measures (i.e., different columns), with one notable exception being the group difference of the self-acceptance treatment regarding the self-efficacy measure, where the probability of superiority is not significant. It is also noteworthy that, on average, significances from the regression and Hedges' G are very close, while for the probability of superiority, there are more instances with differing significances. This likely stems from the fact that the probability of superiority is calculated differently, and it is useful to interpret the results together (where a difference of significance could indicate ambiguity in the result).

Looking at changes that happened from November to February, let me first address significant changes for the primary outcome measures²¹. The group difference of self-acceptance treatment (vs. control) for (the outcome variable) self-acceptance has increased in significance and also in the size of the effect. The group difference of mindfulness treatment for personal growth is new, but negligible due to the weak to nonexistent significance. Group differences of emotional regulation treatment for personal growth that were previously only slightly significant for the regression and Hedges' G, vanished. A group difference of self-acceptance treatment for self-efficacy (visible for the regressions and Hedges' G) increased in size. The group difference of self-acceptance treatment for Flourishing has increased in size but not in significance (except for the probability of superiority) compared to November. The group difference of self-acceptance treatment for environmental mastery increased from no to 10% significance in November to 5 to 1% significance in February and doubled in size²². Lastly, a negative group difference of emotional regulation treatment for the outcome purpose was reversed and a minor group difference of self-acceptance treatment for purpose vanished. All other significances that were not mentioned were similar in November.

To underline the most important changes, the strongest changes for the primary variables are therefore the "effect" of SA treatment²³ on self-acceptance, self-efficacy, and environmental mastery and the reversion of a negative difference in the outcome measure purpose concerning the emotional regulation treatment. It appears from these results tables that the SA treatment has effects on the outcome variables most often. The size of the effects for the self-acceptance treatment affected outcomes, especially when looking at Hedges' G and its standard interpreta-

²¹Note that in the following paragraphs I only use the term "group differences" to describe differences between treatment and control groups, as the measures do not take into account changes over time and therefore I want to avoid using terms like "treatment effects".

²²comparing the standardized regressions

²³Effect in this case means there is a change from before to after treatment.

tion, can be categorized as moderate.

Let me turn to the secondary outcome measures in table 3. The respective table for November is table A.11 in appendix A.1.

Variable	Reg.	Norm. Reg.	Hedges G	CI Low	CI High	Prob. Superiority
Stress (MI)	2.485	0.392	0.370	-0.072	0.811	0.564
Stress (ER)	-2.321	-0.366	-0.345	-0.788	0.100	0.388*
Stress (SA)	-1.781	-0.281	-0.263	-0.701	0.176	0.401
Stress (RA)	-3.044**	-0.480**	-0.458**	-0.903	-0.009	0.359**
Psychosocial functionality (MI)	0.109	0.335	0.328	-0.119	0.773	0.557
Psychosocial functionality (ER)	0.016	0.049	0.048	-0.399	0.494	0.457
Psychosocial functionality (SA)	-0.091	-0.281	-0.274	-0.718	0.171	0.376*
Psychosocial functionality (RA)	0.033	0.101	0.097	-0.349	0.543	0.483
Depressive disorder (MI)	-0.021	-0.042	-0.044	-0.486	0.399	0.430
Depressive disorder (ER)	-0.086	-0.171	-0.178	-0.625	0.270	0.417
Depressive disorder (SA)	-0.141	-0.280	-0.294	-0.738	0.152	0.366*
Depressive disorder (RA)	-0.252**	-0.501**	-0.540**	-0.991	-0.084	0.339**
Subjective wellbeing (MI)	0.747	0.095	0.095	-0.348	0.538	0.546
Subjective wellbeing (ER)	1.765	0.226	0.227	-0.222	0.674	0.572
Subjective wellbeing (SA)	4.414**	0.564**	0.589**	1.034	1.039	0.659**
Subjective wellbeing (RA)	-0.778	-0.099	-0.099	-0.545	0.346	0.456

Table 3: Effect sizes of secondary outcome variables for standard regression, standardized regression, Hedges G, and Probability of Superiority. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

The most notable difference from November to February is a strong negative "effect" from resource activation treatment on stress and depressive disorder, visible in all measures. While there is no difference between the treatment and control groups visible in the November questionnaire, there is a difference in February. Also note that a negative effect in these measures is desirable since a higher stress and depressive disorder score equates to a worse well-being. The group difference of self-acceptance treatment for subjective well-being visible in table 3 was also there in November, although at a lower/no significance and a smaller coefficient. Lastly, a previously existing group difference in psychosocial functionality concerning MI treatment (not visible in the probability of superiority) is no longer present in February. From the secondary outcomes, resource activation (and potentially self-acceptance) has effects on well-being most often. In this case, the size of the effects can also be described as moderate.

Regarding the effects of resource activation treatment on stress and depressive disorder, the effects are stronger and more consistent than for the primary outcomes. For the primary outcomes, while we can observe some changes over time, they are on average slightly less coherent over effect size measures, but the effect of the self-acceptance treatment is significant most often. Still, at this point in time, it is not possible to say much about the validity of the results until further analyses have been conducted. Nevertheless, these tables give a first impression of the results.

Let me also quickly summarize the takeaways from observing the four different outcome measures side-by-side. In general, the (standardized) regression and Hedges' G give similar

results in terms of coefficient size and significance. The probability of superiority, while mostly consistent with the other measures, varies more often in significance from the other outcome measures, and tends to have a lower significance if it does. This could be an indicator that it is useful to include this effect size measure alongside the other measures as a robustness check. Regardless, for this study, looking at the results for these four effect size measures was a mere first step. The purpose of including these effect size measures was, on the one hand, to have a measurement that also appears in other psychological research to facilitate comparability and, on the other hand, to assess how much the results change when deviating from this simple approach, which will be done in the following sections.

5 ANOVA

In the next step, I wanted to use an approach that can handle more than one treatment at once. Collins et al. 2014 propose analyzing factorial trials with factorial ANOVA. In principle, factorial ANOVA allows one to look at all main effects (that is, all main effects of the factors MI, ER, SA, and RA) and all interaction effects between the different factors. Due to the small sample size, I only estimate the main effects here. Note that the results have to be interpreted differently from the previous coefficients. ANOVA reports the F-statistic, which is the ratio of the mean square between (MSB) and the mean square within (MSW). The mean square between measures the variability between groups, i.e., variability that we can explain. The mean square within measures the variability within groups, which represents random variability that we can not explain. Therefore, a higher F-statistic means more variability that we can explain with differences between groups. Still, the F-statistic does not allow estimation of the size of the effect in the same way the coefficients of a treatment dummy in a regression do.

Outcome Variable	F-statistic MI	F-statistic ER	F-statistic SA	F-statistic RA	F-statistic Model
Mindfulness	1.052	0.148	1.704	1.266	0.804
Self-compassion	0.002	4.327**	1.778	3.037*	2.586**
Emotional regulation	0.119	0.667	3.171*	2.024	1.546
Self-acceptance	1.106	2.991*	5.153**	0.016	2.760**
Autonomy	0.129	0.819	0.157	0.370	0.396
Personal growth	4.655**	3.576*	0.200	0.486	1.897
Self-efficacy	0.011	0.528	3.418*	0.297	1.193
Flourishing	0.063	5.536**	3.681*	0.446	2.928**
Environmental mastery	0.460	0.280	7.460***	0.000	2.275*
Positive relations	0.338	1.092	0.721	0.104	0.715
Purpose	0.225	0.312	0.259	0.306	0.256
Stress	3.120*	0.735	2.487	5.562**	2.899**
Subjective wellbeing	0.135	0.459	5.335**	0.039	1.773
Psychosocial functionality	2.880*	0.646	2.123	0.011	1.151
Depressive disorder	0.000	0.216	2.306	6.364**	2.138*

Table 4: *F-statistics for factorial ANOVA without interaction effects in February.* * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

Table 4 shows the results for factorial ANOVA in February without interaction effects. As in the previous section, ANOVA only accounts for one point in time, therefore, it is necessary to compare this table with the table in November to account for changes over time. The results for November can be found in table 5. When comparing the two and looking at significances

of at least 5%, there is an effect from MI on personal growth, from SA on self-acceptance, environmental mastery, and subjective well-being, and an effect of RA on stress and depressive disorder. Additionally, previously existing group differences concerning Purpose with respect to the factors ER and SA vanished, and group differences concerning psychosocial functionality with respect to the factor MI lessened.

Outcome Variable	F-statistic MI	F-statistic ER	F-statistic SA	F-statistic RA	F-statistic Model
Mindfulness	1.402	0.000	1.780	0.032	0.798
Self-compassion	1.449	5.049**	1.200	2.334	2.801**
Emotional regulation	0.663	0.729	1.831	0.286	0.949
Self-acceptance	0.323	2.416	2.020	0.153	1.449
Autonomy	2.735	2.594	0.555	0.007	1.155
Personal growth	1.267	3.639*	0.902	0.056	1.491
Self-efficacy	0.003	0.817	2.371	0.166	1.043
Flourishing	0.955	6.135**	2.442	0.081	2.678**
Environmental mastery	1.106	0.062	2.680	0.052	1.044
Positive relations	1.503	0.312	0.219	0.138	0.663
Purpose	0.243	7.021**	4.718**	0.079	2.577**
Stress	1.490	1.378	2.034	0.003	1.470
Subjective wellbeing	0.021	1.283	2.148	0.785	1.363
Psychosocial functionality	5.231**	0.407	1.944	0.083	1.620
Depressive disorder	0.037	0.709	0.329	0.539	0.423

Table 5: *F-statistics for factorial ANOVA without interaction effects in November.* * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

In addition to the F-statistics of the individual factors, ANOVA also reports the F-statistic of the model to assess whether the model as a whole (i.e., the subdivision into MI, ER, SA, and RA groups) explains a substantial part of the variation. This can help to evaluate whether the division in treatment groups/factors can explain group differences well. Looking at the F-statistic of the model, there are some outcome variables for which the F-statistic of the model is (newly) significant compared to November. For the outcome variables self-acceptance, and stress (and at the 10% level for environmental mastery and depressive disorder), the F-statistic of the model is newly significant after treatment.

Comparing the ANOVA significances to the previous results, they are, in general, similar, with a few varying results. For example, the effect of self-acceptance treatment on self-efficacy has only 10% significance in ANOVA and was not there at all in the November questionnaire. In the previous section, these significances were higher for all measures except for the probability of superiority. In contrast, the group differences in the personal growth measure concerning mindfulness treatment have 5% significance in ANOVA and only 0-10% significance previously. These are some examples of differences in the outcomes between the two sections. In general, the results are similar, but they differ for some values. While the arguably most important effects are visible in all measurements so far, this shows that including all treatment variables at the same time is important, as it changes some results (even if not all of the results are turned on their heads). So far, no one way to measure treatment can be considered superior. While ANOVA takes into account multiple treatment effects, the interpretation of the coefficient is not as straightforward, and ANOVA also assumes homogeneity. Considering that psychological studies also often report ANOVA, and that it was the recommended analysis for factorial trials by Collins et al. 2014, it is still useful to include the ANOVA results in this paper as

well. Nevertheless, for a more detailed analysis it makes sense to deviate from the simpler frameworks, such as effect size measures and ANOVA, and turn to the standard economic approach, regressions, to assess the impact of the treatments. The shortcomings of the previous measures can be circumvented with the following approach.

6 Extended regression approach

In this section, I will present the alternative to the previous approaches, namely an analysis of the outcomes in the form of a regression. The reason for switching to the regression approach is threefold. First, it is more common in economic research. Second, it allows for assessing the change over time rather than focusing on one point in time and having to compare measures. Third, it allows for the inclusion of all four treatments, covariates, and other specifications that will be explained in the following subsections. The simplest regression specification is presented in section 6.1, where I account for the November outcomes and include all treatment effects. The baseline regression can be found in section 6.2, which additionally accounts for covariates not included in the first one. The following two specifications serve as a robustness check and extension of the baseline regressions. Group structures are included in section 6.3 and self-selection is taken into account in section 6.4.

6.1 Including all four intervention components

The first regression specification is the "simplest" regression specification and can be used to compare the results to the previous measures. This regression includes all four treatment dummy variables and, most importantly, the initial score of the respective outcome variable in November. With this approach, the regression already accounts for a change over time. The regression looks as follows

$$\tilde{Y}_i^{j,feb} = \beta_0 + \beta_1 MI_i + \beta_2 ER_i + \beta_3 SA_i + \beta_4 RA_i + \beta_5 \tilde{Y}_i^{j,nov} + \varepsilon_i \quad (11)$$

Where the explanatory variable is $\tilde{Y}_i^{j,feb}$, which is the standardized outcome variable j of individual i in February. Included are the four treatment dummies and the initial score in November, $\tilde{Y}_i^{j,nov}$. Note that I used the standardized outcome variables. Including the initial score allows for accounting for differences over time, contrary to the previous approaches, where a direct inclusion of the November questionnaire answers was not possible. The regression results are summarized in Table 6.

Dependent Variable	MI	ER	SA	RA	Initial score
Mindfulness	-0.070	-0.168	-0.051	0.213	0.901***
Self-compassion	0.253**	0.157	-0.000	0.030	0.791***
Emotional regulation	0.003	0.075	0.201	0.225	0.613***
Self-acceptance	0.121	0.162	0.119	0.010	0.861***
Autonomy	-0.116	-0.044	0.216	0.244	0.681***
Personal growth	0.276*	0.162	-0.222	0.039	0.547***
Self-efficacy	-0.026	0.049	0.068	0.156	0.712***
Flourishing	-0.047	0.230	0.082	0.135	0.636***
Environmental mastery	0.124	0.155	0.226	0.013	0.803***
Positive relations	0.103	0.201	0.041	0.134	0.531***
Purpose	-0.010	0.245	-0.122	-0.117	0.635***
Stress	0.154	-0.112	-0.146	-0.565***	0.689***
Subjective wellbeing	0.083	-0.044	0.119	0.162	0.838***
Psychosocial functionality	-0.070	0.055	-0.100	-0.023	0.778***
Depressive disorder	-0.064	-0.006	-0.217	-0.383**	0.775***

Table 6: Linear regression of standardized outcome variables in February on the four treatment dummies MI, ER, SA and RA and on the initial score in November. Robust standard errors. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

It is immediately evident that a lot of previously existent significant effects vanish, and what remains (with at least 5% significance) is the effect of RA treatment on stress and depressive disorder. Additionally, there is a new effect of mindfulness on self-compassion that was not evident in the previous measures. Concerning the outcomes, the most robust effect over all measures thus far was the effect from resource activation on stress and depressive disorder. The following subsections will investigate whether this also holds for further adjustments of the regression.

The differences between the regression outcome and the previous outcomes have several reasons. First, I was able to include all four treatments (which was also the case in ANOVA, but not the other measures). Second, and most importantly, the initial scores were included, i.e., I concentrated on the differences over time. Third, I was able to include robust standard errors and therefore did not have a homogeneity assumption. The differences in the regression results compared to the previous measures are quite strong. The reason I included this "simple" regression was exactly for this purpose, to compare the results of a still rather simple regression to the previous measurements. A main takeaway from this comparison is therefore that, at least in a factorial design, it makes a great difference which outcome measure is chosen. As the regression approach can account for several treatment effects, heterogeneity, and the initial score, it appears that it should be the favored outcome measure. It is also useful, in cases like this, to report multiple measures and determine which effects are consistent over different measures.

6.2 Including covariates

As a next step, I included several covariates that could possibly influence the results. The regression, therefore, looks similar to before:

$$\tilde{Y}_i^{j,feb} = \beta_0 + \beta_1 MI_i + \beta_2 ER_i + \beta_3 SA_i + \beta_4 RA_i + \beta_5 \tilde{Y}_i^{j,nov} + \gamma \mathbf{X} + \varepsilon_i \quad (12)$$

where the only new addition is the vector of covariates $\mathbf{X}$. This vector includes the demographic variables age²⁴ and gender²⁵, a therapy experience dummy²⁶, and the Big Five personality traits openness, conscientiousness, extraversion, agreeableness, and neuroticism, all of which were captured in the additional questionnaire. Besides the treatment effects, the influence of the Big Five is of particular interest due to their importance and prevalence in psychological research as well as their connection to mental health and well-being outcomes (Bucher, Suzuki, and Samuel 2019, Anglim et al. 2020). This allows one to evaluate which personality traits also have an effect on the change in outcome variables and how the size of the effect compares to the treatment effect. This regression specification also serves as the baseline regression result.

Dependent Variable	MI	ER	SA	RA	Initial Score	OP	CC	EX	AG	NR
Mindfulness	0.017	-0.107	-0.064	0.190	0.860***	-0.004	0.130	-0.074	-0.183**	-0.109
Self compassion	0.232*	0.161	-0.036	0.120	0.664***	0.113	0.147*	-0.056	0.078	-0.176**
Emotional regulation	0.005	0.057	0.180	0.158	0.424**	0.170	0.169	0.046	0.028	-0.175*
Self-acceptance	0.097	0.214	0.040	-0.138	0.698***	0.130*	0.177**	-0.031	-0.043	-0.131**
Autonomy	-0.123	-0.071	0.381*	0.292	0.613***	0.076	-0.125	0.110	-0.005	0.090
Personal growth	0.157	0.122	-0.398**	-0.049	0.458***	0.119	0.137*	-0.014	0.073	-0.047
Self-efficacy	0.061	0.040	0.059	0.074	0.405***	0.149**	0.236***	0.045	0.025	-0.271***
Flourishing	-0.127	0.189	-0.039	0.007	0.448***	0.113	0.193**	-0.044	0.009	-0.122
Environmental mastery	0.227	0.180	0.186	-0.043	0.646***	-0.026	0.195**	0.042	0.158**	-0.189**
Positive relations	0.054	0.146	0.018	0.065	0.519***	0.098	-0.025	-0.040	0.130	-0.070
Purpose	0.035	0.181	0.049	-0.036	0.565***	0.026	-0.160*	-0.033	0.034	-0.092
Stress	0.135	-0.144	-0.067	-0.578***	0.489***	-0.121	-0.317***	0.056	-0.048	0.193**
Subjective wellbeing	0.047	0.010	-0.034	0.083	0.728***	0.055	0.184**	0.033	0.061	-0.076
Psychosocial functionality	-0.035	0.079	-0.040	0.125	0.591***	-0.098	-0.210**	0.030	-0.116	0.124
Depressive disorder	-0.088	-0.088	0.020	-0.428***	0.711***	0.057	-0.191*	-0.135	0.037	0.080

Table 7: Linear regression of the standardized outcome variables on the four treatment dummies MI, ER, SA and RA, the initial score, the Big Five personality traits conscientiousness (CC), extraversion (EX), agreeableness (AG) and neuroticism (NR). The Big Five personality traits are measured on a scale from 1 to 7. Also included but not depicted are the covariates: age (linearly included), gender (dummy) and therapy experience (dummy). * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

The regression results are visible in Table 7. The first noticeable aspect is that the effects of RA on depressive disorder and stress remain with 5% significance compared to the previous regression. These treatment effects were evident in all of the specifications thus far. Receiving RA treatment is associated with a decrease in the outcome variables stress and depressive disorder of about 0.57 and 0.43 standard deviations. Comparing this to the categorization of Hedges' G effect sizes, this could be categorized as a moderate effect size. The effect of MI on self-compassion decreases in significance to 10% while the size of the effect is rather small. There is now also a new, negative effect of the self-acceptance treatment on personal growth that is in the small to moderate size range and has 5% significance. Most likely, this new effect has arisen

²⁴included linearly

²⁵as a dummy

²⁶that indicates whether a participant had previous therapy experience before the course

now that more explanatory variables are included. While a negative treatment coefficient for the outcome variables stress and depressive disorder is positive, a negative treatment coefficient for the outcome variable personal growth could be perceived as negative, although it makes sense that more self-acceptance leads to less of a personal growth mindset and does not necessarily need to be interpreted negatively.

Let me now examine the other explanatory variables. The three demographic variables age, gender, and therapy experience have been excluded from table 7 as they are not relevant for the questions this paper is trying to answer²⁷. Turning to the Big Five personality traits, one has to keep in mind that they are measured on a scale from 1 to 7. The bigger picture seems to be that conscientiousness and neuroticism affect the outcome variables most often, while openness and agreeableness are only significant in a few instances. The fact that extraversion does not affect the results also makes sense given that the context is emotional skills, which require more work with oneself than with others. Looking at the size of the effects for the Big Five, they are small to moderate. But given that the scale goes from 1 to 7, they could arguably be more important than treatment (and are also significant often when treatment is not). While an in-depth analysis of this fact is a topic for another paper, it also confirms that personality traits play a huge role for emotional skill, mental health, and subjective well-being outcomes, and also opens up the discussion whether personality could be a determining factor for treatment success. Looking at the direction of the effect, conscientiousness affects outcomes mostly positively,²⁸ and neuroticism affects outcomes negatively. Most striking are the effects of conscientiousness and neuroticism on self-efficacy, where both traits are significant at the 1% level.

6.3 Taking into account group structures with group-clustered errors

In the TESC program, there were 16 different groups and in total 14 different teachers²⁹. In addition to teachers having different teaching styles, it is conceivable that group dynamics can affect the outcome. Since the sessions also included opening up to others and participants working with each other, groups can play a role in determining the outcome. Unfortunately, including group fixed effects was not feasible due to the limited number of observations. Other alternatives that can handle group effects, such as mixed-design regressions, also require more observations due to additional assumptions about the variance structure and the random effects estimation (Rencher and Schaalje 2008) and were infeasible for that reason. Therefore, the next regression has the same specification as equation 12, but includes group-clustered standard errors. While this does not explicitly estimate the group effects, it at least accounts for residual correlation of observations within the same group (MacKinnon 2019, Cameron and Miller 2015).

²⁷but are nevertheless included in the regression for completeness reasons

²⁸The effect of conscientiousness on psychosocial functionality is an exception. Given that psychosocial functionality also measures the ability to handle stressful and difficult situations and the ability to handle worry and stress, it could be argued that it is not too surprising that conscientiousness is negatively related to this outcome.

²⁹Two teachers taught two groups each.

Although one has to be cautious with small numbers of clusters, it makes sense to include them as a robustness check due to the data being grouped by design of the experiment. The regression results are depicted in Table 8.

Dependent Variable	MI	ER	SA	RA	Initial Score	OP	CC	EX	AG	NR
Mindfulness	0.017	-0.107	-0.064	0.190	0.860***	-0.004	0.130	-0.074	-0.183**	-0.109**
Self-compassion	0.232*	0.161	-0.036	0.120	0.664***	0.113*	0.147**	-0.056	0.078	-0.176***
Emotional regulation	0.005	0.057	0.180	0.158	0.424**	0.170	0.169	0.046	0.028	-0.175*
Self-acceptance	0.097	0.214*	0.040	-0.138	0.698***	0.130**	0.177*	-0.031	-0.043	-0.131*
Autonomy	-0.123	-0.071	0.381	0.292	0.613***	0.076	-0.125	0.110	-0.005	0.090
Personal growth	0.157	0.122	-0.398***	-0.049	0.458***	0.119*	0.137	-0.014	0.073	-0.047
Self-efficacy	0.061	0.040	0.059	0.074	0.405***	0.149*	0.236***	0.045	0.025	-0.271***
Flourishing	-0.127	0.189	-0.039	0.007	0.448***	0.113	0.193**	-0.044	0.009	-0.122***
Environmental mastery	0.227	0.180	0.186	-0.043	0.646***	-0.026	0.195***	0.042	0.158*	-0.189***
Positive relations	0.054	0.146	0.018	0.065	0.519***	0.098	-0.025	-0.040	0.130	-0.070
Purpose	0.035	0.181	0.049	-0.036	0.565***	0.026	-0.160*	-0.033	0.034	-0.092
Stress	0.135	-0.144	-0.067	-0.578***	0.489***	-0.121	-0.317***	0.056	-0.048	0.193**
Subjective wellbeing	0.047	0.010	-0.034	0.083	0.728***	0.055	0.184	0.033	0.061	-0.076
Psychosocial functionality	-0.035	0.079	-0.040	0.125	0.591***	-0.098*	-0.210**	0.030	-0.116	0.124
Depressive disorder	-0.088	-0.088	0.020	-0.428***	0.711***	0.057	-0.191**	-0.135*	0.037	0.080

Table 8: Linear regression with group-clustered standard errors of the standardized outcome variables on the four treatment dummies MI, ER, SA and RA, the initial score, the Big Five personality traits conscientiousness (CC), extraversion (EX), agreeableness (AG) and neuroticism (NR). The Big Five personality traits are measured on a scale from 1 to 7. Also included but not depicted are the covariates: age (linearly included), gender (dummy), and therapy experience (dummy). * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

Looking first at the treatment effects, the effect of SA on personal growth increased in significance, there is a new, but only 10% significant, effect of ER on self-acceptance, and the effect of RA on stress and depressive disorder remained. Since only the standard errors change with the switch to group-clustered errors, the changes here are only in significance, not coefficient size. The previous results do not change substantially, which supports the previous results, especially those effects that remain similar for both specifications. When looking at the Big Five, there are some changes, such as a new effect of neuroticism on mindfulness and flourishing. For the Big Five in general, there is some fluctuation, but the general impression remains. Especially conscientiousness and neuroticism have a significant effect on outcomes and are important to consider. Therefore, including group-clustered errors does not substantially change the main results of the previous specification.

6.4 Taking into account self-selection with inverse probability weighting

Finally, I wanted to address the drop-out problem mentioned in section 3.2 for the last robustness check. With the observations from the first questionnaire and the demographic characteristics, I was able to determine whether drop-out was significantly related to characteristics that were measured in the questionnaire. There was a significant difference between those who dropped out of the sample and those who did not for several variables, such as the variables gender, conscientiousness, and more.

There are several ways to address self-selection. I decided to use inverse probability weighting. This is a common solution for the drop-out problem and has been discussed in several papers, such as Wooldridge 2007 and Seaman and White 2013. Other methods that rely on instrumental variables, such as the Heckman selection correction, were not feasible, since I was not able to satisfy the exclusion restriction, i.e., there was no variable that affected drop-out, but not outcome variables³⁰ (Puhani 2000).

The idea behind inverse probability weighting is as follows. In the first step, the probability of an individual staying in the sample is estimated with a logistic regression, using variables that can affect drop-out. This is depicted in the following regression.

$$D_i = \ln \left(\frac{P}{1 - P} \right) = \beta_0 + \beta_1 \cdot \text{Gender}_i + \beta_2 \cdot \text{Flourishing}_i^{nov} + \beta_3 \cdot \text{Environmental_mastery}_i^{nov} + \beta_4 \cdot \text{Personal_growth}_i^{nov} + \beta_5 \cdot \text{Stress}_i^{nov} + \beta_6 \cdot \text{Therapy_experience}_i + \beta_7 \cdot \text{NR}_i + \beta_8 \cdot \text{CC}_i \quad (13)$$

The explanatory variable is whether an individual stayed in the sample (the dummy variable D_i). This is explained via a logistic regression with all variables that can effect drop-out (and possibly the outcome variables themselves). Therefore, on the right-hand side are all the explanatory variables that can explain drop-out. In this case, I used all variables for which there was a significant difference between those who dropped out of the sample and those who did not³¹, plus other variables that could play a role in affecting drop-out and the outcome variables³². I also tested other specifications of the logistic regression, and ultimately chose this one as the main specification, because all variables that show a significant difference between the drop-outs and the individuals who stayed in the sample are included. Alternative specifications only led to slight differences in results, an alternative IPW specification can be found in appendix B.2.

In the second step, the inverse of the probabilities of staying in the sample (that come from the logistic regression in the first step) are calculated and used as weights in the regression as known from section 6.2³³. Group-clustered standard errors are also included. The resulting regression table can be seen in Table 9.

³⁰A variable that was randomly assigned and only affected drop-out (such as rewards for answering the questionnaires) would have been needed for that.

³¹which were gender, conscientiousness and the initial scores for flourishing, environmental mastery, personal growth

³²i.e., in this case the initial stress score, the therapy experience dummy and the neuroticism score

³³The same ipw approach was also done by Doyle et al. 2016.

Dependent Variable	MI	ER	SA	RA	Initial Score	OP	CC	EX	AG	NR
Mindfulness	-0.056	-0.119	-0.052	0.172	0.864***	0.024	0.123	-0.085	-0.194**	-0.079
Self-compassion	0.171	0.165	-0.097	0.163*	0.624***	0.153**	0.202***	-0.062	0.082	-0.182***
Emotional regulation	-0.141	0.040	0.210	0.103	0.414**	0.232*	0.094	0.027	-0.039	-0.158
Self-acceptance	0.077	0.179	0.072	-0.155	0.690***	0.134**	0.182*	-0.056	-0.022	-0.177***
Autonomy	-0.248	-0.071	0.485*	0.235	0.592**	0.141	-0.214	0.058	-0.031	0.082
Personal growth	0.094	0.117	-0.524***	0.097	0.518***	0.036	0.194*	0.011	0.102	-0.014
Self-efficacy	0.026	0.100	0.057	0.091	0.383***	0.133	0.214***	0.042	0.032	-0.308***
Flourishing	-0.183	0.236	-0.224	0.110	0.450***	0.086	0.282***	-0.041	0.062	-0.127**
Environmental mastery	0.098	0.110	0.209	0.009	0.646***	0.010	0.201***	0.002	0.132*	-0.197***
Positive relations	-0.004	0.146	0.048	0.092	0.520***	0.110	-0.033	-0.056	0.107	-0.062
Purpose	0.009	0.184	0.113	-0.000	0.616***	0.050	-0.187*	-0.074	0.042	-0.133
Stress	0.286*	-0.162	0.017	-0.601***	0.491***	-0.108	-0.319***	0.058	-0.033	0.199**
Subjective wellbeing	-0.070	0.015	0.044	0.117	0.710***	0.065	0.177	0.013	0.059	-0.062
Psychosocial functionality	-0.008	-0.041	-0.004	0.188	0.630***	-0.089	-0.175*	0.046	-0.105	0.124
Depressive disorder	-0.015	-0.107	0.058	-0.507***	0.693***	0.063	-0.211**	-0.118*	0.040	0.064

Table 9: Linear regression with group-clustered standard errors and inverse probability weighting of the standardized outcome variables on the four treatment dummies MI, ER, SA and RA, the initial score, the Big Five personality traits conscientiousness (CC), extraversion (EX), agreeableness (AG) and neuroticism (NR). The Big Five personality traits are measured on a scale from 1 to 7. Also included but not depicted are the covariates: age (linearly included), gender (dummy) and therapy experience (dummy). * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

Looking at the treatment effects, it is evident that the negative effect of SA on personal growth and of RA on stress and depressive disorder remains. There are some fluctuations for outcomes that are only 10% significant that can be disregarded due to the fluctuation in significances throughout the different specifications. The size of the treatment effects increased slightly compared to the previous regression. As for the Big Five variables, there are also some minor changes in significance, but nothing major. The main message remains that the Big Five, especially conscientiousness and neuroticism, play an important role in determining the outcomes, even after the initial scores of the outcome variables are accounted for. The most steadily significant Big Five effects across the different specifications³⁴ are the effect of conscientiousness on self-efficacy, flourishing, environmental mastery, and stress, and the effect of neuroticism on self-compassion, self-efficacy, and environmental mastery. If we look at the outcomes, this means across all regressions, self-efficacy and environmental mastery are affected by the Big 5 variables most often.

In total, this robustness check also does not affect the main result from the benchmark expression. Partly due to the lack of observations, no one specification can be viewed as superior. But all specifications can be judged together to reach a conclusion. What remains constant over all three specifications is the negative effect of SA treatment on personal growth, and, most importantly, the effect of RA treatment on stress as well as depressive disorder. While fluctuations of the Big Five are slightly stronger, the main message still persists. The Big Five, especially conscientiousness and neuroticism, have an effect on the outcome variables that is not to be neglected.

³⁴at least at the 5% level across all regression specifications

7 Discussion of the results

In this section, I summarize and discuss the results and the most important takeaways from this paper. Section 7.1 concentrates on methodological takeaways, section 7.2 summarizes the main outcomes and effects of the study, and section 7.3 discusses limitations of the study as well as possibilities for future research.

7.1 Methodological takeaways

Let me first summarize the methodological takeaways from this paper. I included several outcome measures, ranging from simpler measures to more complicated measures. I started with effect sizes that are common in psychological research and studies. Therefore, I first want to present the benefits of including several measures in section 4. These effect sizes can only measure group differences for two groups (i.e., one factor, treatment vs. control) at one point in time. To compare standard effect sizes to a more standard economic approach, I chose to include a normal and a standardized regression on one treatment dummy. To complement this and the effect size Hedges' G with another effect size that has a different approach, I chose the probability of superiority, which concentrates not on mean differences but on a comparison of outcomes between individuals in the treatment groups and in the control groups. Comparing the results of the four effect sizes to each other, the regression and Hedges' G had mostly similar outcomes, while the probability of superiority differed more often. This is because the calculation of the probability of superiority is conceptually different, but it can suggest that simply reporting Hedges' G to measure effect sizes could be insufficient, and that simply adding another measure can give more insight into the stability of the results.

Let me now highlight the takeaways of going from effect sizes to ANOVA to regression. The first measures, the effect sizes, have several downsides. The most important one is that to interpret them, one has to compare side-by-side the effect sizes from pre-treatment to post-treatment, otherwise, it is not possible to rule out that group differences between treatment and control come from pre-existing group differences. The other disadvantage specific to this study is that the group differences can only be calculated for one treatment at a time. Given that this study uses a factorial design, this simplifies the data analysis, possibly too much. Nevertheless, I included it as a basis for comparison with other, similar, studies, which mainly report Hedges' G . The next step was therefore to include ANOVA results. The reason is that it is one step further than the previous effect size measures as it can take into account all four factors at once. The downside is that it still can only be calculated for one point in time, which means a comparison of pre-treatment and post-treatment ANOVA is necessary. Although this issue is specific to this study due to the factorial design, using ANOVA already eliminated a number of significant group differences, highlighting that taking into account one treatment at a time might not be

complex enough³⁵. The most important takeaway might be taken from the shift from ANOVA to the first regression set-up in section 6.1. The regression, like ANOVA, takes into account all four factors at the same time, but also takes into account initial outcome measures too. This seems to be the crucial point where a lot of previously existent significances vanish. This shows that, while in theory a side-by-side comparison of pre-treatment and post-treatment ANOVA can show changes over time, it makes a huge difference to actually account for initial scores within the outcome measure itself³⁶. Therefore, it seems sensible for future studies, also in the psychological field, to not only rely on measurements that take into account one point in time only.

7.2 Content-specific takeaways and the effects of TESC

Next, I would like to summarize the effects of the course. First, while there is no causality, the average changes in section 3.3 show a generally positive trend in outcome changes. This is important to note because, while this paper explicitly wanted to measure the effect of emotional skill intervention components, the control components (which consisted of social skill topics) could also have played a positive role in changing the outcome variables. This may suggest that teaching "any" type of noncognitive content could be helpful to an extent. It is also noteworthy that the positive trend is visible even though the last questionnaire was done at the end of the semester, where exams usually take place and where students experience more pressure and stress than maybe at the beginning of the semester.

Besides the generally positive trend³⁷ there are two effects in particular that I want to point out, because they stayed consistent throughout all measurements. This is specifically the effect of RA treatment on stress and on depressive disorder. Since these effects are visible throughout all specifications, it seems that this is a robust result. The size of the effect, when compared to standard Hedges' G measurements, for example, is moderate and therefore not to be neglected. The only constraint in the interpretation is that resource activation was the last component, which could have played a role for its effectiveness³⁸. In future studies, it would therefore be interesting to uncouple resource activation treatments from the time aspect. In addition, the results do not allow for drawing conclusions for the optimal composition of emotional skill treatments, due to the lack of significance in the results, which prohibited a thorough comparison of components against each other.

³⁵While the set-up, as known from figure 3, suggests that this simple comparison gives treatment effect of individual factors, a limited number of observations and a therefore unbalanced spread of observations per group, among other things, could be a limiting factor for this procedure.

³⁶I want to note that when comparing ANOVA to a simple regression with only the four factors and without the initial scores, the significances only deviate minimally, showing that the main change comes not from the shift from ANOVA to regression but from including the initial outcome variable scores. This regression can be found in table B.12 in appendix B.1.

³⁷and the generally positive feedback of students about the course

³⁸even though the distribution of students over groups and therefore factors was generally balanced

Additionally, one important finding is that the Big Five seem to play an important role in determining outcomes. While the effect is smaller in comparison to the significant effects of RA treatment, keeping in mind that the Big Five are measured on a scale of 1 to 7, larger differences in conscientiousness or neuroticism can influence the results heavily. And most importantly, since the initial outcome level is accounted for, it seems that the Big Five do not just influence the level of the outcome variables, but also the change over the course of the treatment. The importance of the Big Five is not surprising and in line with psychological research, but it shows the importance of personality for intervention studies. This also opens up interesting questions for future research, where the Big Five could be an important moderator for interventions. It is for instance conceivable that those with certain character traits simply benefit more from (emotional skill) interventions, than others³⁹.

7.3 Limitations and recommendations for future studies

There are several limiting factors worth mentioning for interpreting the results. The most severe limiting factor was the high drop-out rate and the resulting low number of observations. For the complexity of the study design, it would have been necessary to have more observations for a more detailed and more complex data analysis. The 271 originally registered students would have been a good observation number, but that dwindled down to 66 observations that included the first and last questionnaire, as well as the additional one that included the Big Five.

Another fact worth mentioning is that the questionnaires included many outcome variables. Contrary to many studies that include fewer outcome variables that often directly center around mental health and subjective well-being, these types of variables were only secondary outcomes in this study as described in figure 1. Looking only at the secondary outcome variables, the lack of significances in the model would not be as great as when looking at all measured variables.

Two smaller facts worth mentioning is the rather short duration of the treatments and the suboptimal timing of the first questionnaire. While the course as such lasted for one semester, the individual components may have been too short for a stronger effect. Additionally, the first questionnaire was sent out in week 4 of the intervention, meaning some lessons had already taken place and some groups were almost finished with the mindfulness intervention. This might be an explanation as to why there was no visible effect for mindfulness treatment.

For future studies, a bigger sample size or a reduced complexity⁴⁰ would be needed to gain more insight on the effectiveness of the individual components compared against each other. It would also be interesting to look at the resource activation component again, decoupled from being the last component of a course. Lastly, a more in-depth analysis of the effect of the Big

³⁹This is an idea that has already received attention by psychological research, even though there is no clear consistency in results yet (see e.g. Giluk 2009).

⁴⁰including a longer time frame for each component

Five on treatment success, as a moderator and interaction effect with treatment, could offer valuable insights into the role of personality in the effectiveness of (emotional skill) interventions.

8 Conclusion

The main objective of this study was to determine the effectiveness of the individual intervention components of TESC (mindfulness, emotional regulation, self-acceptance, and resource activation) in improving measures of emotional skills themselves and additional well-being variables. To determine this, I used several effect size measures, as well as ANOVA and a regression approach. The two treatment effects that were consistently significant were the effect of the resource activation treatment on stress and depressive disorder. I also used group-clustered errors and inverse probability weighting as robustness checks to account for group effects and drop-outs of the course. The two effects were significant in all specifications, including the robustness checks. Due to a limited number of observations, it was not possible to draw any further conclusions on the other treatment components and their optimal composition. It is also noteworthy that, while it is not possible to establish causality, the general trend of all outcome variables is positive over the duration of the course. This general improvement suggests potential benefits of participation in interventions like TESC.

The analysis also revealed that personality traits, especially measures for conscientiousness and neuroticism, play an additional crucial factor for the outcome variables after the intervention, even when accounting for the baseline scores.

The different effect sizes used in section 4.2 showed similar results, but the probability of superiority deviated from the other results more often, indicating that it could be used as an additional indicator of robustness when choosing to report the other measures. The comparison of the different outcome measures also showed that a regression approach that directly includes the baseline outcome variable scores appears most suitable, as, compared to the previous measures, many treatment effects vanish.

The main limitation in this study was the small number of observations, given the complexity of the intervention structure. For future studies, trying out a similar set-up with more participants and longer time frames for individual intervention components could be beneficial. It would also be interesting to examine further the effect of personality traits as moderators, i.e., check for interaction effects between treatment and personality traits.

In general, the findings of this study highlight the potential of structured interventions such as TESC to foster emotional competencies and well-being and underscore the importance of personality traits for outcomes. More research is needed to refine the optimal training composition and determine the effect of personality traits as moderators.

References

- Anglim, Jeromy, Sharon Horwood, Luke D Smillie, Rosario J Marrero, and Joshua K Wood (2020). “Predicting psychological and subjective well-being from personality: A meta-analysis.” In: *Psychological bulletin* 146.4, p. 279.
- Bai, Yuehao (2022). “Optimality of matched-pair designs in randomized controlled trials”. In: *American Economic Review* 112.12, pp. 3911–3940.
- Bruhn, Miriam and David McKenzie (2009). “In pursuit of balance: Randomization in practice in development field experiments”. In: *American economic journal: applied economics* 1.4, pp. 200–232.
- Bucher, Meredith A, Takakuni Suzuki, and Douglas B Samuel (2019). “A meta-analytic review of personality traits and their associations with mental health treatment outcomes”. In: *Clinical psychology review* 70, pp. 51–63.
- Cameron, A Colin and Douglas L Miller (2015). “A practitioner’s guide to cluster-robust inference”. In: *Journal of human resources* 50.2, pp. 317–372.
- Cohen, J (1988). “Statistical power analysts for the behavioral sciences. 2nd edn Hillsdale”. In: *NJL Erlbaum Associates*.
- Collins, Linda M, John J Dziak, Kari C Kugler, and Jessica B Trail (2014). “Factorial experiments: efficient tools for evaluation of intervention components”. In: *American journal of preventive medicine* 47.4, pp. 498–504.
- Costa Jr, Paul T and Robert R McCrae (1992). “Four ways five factors are basic”. In: *Personality and individual differences* 13.6, pp. 653–665.
- Deming, David J (2017). “The growing importance of social skills in the labor market”. In: *The Quarterly Journal of Economics* 132.4, pp. 1593–1640.
- Doyle, Orla, Edel McGlanaghy, Christine O’Farrelly, and Richard E Tremblay (2016). “Can targeted intervention mitigate early emotional and behavioral problems?: generating robust evidence within randomized controlled trials”. In: *PLoS One* 11.6, e0156397.
- Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007). “Using randomization in development economics research: A toolkit”. In: *Handbook of development economics* 4, pp. 3895–3962.
- Edin, Per-Anders, Peter Fredriksson, Martin Nybom, and Björn Öckert (2022). “The rising return to noncognitive skill”. In: *American Economic Journal: Applied Economics* 14.2, pp. 78–100.
- Erceg-Hurn, David M and Vikki M Miroseovich (2008). “Modern robust statistical methods: an easy way to maximize the accuracy and power of your research.” In: *American Psychologist* 63.7, p. 591.
- Firpo, Sergio, Miguel N Foguel, and Hugo Jales (2020). “Balancing tests in stratified randomized controlled trials: A cautionary note”. In: *Economics Letters* 186, p. 108771.
- Gassmann, Daniel and Klaus Grawe (2006). “General change mechanisms: The relation between problem activation and resource activation in successful and unsuccessful therapeutic inter-

- actions". In: *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* 13.1, pp. 1–11.
- Gignac, Gilles E and Eva T Szodorai (2016). "Effect size guidelines for individual differences researchers". In: *Personality and individual differences* 102, pp. 74–78.
- Giluk, Tamara L (2009). "Mindfulness, Big Five personality, and affect: A meta-analysis". In: *Personality and Individual Differences* 47.8, pp. 805–811.
- Grissom, Robert J and John J Kim (2001). "Review of assumptions and problems in the appropriate conceptualization of effect size." In: *Psychological Methods* 6.2, p. 135.
- Gross, James J (2015). "Emotion regulation: Current status and future prospects". In: *Psychological inquiry* 26.1, pp. 1–26.
- Heckman, James J (2006). "Skill formation and the economics of investing in disadvantaged children". In: *Science* 312.5782, pp. 1900–1902.
- Heckman, James J and Tim Kautz (2012). "Hard evidence on soft skills". In: *Labour economics* 19.4, pp. 451–464.
- Hedges, Larry V (1981). "Distribution theory for Glass's estimator of effect size and related estimators". In: *journal of Educational Statistics* 6.2, pp. 107–128.
- Heintzelman, Samantha J, Kostadin Kushlev, Lesley D Lutes, Derrick Wirtz, Jacqueline M Kanippayoor, Damian Leitner, Shigehiro Oishi, and Ed Diener (2020). "ENHANCE: Evidence for the efficacy of a comprehensive intervention program to promote subjective well-being." In: *Journal of Experimental Psychology: Applied* 26.2, p. 360.
- Hodzic, Sabina, Jana Scharfen, Pilar Ripoll, Heinz Holling, and Franck Zenasni (2018). "How efficient are emotional intelligence trainings: A meta-analysis". In: *Emotion Review* 10.2, pp. 138–148.
- Kabat-Zinn, Jon (2015). "Mindfulness". In: *Mindfulness* 6.6, pp. 1481–1483.
- MacKinnon, James G (2019). "How cluster-robust inference is changing applied econometrics". In: *Canadian Journal of Economics/Revue canadienne d'économie* 52.3, pp. 851–881.
- Puhani, Patrick (2000). "The Heckman correction for sample selection and its critique". In: *Journal of economic surveys* 14.1, pp. 53–68.
- Recabarren, Romina E, Claudie Gaillard, Matthias Guillod, and Chantal Martin-Soelch (2019). "Short-term effects of a multidimensional stress prevention program on quality of life, well-being and psychological resources. A randomized controlled trial". In: *Frontiers in psychiatry* 10, p. 88.
- Rencher, Alvin C and G Bruce Schaalje (2008). *Linear models in statistics*. John Wiley & Sons.
- Ryff, Carol D (1989). "Happiness is everything, or is it? Explorations on the meaning of psychological well-being." In: *Journal of personality and social psychology* 57.6, p. 1069.
- Schupp, Jürgen and Jean-Yves Gerlitz (2014). "Big five inventory-soep (bfi-s)". In: *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. Vol. 10.
- Seaman, Shaun R and Ian R White (2013). "Review of inverse probability weighting for dealing with missing data". In: *Statistical methods in medical research* 22.3, pp. 278–295.

- Seppälä, Emma M, Christina Bradley, Julia Moeller, Leilah Harouni, Dhruv Nandamudi, and Marc A Brackett (2020). “Promoting mental health and psychological thriving in university students: a randomized controlled trial of three well-being interventions”. In: *Frontiers in psychiatry* 11, p. 534776.
- Tran-Huu, Sarah F (2023). *Das Studienmodul TESK an der JGU Mainz. Hintergrund und Konzept für Interessierte zu unserem Training emotionaler und sozialer Kompetenzen (TESK)*. Accessed October 7, 2025. URL: https://www.macro.economics.uni-mainz.de/files/2023/06/Hintergrund-Konzept_TESK_Juni-2023.pdf (visited on 10/07/2025).
- Tran-Huu, Sarah F (2025). *Resilienz und Psychologisches Wohlbefinden Junger Erwachsener Fördern*. Springer.
- Van Agteren, Joep, Matthew Iasiello, Laura Lo, Jonathan Bartholomaeus, Zoe Kopsaftis, Marissa Carey, and Michael Kyrios (2021). “A systematic review and meta-analysis of psychological interventions to improve mental wellbeing”. In: *Nature human behaviour* 5.5, pp. 631–652.
- Wooldridge, Jeffrey M (2007). “Inverse probability weighted estimation for general missing data problems”. In: *Journal of econometrics* 141.2, pp. 1281–1301.

Appendix

A Effect size measures

A.1 Effect size tables in November

Variable	Reg.	Norm. Reg.	Hedges-G	CI Low	CI High	Prob. Super.
Mindfulness (MI)	-0.179	-0.240	-0.242	-0.672	0.189	0.421
Mindfulness (ER)	0.066	0.088	0.089	-0.342	0.519	0.525
Mindfulness (SA)	0.209	0.281	0.284	-0.146	0.712	0.576
Mindfulness (RA)	-0.074	-0.099	-0.100	-0.527	0.329	0.480
Self-compassion (MI)	-0.225	-0.329	-0.320	-0.750	0.113	0.418
Self-compassion (ER)	0.398**	0.581**	0.581**	0.140	1.018	0.693***
Self-compassion (SA)	0.158	0.231	0.223	-0.206	0.650	0.559
Self-compassion (RA)	0.187	0.273	0.264	-0.166	0.693	0.602
Emotional regulation (MI)	-0.117	-0.194	-0.188	-0.617	0.243	0.445
Emotional regulation (ER)	0.166	0.274	0.266	-0.167	0.697	0.586
Emotional regulation (SA)	0.183	0.302	0.295	-0.135	0.723	0.566
Emotional regulation (RA)	0.036	0.059	0.057	-0.371	0.484	0.535
Self-acceptance (MI)	0.281	0.080	0.083	-0.346	0.512	0.498
Self-acceptance (ER)	1.230*	0.351*	0.370	-0.065	0.803	0.564
Self-acceptance (SA)	1.291*	0.368*	0.389	-0.042	0.819	0.551
Self-acceptance (RA)	-0.465	-0.133	-0.138	-0.566	0.291	0.436
Autonomy (MI)	1.001	0.277	0.288	-0.144	0.718	0.527
Autonomy (ER)	0.939	0.260	0.270	-0.163	0.701	0.533
Autonomy (SA)	-0.314	-0.087	-0.089	-0.516	0.338	0.423
Autonomy (RA)	0.185	0.051	0.053	-0.375	0.481	0.475
Personal growth (MI)	0.473	0.168	0.189	-0.241	0.618	0.459
Personal growth (ER)	1.010*	0.360*	0.410	-0.026	0.844	0.520
Personal growth (SA)	0.697	0.248	0.280	-0.150	0.708	0.489
Personal growth (RA)	0.049	0.017	0.020	-0.408	0.447	0.413
Self-efficacy (MI)	-0.114	-0.022	-0.024	-0.453	0.405	0.434
Self-efficacy (ER)	1.192	0.235	0.254	-0.179	0.685	0.544
Self-efficacy (SA)	1.807*	0.356*	0.389	-0.043	0.819	0.591
Self-efficacy (RA)	-0.710	-0.140	-0.150	-0.578	0.278	0.421
Flourishing (MI)	0.866	0.117	0.134	-0.295	0.563	0.481
Flourishing (ER)	3.486**	0.470**	0.561**	0.121	0.998	0.639**
Flourishing (SA)	2.787**	0.376**	0.442**	0.009	0.873	0.567
Flourishing (RA)	-0.757	-0.102	-0.117	-0.545	0.311	0.457
Environmental mastery (MI)	-0.808	-0.218	-0.217	-0.646	0.214	0.380*
Environmental mastery (ER)	0.563	0.152	0.151	-0.281	0.581	0.518
Environmental mastery (SA)	1.344*	0.362*	0.365	-0.067	0.794	0.573
Environmental mastery (RA)	-0.453	-0.122	-0.121	-0.549	0.307	0.409
Purpose (MI)	0.119	0.043	0.049	-0.380	0.477	0.422
Purpose (ER)	-1.203**	-0.434**	-0.511**	-0.946	-0.072	0.316***
Purpose (SA)	0.910*	0.329*	0.381	-0.051	0.810	0.544
Purpose (RA)	-0.016	-0.006	-0.007	-0.434	0.421	0.437
Positive relations (MI)	-1.126	-0.295	-0.299	-0.729	0.134	0.395
Positive relations (ER)	0.747	0.196	0.197	-0.235	0.627	0.518
Positive relations (SA)	0.420	0.110	0.110	-0.317	0.537	0.500
Positive relations (RA)	-0.439	-0.115	-0.115	-0.543	0.313	0.447

Table A.10: Effect sizes of primary outcome variables in November for standard regression, standardized regression, Hedges G, and Probability of Superiority. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

Variable	Reg.	Norm. Reg.	Hedges-G	CI Low	CI High	Prob. Superiority
Stress (MI)	1.727	0.272	0.295	-0.137	0.725	0.576
Stress (ER)	-2.096	-0.330	-0.360	-0.792	0.075	0.361*
Stress (SA)	-1.905	-0.300	-0.326	-0.754	0.105	0.397
Stress (RA)	0.335	0.053	0.057	-0.371	0.484	0.493
Psychosocial functionality (MI)	0.139**	0.428**	0.458**	0.022	0.891	0.606
Psychosocial functionality (ER)	0.002	0.005	0.005	-0.425	0.435	0.424
Psychosocial functionality (SA)	-0.073	-0.224	-0.235	-0.662	0.194	0.385*
Psychosocial functionality (RA)	0.002	0.006	0.006	-0.421	0.434	0.445
Depressive disorder (MI)	0.027	0.053	0.062	-0.368	0.490	0.479
Depressive disorder (ER)	-0.092	-0.183	-0.215	-0.645	0.217	0.382
Depressive disorder (SA)	-0.053	-0.106	-0.124	-0.551	0.303	0.444
Depressive disorder (RA)	-0.059	-0.117	-0.137	-0.565	0.292	0.423
Subjective wellbeing (MI)	-0.493	-0.063	-0.061	-0.490	0.368	0.485
Subjective wellbeing (ER)	2.462	0.315	0.309	-0.125	0.741	0.576
Subjective wellbeing (SA)	3.095*	0.396*	0.391	-0.041	0.821	0.602
Subjective wellbeing (RA)	-2.038	-0.260	-0.255	-0.684	0.175	0.443

Table A.11: Effect sizes of secondary outcome variables in November for standard regression, standardized regression, Hedges G, and Probability of Superiority. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

B Extended regression approach

B.1 Regression on the four dummies

As explained in the main text, a regression with only the four dummies was omitted in the main text due to the similarity in concept and result to ANOVA. The results table can be found in this table B.12.

Dependent Variable	MI	ER	SA	RA
Mindfulness	-0.279	-0.099	0.298	0.251
Self-compassion	-0.030	0.491**	0.273	0.363
Emotional regulation	-0.077	0.226	0.398*	0.316
Self-acceptance	0.184	0.385*	0.473**	-0.089
Autonomy	0.063	0.224	0.064	0.113
Personal growth	0.377**	0.399**	-0.142	0.081
Self-efficacy	-0.055	0.158	0.341*	0.073
Flourishing	0.049	0.492***	0.359*	0.119
Environmental mastery	-0.145	0.159	0.619***	-0.004
Positive relations	-0.093	0.189	0.186	0.083
Purpose	-0.075	-0.032	0.088	-0.114
Stress	0.368	-0.274	-0.398*	-0.583**
Subjective wellbeing	0.077	0.152	0.519**	-0.050
Psychosocial functionality	0.354	0.191	-0.406*	-0.032
Depressive disorder	-0.046	-0.090	-0.391*	-0.597***

Table B.12: Regression with only the treatment dummies. * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

This can be compared with ANOVA because there are only four explanatory variables. The method of estimation is different, but the significances are fairly similar. At least at the five percent level, there are several effects, such as mindfulness on personal growth, emotional regulation treatment on self-compassion, personal growth, and flourishing, self-acceptance on

self-acceptance, environmental mastery and subjective well-being, and resource activation on stress and depressive disorder. Comparing this directly with the ANOVA results in table 4, the significances are very similar.

B.2 Alternative inverse probability weighting specification

In the process of specifying the logistic regression for the inverse probability weighting specification, I also tested different regression specification. One of which was the following regression:

$$D_i = \beta_0 + \beta_1 \cdot \text{Gender}_i + \beta_2 \cdot \text{Stress}_i^{\text{nov}} + \beta_3 \cdot \text{Therapy_experience}_i + \beta_4 \cdot \text{NR}_i + \beta_5 \cdot \text{CC}_i$$

which is a shortened version of the main specification. I chose this as an alternative, because the initial scores flourishing, environmental mastery and personal growth are not included. But even changing this, the main results only change slightly, as is visible in table B.13

Dependent Variable	MI	ER	SA	RA	Initial Score	OP	CC	EX	AG	NR
Mindfulness	-0.034	-0.108	-0.058	0.199	0.854***	0.007	0.136*	-0.072	-0.180*	-0.099
Self-compassion	0.170	0.166	-0.122	0.179*	0.626***	0.136*	0.218***	-0.052	0.091	-0.180***
Emotional regulation	-0.109	0.031	0.233	0.089	0.402**	0.198	0.139	0.023	-0.015	-0.182*
Self-acceptance	0.048	0.169	0.074	-0.136	0.707***	0.124**	0.157*	-0.045	-0.035	-0.166***
Autonomy	-0.242	-0.076	0.455	0.253	0.605***	0.109	-0.167	0.055	-0.024	0.073
Personal growth	0.100	0.135	-0.537***	0.085	0.502***	0.028	0.201*	0.030	0.102	-0.013
Self-efficacy	-0.003	0.058	0.017	0.109	0.399***	0.115	0.218***	0.051	0.033	-0.285***
Flourishing	-0.195	0.235	-0.252	0.119	0.439***	0.062	0.292**	-0.013	0.049	-0.122**
Environmental mastery	0.149	0.156	0.212	0.013	0.622***	-0.003	0.204***	0.027	0.143*	-0.217***
Positive relations	-0.006	0.146	0.020	0.067	0.504***	0.091	-0.016	-0.033	0.123	-0.065
Purpose	-0.014	0.178	0.094	0.027	0.589***	0.051	-0.198**	-0.075	0.033	-0.118
Stress	0.274*	-0.171	0.007	-0.647***	0.494***	-0.075	-0.339***	0.026	-0.056	0.197**
Subjective wellbeing	-0.041	-0.000	0.025	0.101	0.731***	0.057	0.183	0.012	0.060	-0.076
Psychosocial functionality	0.052	-0.013	-0.006	0.156	0.590***	-0.072	-0.205**	0.043	-0.111	0.145
Depressive disorder	-0.017	-0.090	0.102	-0.523***	0.689***	0.091	-0.230**	-0.124*	0.033	0.068

Table B.13: Linear regression with group-clustered standard errors and an alternative specification for the inverse probability weighting of the standardized outcome variables on the four treatment dummies MI, ER, SA and RA, the initial score, the Big Five personality traits conscientiousness (CC), extraversion (EX), agreeableness (AG) and neuroticism (NR). The Big Five personality traits are measured on a scale from 1 to 7. Also included but not depicted are the covariates: age (linearly included), gender (dummy), and therapy experience (dummy). * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.01$

When comparing the two tables, there are only very minor changes compared to the main specification. Therefore, it does not seem that the exact specification matters greatly and supports the statement that the proposed main specification is a reasonable one for the purpose of this study.