Effects of Mindfulness on Conceptual Modeling Performance: a Series of Experiments

Beatriz Bernárdez, Amador Durán, José A. Parejo, Natalia Juristo, Antonio Ruiz–Cortés

Abstract—**Context**. Mindfulness is a meditation technique whose main goal is keeping the mind calm and educating attention by focusing only on one thing at a time, usually breathing. The reported benefits of its continued practice can be of interest for Software Engineering students and practitioners, especially in tasks like conceptual modeling, in which concentration and clearness of mind are crucial. **Goal**. In order to evaluate whether Software Engineering students enhance their conceptual modeling performance after several weeks of mindfulness practice, a series of three controlled experiments were carried out at the University of Seville during three consecutive academic years (2013–2016) involving 130 students. **Method**. In all the experiments, the subjects were divided into two groups. While the experimental group practiced mindfulness, the control group was trained in public speaking as a placebo treatment. All the subjects developed two conceptual models based on a transcript of an interview, one before and another one after the treatment. The results were compared in terms of conceptual modeling quality (measured as effectiveness, i.e. the percentage of model elements correctly identified) and productivity (measured as efficiency, i.e. the number of model elements correctly identified per unit of time). **Results**. The statistically significant results of the series of experiments revealed that the subjects who practiced mindfulness developed slightly better conceptual models (their quality was 8.16% higher) and they did it faster (they were 46.67% more productive) than the control group, even if they did not have a previous interest in meditation. **Conclusions**. The practice of mindfulness improves the performance of Software Engineering students in conceptual modeling, especially their productivity. Nevertheless, more experimentation is needed in order to confirm the outcomes in other Software Engineering tasks and populations.

Index Terms—Mindfulness, Conceptual Modeling, Experiment Replication, Family of Experiments, Software Engineering Education

✦

1 INTRODUCTION

THERE is a growing interest in the Software Engineering community about the influence of psychological aspects on the software process. Recent works have studied the impact, among others, of motivation [1], emotional intelligence [2], moods and emotions [3] and *being in the zone*, also known as *flow* [4]. Remarkably, developers perceive their days as productive when they complete many or big tasks without significant interruptions or context switches [5]. External interruptions can be avoided by switching off mobile phones and messaging applications, but internal interruptions due to lack of concentration or mind wandering are sometimes difficult to deal with in stressful environments.

In 1979, Jon Kabat–Zinn started to apply a meditation technique—known as *mindfulness*—as a therapeutic treatment for stress. His first two books about mindfulness and its benefits began its popularization in Western countries in the 1990s [6], [7]. Among other effects such as the reduction of symptoms associated with stress, anxiety or insomnia, mindfulness has been reported to be useful for educating attention and enhancing mental clarity, thus improving problem–solving capabilities [8], [9]. Lately, mindfulness has entered the mainstream media, getting the attention not only of many people but also of several relevant companies, including Google, Intel and Goldman Sachs among others [10], [11].

Considering the reported benefits of mindfulness on several cognitive processes, educators around the world have begun to use it as a learning tool for students across a wide variety of age and education levels [12], including higher education [13]. In our case, after teaching Requirements Engineering for more than 20 years, we think that our Software Engineering students, i.e. future software developers, can also benefit from practicing mindfulness. We have focused on Requirements Engineering because of our research and professional experience in the topic and because we teach related courses. Among the different tasks in Requirements Engineering, we have chosen conceptual modeling because we think that some of the reported enhancements due to the practice of mindfulness such as sustained attention [14], working memory [15], equanimity[1], and mental clarity [16] could help our students to build up complex mental representations of problems such as those needed during conceptual modeling and, probably, other Software Engineering tasks.

For the purpose of checking our intuition, the following research question was stated:

• B. Bernárdez, A. Durán, J. A. Parejo, and A. Ruiz–Cortés are with the E.T.S.I. Informática, Universidad de Sevilla, Spain. E-mail: beat@us.es, amador@us.es, japarejo@us.es, aruiz@us.es

• N. Juristo is with the E.T.S.I. Informática, Universidad Politécnica de Madrid, Spain. E-mail: natalia@fi.upm.es

1. Equanimity is related to response inhibition and emotion regulation processes [16], which could improve conceptual modeling performance helping students to deal with frustrations in figuring out the correct solution.
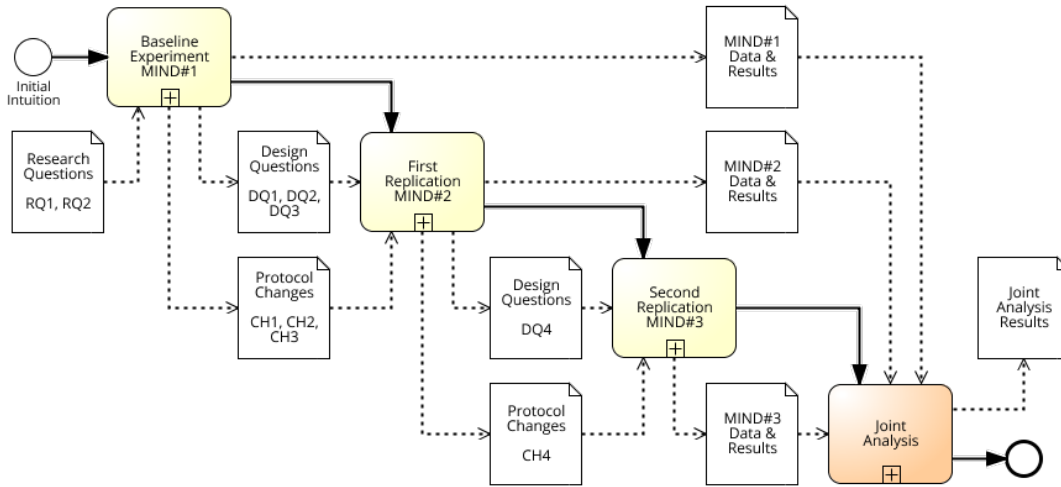
Figure 1. Iterative process of refinement and improvement of the series of controlled experiments reported in this article
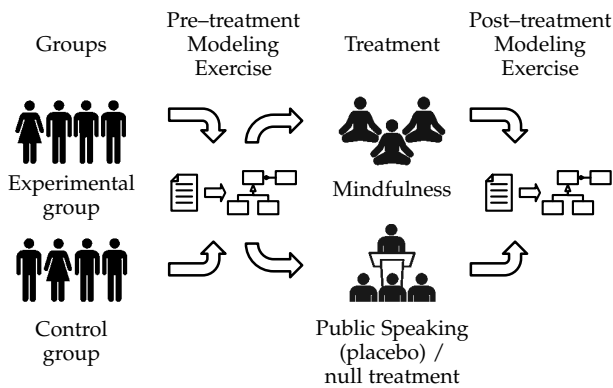

Figure 2. High–level design of the controlled experiments

RQ   *Does the practice of mindfulness impact conceptual modeling performance?*

This research question was split into two more concrete questions for the sake of experimental operationalization:

RQ$_1$ *Does the practice of mindfulness impact conceptual modeling quality?*

RQ$_2$ *Does the practice of mindfulness impact conceptual modeling productivity?*

In order to answer these research questions, an iterative process which consisted of a baseline experiment and two internal replications was carried out during the first semesters of the 2013–2014 [17], 2014–2015 [18] and 2015–2016 academic years. In each iteration, some *design questions* were raised and the experimental settings were refined and improved introducing some changes, as depicted in Figure 1. We followed an approach of series of experiments similar to those applied in Medicine, where they are carried out to empirically validate a new treatment or drug, because an isolated experiment is not sufficient to establish a theory [19]. In the case of Software Engineering, replications are usually performed to gain confidence in the results of previous studies or to understand the sources of variability that influence a given result [20]. In the series of experiments, a group of students in the Software Engineering Degree at the University of Seville attended mindfulness sessions during several weeks, whereas a second group of students attended a *placebo* public speaking workshop. As shown in Figure 2, two conceptual modeling exercises were performed by the students, one before and another after the treatment, and their performance was compared in terms of quality and productivity.

Having published the results of the baseline experiment [17] and the first replication [18], the contribution of this article is twofold. On one hand, describing the iterative process and the evolution of the experimental protocol, including descriptions and motivations of all the performed changes, so other researches can replicate it totally or partially using this information and the lab–pack available at https://exemplar.us.es/demo/BernardezMindfulnessTSE. Note that the number of replications in Software Engineering is still low and the lack of lab–packs is one of the main causes, as commented in [21].

On the other hand, presenting not only a narrative synthesis [22] of the results of the three individual experiments but also a thorough joint analysis of all the data collected during the 3–year study. The joint analysis has made it possible not only to detect the small effect of mindfulness on conceptual modeling quality, which went unnoticed in the analysis of the individual experiments, but also to answer some design questions raised during the series of experiments that could only be answered using the joint data from the three experiments.

The rest of the article is organized as follows. In Section 2, mindfulness is briefly described and the benefits of its practice in various contexts are commented. In Section 3, the experimental settings common to the series of experiments are described. In Section 4, a summary of the results of the individual experiments together with the motivation and descriptions of the changes in the experimental protocol are presented. In Section 5, the joint analysis of the series of experiments is discussed. The threats to validity are analyzed in Section 6 and the related work is commented in Section 7. Finally, the lessons learned are discussed in Section 8 and the conclusions and the future work are discussed in Section

Table 1
Usual Steps for a Mindfulness Session

Step	Description
1	Imagine a thread extending from the top of your head, pulling your back, neck and head straight up towards the ceiling in a straight line. Sit tall.
2	Use a timer to set a time limit, e.g. 10 or 15 minutes.
3	Close your eyes and scan your body, focusing and relaxing each body part one at a time, from your feet to your head.
4	Take three slow abdominal deep breaths.
5	Begin to breathe normally, but focusing on your breathing.
6	When thoughts come to you, simply acknowledge them, set them aside, and return your attention to your breathing.
7	Enjoy the rare chance to let your mind simply *be*.
8	Once the timer has rung, undo the posture, bring your conscious attention back to your surroundings and open your eyes slowly.

9. For the sake of completeness, the supplemental material includes the unpublished details of the second replication.

2   MINDFULNESS: AN OVERVIEW

Mindfulness is a meditation technique with roots in Buddhism and other contemplative traditions but without religious connotations. A mindfulness session[2] consists of drawing away to a quiet place to meditate during a certain amount of time, e.g. starting with 3, 5, or 10 minutes, as recommended in [23] for beginners, and increasing duration as the familiarity with the practice augments. During a mindfulness session, meditators try to focus their attention only on one thing, usually breathing, which is the usual meditation support because of its unavoidability. When some thoughts come to mind, they must be acknowledged and put aside without judging, then returning the attention to breathing again. Based on [23], [24], the usual steps of a mindfulness session are summarized in Table 1.

Apart from other salutary effects [25], one of the benefits of mindfulness is to transfer the state of consciousness achieved during meditation to ordinary activities, i.e. being aware and focused in daily life, staying in the present moment rather than rehashing the past or imagining the future. By regulating attention and reducing mental wandering while performing tasks, it helps us to perceive our environment clearly and to solve problems more efficiently.

Some research findings in Psychology and Neuroscience recently compiled in [26] reveal that even a few weeks of continued practice of mindfulness is sufficient to produce changes in the level of mind–wandering, i.e. practitioners become more focused and their objective parameters of attention begin to change. At a neuroscientific level, the practice of mindfulness is seen as a systematic training that can transform some of our mental habits by means of *neuroplasticity*, i.e. the continuous modification of our

brain through millions of cell–to–cell synaptic connections in response to our daily experiences [27].

2.1   Psycho–Social Benefits of Mindfulness

Apart from the benefits for stress reduction reported by Kabat–Zinn [28], other mindfulness–based therapeutic programs have been successfully applied to individuals prone to anxiety and other chronic diseases [29], [30]. For example, in [31] 136 heterogeneous patients were studied, showing that after two months of daily 20–minute practice, a significant percentage experienced better personal well–being in terms of mental clarity, equanimity, and self–compassion. For the interested reader, many other studies about the benefits of mindfulness at a personal level have been compiled by Sedlmeier et al. in [25].

With respect to social aspects, the main benefits of mindfulness are related to empathy, emotion regulation and emotional intelligence in general, as reported in [8] among others. Benefits for labor relations, especially in stressful working areas like health or teaching, have also been reported [30], [32], showing lower levels of emotional exhaustion, improved life satisfaction and self–efficacy in those practicing mindfulness.

2.2   Mindfulness in Higher Education

The interest in integrating meditation into higher education is growing [13]. The practice of mindfulness in university classrooms began with the objective of improving some skills like, in the case of Law students, listening comprehension, legal conflict resolution and negotiation abilities [33]. Several approaches have been applied for introducing mindfulness in higher education. Some use breathing as a support, while others use sounds or a visual object. Some professors invite their students to meditate some minutes at the beginning and end of their lessons while others prefer meditation retreats of several days, weekly sessions, meditation readings, etc. [33]. For example, the Harvard University offers guided meditations to its students several times a week through the *Your Wellbeing* program [34]. Weekly meditation sessions, courses, and retreats are also offered at the UC Berkeley School of Law [35]. Other universities like Texas at Austin and Cambridge offer similar services such as the *MindBody Labs* [36] or *Mindfulness at Cam* [37] respectively. Specifically at Medical Schools in the USA, a recent survey has reported that 79% of them offer some mindfulness–related activities [38].

Apart from the aforementioned introductory programs, some empirical studies about mindfulness in higher education with healthy students, i.e. students without high levels of anxiety or stress, have been reported. For example, the outcomes of a qualitative study [39] about the influence of mindfulness in a 15–week course with graduate students included an increase of their mental clarity, organization, awareness, and acceptance of emotions and personal issues. In [15], a controlled experiment based on Graduate Record Examinations showed a great improvement in the group that attended 15 mindfulness sessions. In [12], three experiments performed in several USA universities showed that the practice of mindfulness improves student knowledge retention during lessons.

---

2. According to [16], there are three ways of practicing mindfulness: *focused attention* (FA), *open monitoring* (OM) and *ethical enhancement* (EE). Since FA is the usual practice for beginners and it is oriented to educate attention, we decided to use it with our students in the series of experiments.

2.3 Mindfulness in Software Engineering

In some software industries, the practice of mindfulness is fostered claiming improvements in employee relationships, such as reacting less emotionally, or in concentration [40]. Particularly in Google, engineer Chade–Men Tan has developed the *Search Inside Yourself* mindfulness course which is offered since 2007, has a six–month wait list and has been taken by more than 1,500 employees [41]. The course consists of 19 sessions or an intensive two–and–a–half day retreat, and it has three main modules: attention training, self–knowledge development, and creating mental habits [42]. Participants of the program have reported being calmer, more patient, and better able to listen. They also say the program helped them better handle stress and defuse emotions [43]. Following the same trend, Intel began offering its *Awake@Intel* mindfulness program in 2012. On average, participants reported a decrease in stress and an increase in well–being, mental clarity, creativity, the ability to focus and the quality of relationships at work [10].

In the agile software process community, not only the practice of mindfulness has been recommended in order to create a good atmosphere in work groups, meetings, and interactions with customers and users [44], [45], but also an empirical study on the effects of mindfulness in agile developers has been recently carried out reporting positive effects [46] (see Section 7 for details).

3 COMMON EXPERIMENTAL SETTINGS

Following [47], [48], the settings common to the series of experiments are described in this section so other researchers interested in replicating this study, or performing a similar one, can have all the needed information. Differential settings of each experiment in the series and the evolution of the experimental protocol are described in the next section.

For the sake of readability, the baseline experiment and the two internal replications are referred to as MIND#1, MIND#2 and MIND#3 respectively in the rest of the article.

3.1 Goals

Following [49], the Goal–Question–Metric template is used to describe the main goal of the series of experiments according to our main research question:

Analyze the practice of mindfulness
for the purpose of evaluating its effects
with respect to conceptual modeling performance
from the point of view of the researchers
in the context of second–year students of the Degree in Software Engineering at the University of Seville.

3.2 Participants

In the series of experiments, all subjects were second–year students of the Degree in Software Engineering at the University of Seville. They were enrolled in the *Introduction to Software Engineering and Information Systems* (ISEIS) annual course, in which they learn about conceptual modeling for the first time.

All subjects were asked whether they had previously practiced mindfulness or not in order to discard experienced

Table 2
Participant Flow in the Series of Experiments. Numbers in square brackets show the percentage with respect to the number of initially invited students. Numbers within parenthesis show the number of subjects in the experimental and control group respectively.

Exp.	ISEIS students	Interested students	Pre–exercise	Post–exercise	Sample
MIND#1	87 [100%]	75 [86%] (38+37)	42 [48%] (18+24)	35 [40%] (17+18)	32 [37%] (16+16)
MIND#2	95 [100%]	84 [88%] (40+42)	82 [86%] (40+42)	57 [60%] (28+29)	53 [56%] (27+26)
MIND#3	103 [100%]	71 [69%] (36+35)	61 [59%] (30+31)	49 [46%] (26+23)	45 [44%] (23+22)

practitioners that could be blurred in the treatment effect, although no subject was discarded for this reason in any of the experiments. On the other hand, subjects with prior practice in conceptual modeling would have performed better than the rest. In order to avoid this situation and have a sample as homogeneous as possible, 6 repeater students were discarded from the samples (1 out of 33 in MIND#1, 3 out of 56 in MIND#2 and 2 out of 47 in MIND#3). For the same reason, students who had worked for software development companies or participated in software projects applying conceptual modeling were also discarded. Last but not least, students who claimed to be in treatment for anxiety were also discarded from the sample (only one student was discarded for this reason).

Table 2 shows the flow of participants in the series of experiments, indicating the number of ISEIS students who (i) were invited to freely participate, (ii) accepted participating, (iii) performed the pre–treatment exercise, (iv) performed the post–treatment exercise and (v) fulfilled the following criteria: (a) to attend at least 75% of the mindfulness sessions, (b) to do both pre– and post–treatment exercises and (c) to create a draft of information requirements and used the right notation in the conceptual modeling exercises, i.e. UML class diagrams.

3.3 Experimental Process

The experimental process, which is graphically represented in Figure 3, is described in the following subsections.

3.3.1 Student Training

During the execution of the experiments, not only did the participants take the scheduled ISEIS lessons, but they also worked in teams on their semester projects, where they allocate themselves into teams and look for a real organization, usually a small business or a nonprofit organization. Then, they perform a requirements elicitation process (interviews, document analysis, etc.), develop a draft requirements specification and a conceptual model, and transform the conceptual model into a relational database schema.
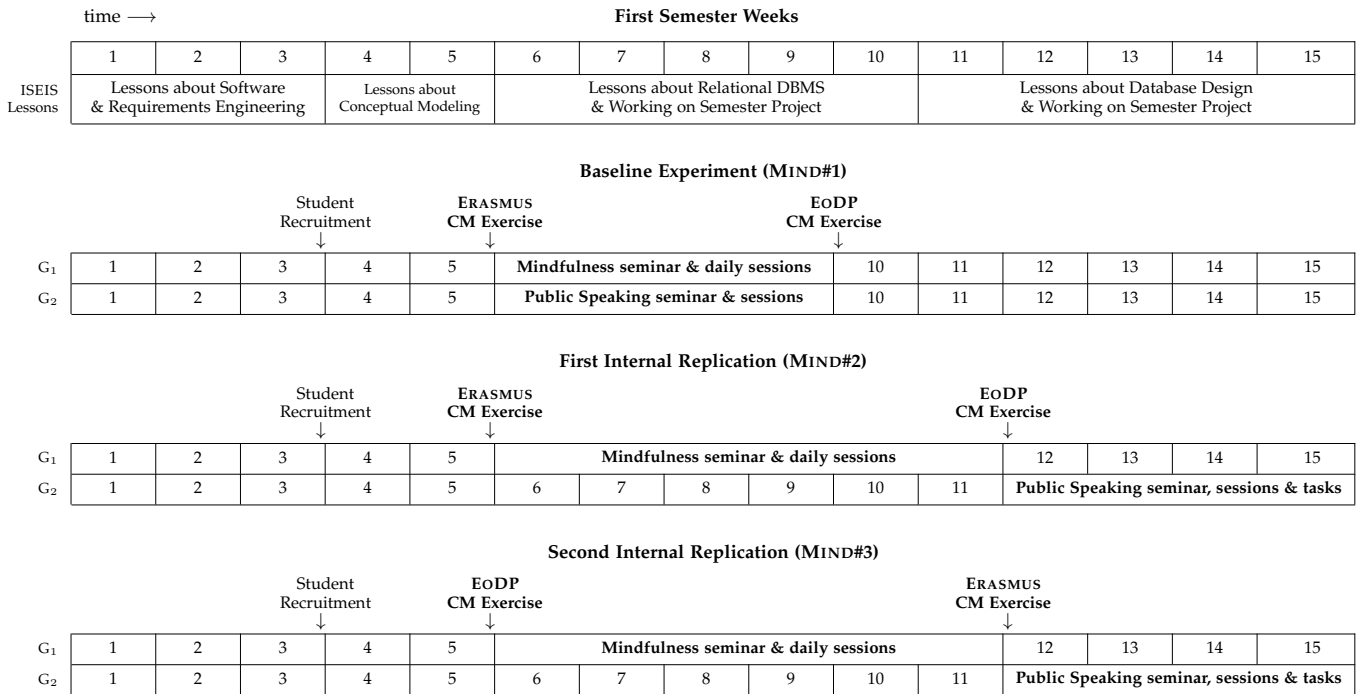
time →						First Semester Weeks								
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

| ISEIS Lessons | Lessons about Software & Requirements Engineering | | | Lessons about Conceptual Modeling | | Lessons about Relational DBMS & Working on Semester Project | | | | | Lessons about Database Design & Working on Semester Project | | | | |

Baseline Experiment (MIND#1)

			Student Recruitment ↓		ERASMUS CM Exercise ↓				EoDP CM Exercise ↓					
G_1	1	2	3	4	5	Mindfulness seminar & daily sessions				10	11	12	13	14	15
G_2	1	2	3	4	5	Public Speaking seminar & sessions				10	11	12	13	14	15

First Internal Replication (MIND#2)

			Student Recruitment ↓		ERASMUS CM Exercise ↓						EoDP CM Exercise ↓			
G_1	1	2	3	4	5	Mindfulness seminar & daily sessions						12	13	14	15
G_2	1	2	3	4	5	6	7	8	9	10	11	Public Speaking seminar, sessions & tasks			

Second Internal Replication (MIND#3)

			Student Recruitment ↓		EoDP CM Exercise ↓						ERASMUS CM Exercise ↓			
G_1	1	2	3	4	5	Mindfulness seminar & daily sessions						12	13	14	15
G_2	1	2	3	4	5	6	7	8	9	10	11	Public Speaking seminar, sessions & tasks			

Figure 3. Experimental Processes across the Series of Experiments. G_1 and G_2 are the experimental and control groups respectively.

3.3.2 Student Recruitment

Student recruitment took place during the third week of the first semester. It consisted of a motivating presentation about (i) participating voluntarily in the ongoing research, (ii) the personal and professional benefits of learning non–technical skills such as mindfulness and public speaking, (iii) the expected commitment as a participant, i.e. attending the workshop sessions and doing both conceptual modeling exercises, and (iv) the 5% bonus for doing so, i.e. half a point on a 10–point scale, which was granted only if the student passed the ISEIS course. After the initial presentation, all the students filled out manually a questionnaire in which they stated their interest in the proposed workshops and their degree of commitment.

It is worthwhile to mention that in order to avoid bias, the concept of being subjects of an ongoing experiment was intentionally not mentioned either during the recruitment or during the experiment runs. They were informed that they could participate in the ongoing research—without specifying goals or topics of the research—and that the training workshops were interesting for their professional lives, not mentioning that they could improve their conceptual modeling performance in any way. They were also informed of the two requirements to obtain the bonus. The first one was to perform some conceptual modeling exercises during the scheduled lessons, although their ISEIS grades would not be affected for the results. The second requirement was to attend at least 75% of the workshop sessions, which would be controlled using the sign–in sheets designed for that purpose.

3.3.3 Task 1: Pre–Treatment Exercise

At the end of the fifth week of the semester, the two groups of subjects did the *pre–treatment* exercise the same

day during a 2–hour lesson. The aim of the exercise was to develop a conceptual model after analyzing a transcript of an interview in which a requirements engineer asks a customer about some problem domain concepts, the way they currently perform their business tasks and their expectations about the information system to be developed. An excerpt from one of the exercises is shown in Figure 4.

The steps taken during the performance of this task were the following:

1) The interview transcript and two blank sheets of paper are handed in to the students.
2) A volunteer is requested among the students to read the transcript aloud together with the professor, each of them playing the role of the requirements engineer and the customer respectively.
3) The students write down the beginning time of the exercise and start developing the conceptual model by creating a draft of the explicit and implicit information requirements in the interview transcript. Then, they identify the model elements and develop the corresponding UML class diagrams.
4) Once they have finished, the students write down the ending time and hand their exercises and the interview transcript out to the professor.

3.3.4 Treatment Administration

The next week after the pre–treatment exercise, all subjects attended an introductory seminar about the workshop corresponding to their groups. Those seminars took about one hour and they were delivered out of ordinary schedule. Then, the treatment sessions were delivered during the recess between classes and took place always in the same conditions for each group, i.e. same classroom, same hour.

> **Q**: *Let's start by the time a student at the University of Seville (USE) thinks about applying for an Erasmus grant.*
> **A**: *Well, interested students usually browse and query the Erasmus destinations offered by their centers (i.e. faculties or schools) for the next academic year. Each destination has associated information from the host university (e.g. University of Berlin), the student profile (e.g. "last–course students with less than 60 credits left"), the number of students accepted in the exchange, the number of months to stay, etc. If a student finds an appealing destination that fits her profile, then she must submit an Erasmus Program Application (EPA) to the International Relations Service (IRS) of the USE. We want both the browsing of destinations and the submission of EPAs to be made using the web application to be developed.*
> **Q**: *What information is recorded in the EPA?*
> **A**: *Basically, student data (Tax ID number, name, address, studies in which is enrolled, etc.), requested destinations (up to 12 destinations can be applied for, indicating preference), and qualified foreign languages.*
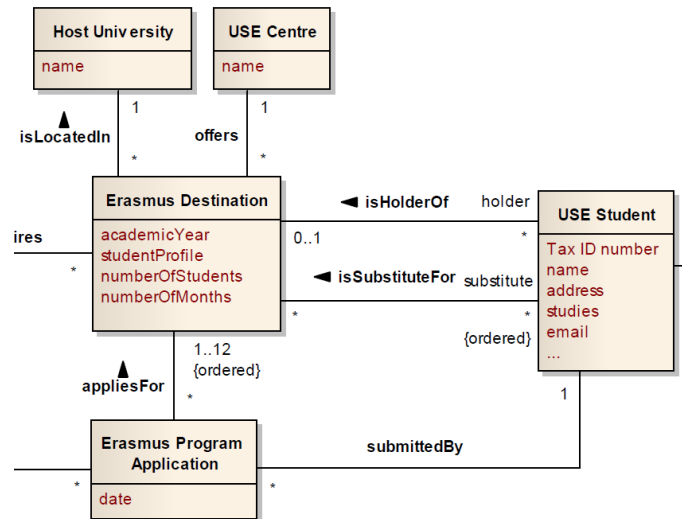
Figure 4. Excerpt from one of the conceptual modeling exercises (ERASMUS) used in the series of experiments (borrowed from [18])

In the mindfulness workshops, the sessions were face–to–face, four days a week. All the sessions followed the same dynamics: the students and the researcher responsible for conducting the session met in a classroom; they all sat down, lights were turned off and curtains were drawn letting only some dim light in the room; when they all were in silence, an alarm was programmed; during the first five minutes, the subjects were guided in their body scan; then, during the remaining time, they were invited to focus solely on their breathing. Sometimes, the researcher asked "where is your mind now?" in order to re–focus them on breathing. In the event some students were late, they were instructed to enter the room making as less noise as possible and sit on one of the chairs that were intentionally left empty near the door.

In the public speaking workshops, the subjects were given some basic guidelines on how to prepare a talk, some notions on non–verbal communication and some seminal talks were commented. Later, they were invited to look for related videos on the Internet and to prepare a script for a public presentation on a topic of their interest.

3.3.5 Task 2: Post–treatment Exercise

Once the mindfulness workshop was finished, both groups did the *post–treatment* exercise in separate classrooms. They followed the same procedure as described for Task 1 with the only difference that in the experimental group the initial 15 minutes were dedicated to group meditation. Then, the subjects did the exercise individually.

3.4 Experimental Material

The material used in the conduction of the series of experiments, available in the lab–pack at https://exemplar.us.es/demo/BernardezMindfulnessTSE, consisted of the following items:

- The motivating presentation slides used for student recruitment.

- The questionnaire used for student recruitment.
- The seminar slides.
- The conceptual modeling exercises and their corresponding solutions used in task 1 and task 2, about ERASMUS grants and *End–of–Degree* (EOD) projects, which were selected from a number of conceptual modeling exercises that we have been using successfully in our Requirements Engineering courses for more than 10 years and are therefore extensively checked and reviewed.

3.5 Factors

The two factors, i.e. independent variables, in the series of experiments are the following:

group It represents the *group* the subjects were assigned during the training workshops, i.e. the *mindfulness* group and the *control* group.

time It represents the moment in *time* when the subjects were measured by performing the conceptual modeling exercises, i.e. *pre–treatment* and *post–treatment*.

3.6 Response Variables and Metrics

According to research questions RQ_{1-2}, conceptual modeling quality and productivity were respectively measured using the response variables *effectiveness* and *efficiency*. These two variables are based on *semantic quality* (SEM_Q), which refers to how faithfully the modeled system is represented [50]. In our case, how similar the models developed by the students were with respect to a *reference* model developed by the researchers. Assuming UML class diagrams as the modeling language, model elements were considered as correctly identified if there existed identical or semantically equivalent counterparts in the reference model. The metric for computing SEM_Q was the following:

$$SEM_Q = CLASS_{OK} - \frac{CLASS_{KO}}{2} + ASSOC_{OK} + ATTR_{OK}$$

in which CLASS_{OK}, ASSOC_{OK}, and ATTR_{OK} are the number of correctly identified classes, associations, and attributes respectively. CLASS_{KO} is the number of incorrectly identified classes, a correction factor introduced to penalize spurious model elements. Note that since SEM_Q is a ratio variable, its minimum valid value is 0, i.e. in the rare case the expression above took a negative value, its value would be assigned to zero.

Once SEM_Q is defined, conceptual modeling effectiveness is defined as the percentage of semantic quality achieved by a subject:

$$\text{EFFECTIVENESS} = \frac{\text{SEM}_Q}{\text{CLASS}_R + \text{ASSOC}_R + \text{ATTR}_R}$$

where CLASS_R, ASSOC_R, ATTR_R are respectively the number of classes, associations, and attributes in the reference model, i.e. their sum is the maximum value of semantic quality.

Consequently, conceptual modeling efficiency is defined as the quotient between the achieved semantic quality and the amount of time in minutes spent by a subject in finishing a conceptual modeling exercise:

$$\text{EFFICIENCY} = \frac{\text{SEM}_Q}{\text{T}_{end} - \text{T}_{begin}}$$

3.7 Context Variables

Some context variables or parameters were identified during the series of experiments. As recommended in [51], parameters and how they were controlled are defined below in order to facilitate experiment replication.

3.7.1 ISEIS *Scheduled Lessons Taught to Subjects*

To avoid any differences in the ISEIS lessons taught to the subjects, all of them had the same professors and the same content taught at the same pace.

3.7.2 *Complexity and Familiarity of the Exercises*

To properly compare the results of the conceptual modeling exercises before and after the treatment, they had to be of similar complexity and the level of familiarity of the subjects with the problem domains had to be similar as well. Both exercises were chosen with a similar complexity, which was considered using the metrics in Table 3. Considering the limited time the subjects had to develop the exercises (2 hours), unfamiliar problem domains could have caused the outcome to be skewed. Since all participants were potential candidates for an ERASMUS grant and had to develop an EOD project (EODP) to finish their studies, the two conceptual modeling exercises were about hypothetical information systems for the management of ERASMUS grants and EOD projects (see the lab–pack at https://exemplar.us.es/demo/BernardezMindfulnessTSE for details).

3.8 Experimental Design and Data Analysis

Considering the experimental process and the identified variables, the chosen experimental design was the 2×2 *mixed factorial* [52], also known as *pre–post* design. In this design—which is common in Medicine or Psychology when the evolution of patients under a given therapeutic treatment needs to be studied—each subject is assigned to one

Table 3
Structural Measures of the Conceptual Modeling Exercises

Structural measures	ERASMUS	EODP
Number of words in the interview transcript	951	1223
Number of classes (CLASS_R)	8	8
Number of associations (ASSOC_R)	10	10
Number of attributes (ATTR_R)	17	24
Average number of attributes per class	2,29	3

single treatment, usually including a placebo, and two repeated measures on the response variables are taken before and after the treatment administration.

In our case, the two factors with two levels are *time*, which is a *within–subjects* factor, i.e. each subject is tested at each level of the factor; and *group*, which is a *between–subjects* factor, i.e. different groups of subjects are used for both levels of the factor. Since all the experiments in the series followed the pre–post experimental design, the effect of the treatment over time was consequently analyzed using 2 (*group*) $\times$ 2 (*time*) mixed–model ANOVA tests, as recommended in [53], [54].

3.9 Statistical Hypotheses

The tested hypotheses associated with the pre–post design are enunciated below. Each group of two hypotheses corresponds to each response variable.

$H_{0,1}$: There is no difference in the conceptual modeling *effectiveness* of subjects between the pre– and post–treatment exercises.

$H_{0,2}$: There is no difference in the conceptual modeling *effectiveness* between the subjects who received the mindfulness treatment and those who did not.

$H_{0,3}$: There is no difference in the conceptual modeling *efficiency* of subjects between the pre– and post–treatment exercises.

$H_{0,4}$: There is no difference in the conceptual modeling *efficiency* between the subjects who received the mindfulness treatment and those who did not.

4 RESULTS AND PROTOCOL EVOLUTION

In this section, we discuss the differential settings (summarized in Table 4), the results[3] (summarized in Table 5), the changes in protocol and operationalization (enumerated in Table 6) and the additional design questions raised during the series of experiments. For the interested reader, detailed information on each experiment is available in [17] for MIND#1, in [18] for MIND#2 and in the supplemental material for MIND#3.

---

3. The results of an ANOVA test are usually reported as (F, p) pairs, where F is the test statistic and p is the significance level of the test, usually known as p–value. The numbers in parenthesis after F are the degrees of freedom of the *between–subjects* and *within–subjects* factors respectively.

Table 4
Differential Settings of the Experiments in the Series

Experimental setting	MIND#1	MIND#2	MIND#3
Subject assignment	Subject preference	Random	Random
Control group treatment	Public Speaking	Null	Null
Mindfulness period & session duration	4 weeks 10 min.	6 weeks 12 min.	6 weeks 12 min.
Pre–treatment exercise	ERASMUS	ERASMUS	EODP
Post–treatment exercise	EODP	EODP	ERASMUS

Table 5
Outcomes of the Experiments in the Series

Outcome	MIND#1	MIND#2	MIND#3
CM Effectiveness Improvement	Not significant ($F_{(1,30)} = 2.713$, $p = 0.110$)	Not significant ($F_{(1,51)} = 2.908$, $p = 0.094$)	Not significant ($F_{(1,43)} = 2.614$, $p = 0.113$)
CM Effectiveness Effect Size	Medium ($\eta_p^2 = 0.086$)	Small ($\eta_p^2 = 0.054$)	Small ($\eta_p^2 = 0.057$)
CM Efficiency Improvement	Highly significant ($F_{(1,30)} = 17.001$, $p = 0.000$)	Highly significant ($F_{(1,51)} = 8.698$, $p = 0.005$)	Highly significant ($F_{(1,43)} = 61.602$, $p = 0.000$)
CM Efficiency Effect Size	Large ($\eta_p^2 = 0.370$)	Large ($\eta_p^2 = 0.146$)	Large ($\eta_p^2 = 0.590$)
Sample size	32	53	45

4.1 Differential Settings and Results of MIND#1

As summarized in Table 4, in the baseline experiment, students were assigned to groups according to their preferences in order to minimize dropout. As shown in Figure 3, the subjects in the control group attended the placebo public speaking workshop at the same time the subjects in the experimental group attended the mindfulness sessions, so all of them felt they were treated equally. The mindfulness workshop took place over 4 weeks with a duration of 10 minutes per session, which seemed to be a reasonable amount of time at that moment. With respect to the order in which the conceptual modeling exercises were performed in MIND#1, it was set for no particular reason as the pre–treatment exercise being about ERASMUS grants and the post–treatment one about EOD project management.

The box and profile plots corresponding to the results of MIND#1 are depicted in Figures 5a and 6a. As shown in Table 5, the mixed–model ANOVA tests for two factors performed in MIND#1[4] showed that, although the effect of mindfulness on effectiveness was positive and *medium*

according to the guidelines[5] in [55], there was no evidence to support that the subjects who practiced mindfulness were more effective, since the differences generated by the treatment did not reach statistical significance ($F_{(1,30)} = 2.713$, $p = 0.110$). Conversely, the subjects who practiced mindfulness were significantly more efficient than the subjects who did not, probably because of the skills improved during the mindfulness workshop. Specifically, the effect size on efficiency was *large* and the differences generated by the treatment were highly significant ($F_{(1,30)} = 17.001$, $p < 0.01$).

Also worth mentioning is that we observed that the students were very interested and participated actively in the experimental tasks regardless of the workshop they attended—some of them even asked to participate in both workshops.

4.2 Protocol Evolution after MIND#1

Although both workshops were carefully presented as equally interesting during the student recruitment task (see Section 3.3.2), it was possible that the students who chose the mindfulness workshop were more motivated than those who chose the public speaking workshop, i.e. there was a selection and assignment bias threat to the validity of the experiment. As a result, the first design question was raised:

DQ$_1$ *Is students' motivation relevant?*

i.e. is there any difference in the dropout percentage or the effect size when the students practice mindfulness of their own volition than when they practice it by random assignment? In order to (i) answer this question, (ii) mitigate the aforementioned threat, and (iii) avoid the limitations in experimental design and data analysis due to not having a random assignment of subjects to groups, we decided to change the experiment protocol and use random assignment in subsequent replications (see CH$_1$ in Table 6).

The second evolutionary change after MIND#1 was the result of the feedback obtained after its presentation at the ESEM conference [17], where some questions were posed about the potential effects of the public speaking workshop in the experiment outcomes, despite our original intention of using it as a placebo. An additional design question was therefore raised:

DQ$_2$ *Is public speaking actually a placebo?*

i.e. is there any difference in the effect size when the control group receives public speaking training as a placebo than when they receive no treatment at all? In order to answer this question and discard public speaking as an alternative treatment, we decided to postpone the public speaking workshop after the post–treatment exercise in future replications, so the control group was formally administered a *null* treatment (see CH$_2$ in Table 6). We decided to postpone the workshop rather than eliminate it to prevent students from feeling that they were not all being treated in the same way and they all were learning useful skills for their professional life regardless of the assigned group.

The motivation behind the third change was to assess whether the lack of statistically significant results in conceptual modeling effectiveness in MIND#1 was motivated

---

4. Assumptions of normality and homogeneity of variances were verified before applying ANOVA tests in the three experiments.

5. In [55], the thresholds for classifying effect sizes according to η_p^2 are 0.01 for *small*, 0.06 for *medium* and 0.14 for *large*.

Table 6
Summary of Changes in the Replications (MIND#2 and MIND#3) with respect to the Baseline Experiment (MIND#1)

| Change Number	Change Description	Changed in	Change Motivation	Altered Dimension	Design Question
CH_1	Assignment of subjects to experimental groups randomized	MIND#2 MIND#3	To mitigate the selection and assignment bias threat and avoid statistical analysis limitations	Protocol	DQ_1
CH_2	Public Speaking workshop postponed after post–treatment exercise	MIND#2 MIND#3	To mitigate the potential placebo disturbing factor on experiment outcomes	Operationalization	DQ_2
CH_3	Number and duration of mindfulness sessions increased	MIND#2 MIND#3	To achieve a statistically significant improvement of conceptual modeling effectiveness	Operationalization	DQ_3
CH_4	Order of the conceptual modeling exercises swapped	MIND#3	To study the potential impact of the conceptual modeling exercise on experiment outcomes	Protocol	DQ_4
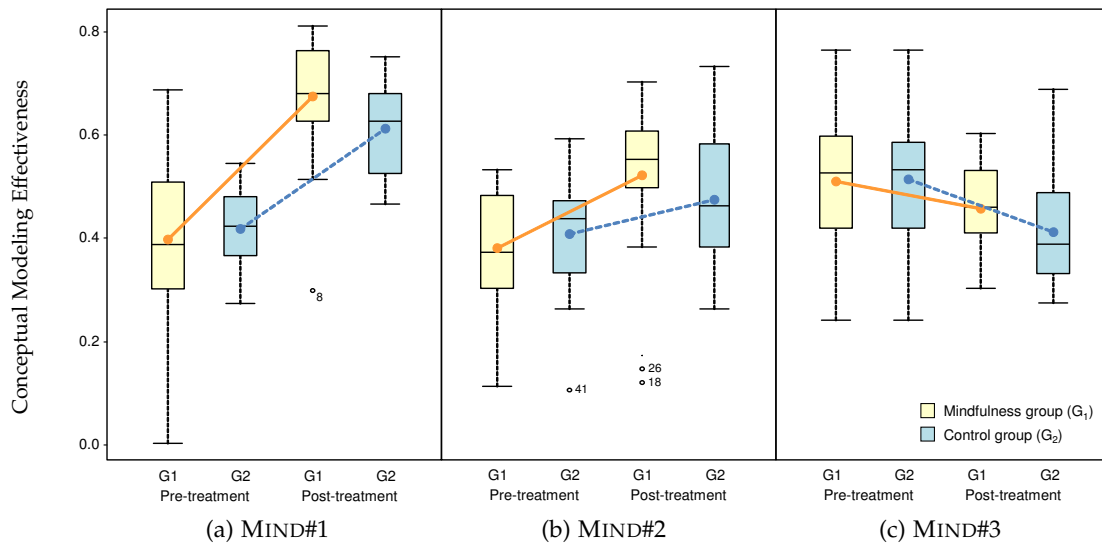
Figure 5. Box and Profile Plots of Conceptual Modeling Effectiveness in the Series of Experiments
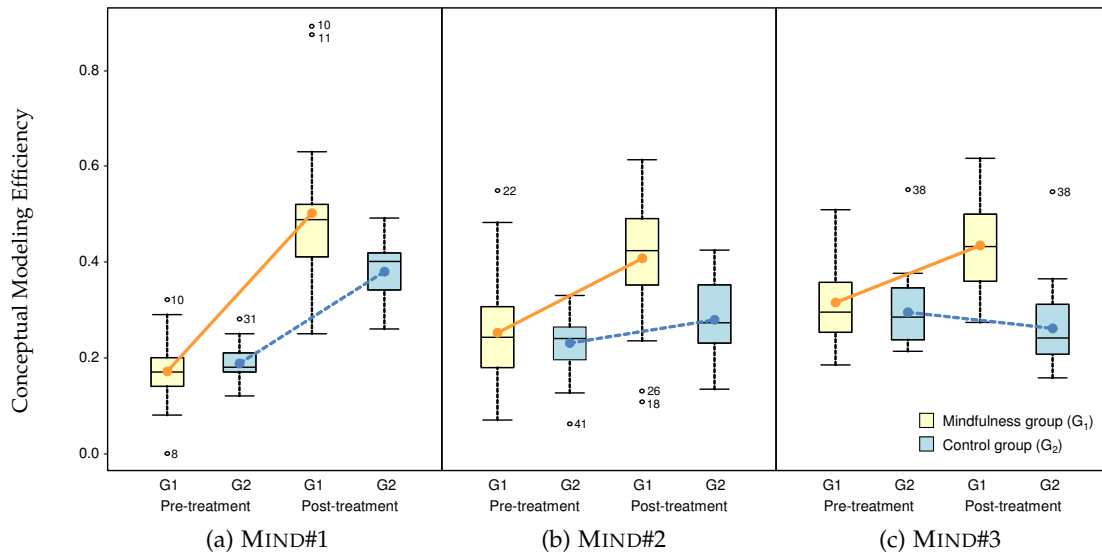

Figure 6. Box and Profile Plots of Conceptual Modeling Efficiency in the Series of Experiments

by an insufficient number and duration of the mindfulness sessions. As a result, we raised another design question:

DQ$_3$ *Has the mindfulness treatment been long enough?*

i.e. is there any difference in the effect size (especially in effectiveness) when the students receive longer mindfulness sessions during more weeks? To answer this question, we decided to extend the mindfulness workshop from 4 to 6 weeks and sessions from 10 to 12 minutes in future replications (see CH$_3$ in Table 6). With respect to the number of weeks, it was limited not only by the required lessons on conceptual modeling so the students could perform the exercises, but also by the Christmas break, which would interrupt the mindfulness sessions. Therefore, the maximum number of weeks was limited to six weeks, from week 6 to week 11 (see Figure 3). Regarding the duration of mindfulness sessions, we were limited by the 20–minute recess between lessons. Considering that students had to come to the meditation room and go back to their classrooms, mindfulness sessions had to be shorter than 15 minutes.

4.3   Differential Settings and Results of MIND#2

The result of applying changes CH$_1$ to CH$_3$ in MIND#2 resulted in its differential settings consisting of a random assignment of subjects to groups, a postponed public speaking workshop, i.e. a null treatment for the control group, and a mindfulness workshop extended to 6 weeks with 12–minute sessions. All the other experiment settings were the same as in MIND#1.

In MIND#2, the tests showed similar outcomes to those of MIND#1, i.e. a *small* (but close to *medium*) positive effect of mindfulness on conceptual modeling effectiveness that was not statistically significant either (F(1,51) = 2.908, $p = 0.094$) and a *large* effect that was again highly significant on efficiency (F(1,51) = 8.698, $p < 0.01$).

As can be seen in Figures 5a–5b and 6a–6b, the absolute values of effectiveness and efficiency in MIND#2 were lower than in MIND#1. To identify whether there were significant differences between MIND#1 and MIND#2 subjects, two two–sample t–tests were performed and highly significant differences were detected in both response variables, i.e. the conceptual modeling performance of the sample in MIND#1 was significantly better than in MIND#2. Considering that the contents and pace of the ISEIS lessons were the same in both experiments, and that the average grade and the passing percentage were 6.2 and 81% in MIND#1 whereas they were 5.2 and 73% respectively in MIND#2, the detected dissimilarity was attributed to the intrinsic students' variability on each academic year and the *random sampling effect* [56] due to the difference of sample sizes between MIND#1 and MIND#2 (32 and 53 subjects respectively).

With respect to the design questions about the relevance of the students' motivation (DQ$_1$), the public speaking workshop being an actual placebo (DQ$_2$) and the treatment being long enough (DQ$_3$), and considering (i) that the dropout percentage was lower in MIND#2 (33.33%) than in MIND#1 (56.00%, see *Dropout* paragraph in Section 6.2); and (ii) that the outcomes were similar not only in statistical significance but also in effect sizes (see Table 5), we concluded that, in MIND#2, there was no evidence to consider

either students' initial motivation or the effect of the public speaking workshop as relevant factors. Similarly, there was also no evidence to consider that increasing the duration of the mindfulness workshop and sessions from 4 to 6 weeks and 10 to 12 minutes respectively was a relevant factor.

4.4   Protocol Evolution after MIND#2

In spite of the outcomes of MIND#2 regarding design questions DQ$_1$ (students' motivation), DQ$_2$ (placebo control condition) and DQ$_3$ (duration of treatment), we decided to keep changes CH$_1$ to CH$_3$ in the experiment protocol in order to confirm the obtained results in the second replication.

Besides that, both results of MIND#1–2 showed highly significant differences in both response variables before and after the intervention ($p < 0.01$ in all cases). Although these results might be due to the combined effect of students' learning along the ISEIS course and the treatment, an additional design question was raised in order to check whether some unnoticed characteristics of the experimental tasks, i.e. the conceptual modeling exercises, had also had some effect on the response variables.

DQ$_4$ *Do the conceptual modeling exercises used in the experimental tasks have any influence on the results?*

To answer this question, we decided to change the order of the conceptual modeling exercises in the second replication (see CH$_4$ in Table 6).

4.5   Differential Settings and Results of MIND#3

The only differential setting in MIND#3 with respect to MIND#2 was the swapping of the pre– and post–treatment exercises, i.e. the former was about EOD projects and the latter about ERASMUS grants.

In MIND#3, the mixed–model ANOVA tests showed similar results to those of MIND#2, i.e. an almost *medium* but *small* effect of mindfulness on effectiveness that was not statistically significant (F(1,43) = 2.614, $p = 0.113$), and a *large* effect on efficiency that was highly significant (F(1,43) = 61.602, $p < 0.01$).

Nevertheless, as shown in Figures 5c–6c, the average effectiveness in the post–treatment exercise was worse than in the pre–treatment exercise, and the efficiency increased for the subjects who practiced mindfulness but it decreased subtly for the others. Considering DQ$_4$, we interpreted this result was likely due to the fact that the EODP exercise was somehow easier to solve for the subjects than the ERASMUS exercise, despite their similar complexity measures (see Table 3), i.e. that the conceptual modeling exercises—or at least, the order in which they were performed—might have had a relevant influence on the experiment outcomes. However, the practice of mindfulness still seemed to be beneficial, since the experimental group scored higher than the control group, i.e. their effectiveness was impacted less by the conceptual modeling exercises, and their efficiency increased after the treatment whereas it decreased slightly for the control group.

5 JOINT ANALYSIS

Given the results of the individual experiments, since the power of statistical tests is strongly reinforced when the sample size increases, we decided to perform a joint analysis of the data to test (i) whether the improvement in conceptual modeling effectiveness was actually so subtle or just negligible; and (ii) whether the order in which the conceptual modeling exercises were performed actually had the influence questioned in DQ_4.

We used the two more common approaches to the joint analysis of a series of experiments [57], i.e. *pooled data meta–analysis*, in which the data of several studies are analyzed as if they come from a single study; and *aggregated data meta–analysis*, which usually focuses on the analysis of the effect size and its variation among studies.

5.1 Pooled Data Meta–Analysis

In pooled data meta–analysis, the datasets from individual studies are combined but introducing *moderator variables* to study the strength of the relationship between factors and response variables. Apart from improving the power of statistical tests, the increased sample size in this kind of meta–analysis enables the detection of smaller significant effects.

5.1.1 Moderator Variables for Pooled Data Meta–Analysis

One moderator variable, *ord*, representing the order in which the conceptual modeling exercises were performed in each experiment,[6] was introduced to study its potential influence on the results, as questioned in DQ_4. The two levels for this moderator variable were ERASMUS→EODP and EODP→ERASMUS.

5.1.2 Hypotheses for Pooled Data Meta–Analysis

Considering the moderator variable introduced, the pooled data meta–analysis tests all the hypotheses in Section 3.9 together with the following ones:

$H_{0,5}$: There is no difference in the conceptual modeling effectiveness between the subjects who performed the exercises in the ERASMUS →EODP order and those who performed them in the inverse order.

$H_{0,6}$: There is no difference in the conceptual modeling efficiency between the subjects who performed the exercises in the ERASMUS→EODP order and those who performed them in the inverse order.

5.1.3 Descriptive Statistics of Pooled Data

The descriptive statistics of the pooled data under different experimental conditions are shown in Table 7 and the distributions of the response variables are depicted as box plots in Figures 7 and 8, where it can be seen that after the intervention, the maximum, minimum, medians and intermediate quartiles have higher values in the mindfulness group than in the control group.

In the case of effectiveness, Table 7 shows a moderate increase of 8.16% of the mindfulness group with respect

6. Hereinafter referred to as *the order* for the sake of readability.

to the control group (column 2, rows 15 and 16), whereas for efficiency the observed increase is 46.67% (column 8, same rows). This large difference in efficiency can be also seen in Figure 8, where the bottom of the post–treatment box of the mindfulness group has almost the same value than the top of the same box of the control group, i.e. after the intervention, approximately 75% of the subjects who practiced mindfulness were more efficient than 75% of the subjects who did not.

Note also that both response variables were on average higher for the EODP exercise than for the ERASMUS exercise (columns 2 and 8, rows 1–2), as questioned in DQ_4. Nevertheless, it is also remarkable that mindfulness had a positive impact in all the experiments, as can be seen in the medians of both response variables (columns 4 and 10, rows 23–28).

5.1.4 Hypothesis Testing for Pooled Data Meta–Analysis

After performing the normality and homoscedasticity tests to determine the applicability of either parametric or non–parametric statistical tests, two mixed–model ANOVAs were carried out for each response variable[7].

On one hand, consistently with the analysis performed in the individual experiments, a 2 (*group*) × 2 (*time*) mixed–model ANOVA was carried out to observe the effect of the treatment as described by the *group*time* interaction.

On the other hand, with the goal of answering DQ_4, a three–factor 2 (*ord*) × 2 (*group*) × 2 (*time*) mixed–model ANOVA was conducted to observe the effect of the order and its influence on the effect of the treatment, as described by the *ord*time* and *ord*group*time* interactions respectively. In the light of the results of the three–factor ANOVA, two additional analyses were performed for each response variable. The first one was a Bayesian analysis to test whether the triple *ord*group*time* interaction was negligible or not. The second one was a 2 (*group*) × 2 (*time*) mixed–model ANCOVA with *ord* as a covariate to study the effect of the treatment while controlling for the effect of the order.

Tests for Conceptual Modeling Effectiveness. The results of the 2×2 mixed–model ANOVA in Table 8 show the most relevant finding of the pooled data meta–analysis, namely that the practice of mindfulness has a *small* but statistically significant effect on conceptual modeling effectiveness, as described by the *group*time* interaction (F(1,128) = 4.341, $p < 0.05$). This relevant finding contrasts with the results of the three individual studies, where the observed practical difference in effectiveness did not reach statistical significance, probably due to the combination of small effects and sample sizes. This outcome shows how the increased sample size in a pooled data meta–analysis can detect small effects that go unnoticed in isolated experiments.

Regarding the effect of the order, the results of the 2×2×2 mixed–model ANOVA in Table 9 show that, although it has a highly significant effect, as described by the *ord*time* interaction (F(1,126) = 75.685, $p < 0.01$), its interaction with the effect of the treatment, as described by the triple

7. Other additional analyses were performed but are not included in this section for the sake of brevity. They are available in the lab–pack and their results are commented in the supplemental material for the interested reader.

Table 7
Descriptive Statistics of Conceptual Modeling Effectiveness and Efficiency per Experimental Condition in Pooled Data.
PRE stands for *pre–treatment*, POST for *post–treatment*, PS for *public speaking*, MF for *mindfulness*,
EO for Erasmus->EoDP, and OE for EoDP→Erasmus.

| Experimental Conditions		n	Conceptual Modeling Effectiveness						Conceptual Modeling Efficiency						row
			mean	sd	median	min	max	range	mean	sd	median	min	max	range	
Grouped by	EoDP	130	**0.54**	0.14	0.54	0.12	0.81	0.69	**0.36**	0.13	0.35	0.11	0.89	0.79	1
Exercise	Erasmus	130	**0.41**	0.11	0.41	0.00	0.69	0.69	**0.26**	0.12	0.24	0.00	0.62	0.62	2
Grouped by	PRE	130	0.44	0.13	0.44	0.00	0.76	0.76	0.25	0.09	0.24	0.00	0.55	0.55	3
Time	POST	130	0.51	0.14	0.51	0.12	0.81	0.69	0.37	0.14	0.37	0.11	0.89	0.79	4
Grouped by	EoDP–PS	64	0.52	0.13	0.52	0.24	0.76	0.52	0.31	0.09	0.29	0.13	0.55	0.42	5
Exercise &	Erasmus–PS	64	0.41	0.10	0.41	0.10	0.69	0.59	0.23	0.07	0.22	0.06	0.55	0.48	6
Treatment	EoDP–MF	66	0.55	0.15	0.56	0.12	0.81	0.69	0.40	0.15	0.40	0.11	0.89	0.79	7
	Erasmus–MF	66	0.41	0.13	0.42	0.00	0.69	0.69	0.30	0.14	0.28	0.00	0.62	0.62	8
Grouped by	EO–PRE	85	**0.40**	0.12	0.41	0.00	0.69	0.69	**0.22**	0.08	0.20	0.00	0.55	0.55	9
Order &	EO–POST	85	**0.55**	0.14	0.56	0.12	0.81	0.69	**0.38**	0.14	0.38	0.11	0.89	0.79	10
Time	OE–PRE	45	**0.51**	0.12	0.52	0.24	0.76	0.52	**0.31**	0.08	0.28	0.19	0.55	0.37	11
	OE–POST	45	**0.43**	0.10	0.43	0.27	0.69	0.41	**0.35**	0.13	0.32	0.16	0.62	0.46	12
Grouped by	PRE–PS	64	0.44	0.12	0.44	0.10	0.76	0.66	0.24	0.08	0.24	0.06	0.55	0.49	13
Time &	PRE–MF	66	0.43	0.15	0.45	0.00	0.76	0.76	0.25	0.11	0.25	0.00	0.55	0.55	14
Treatment	POST–PS	64	**0.49**	0.14	0.48	0.26	0.75	0.49	**0.30**	0.09	0.29	0.13	0.55	0.41	15
	POST–MF	66	**0.53**	0.14	0.54	0.12	0.81	0.69	**0.44**	0.13	0.43	0.11	0.89	0.79	16
Grouped by	Mind#1–EoDP	32	0.64	0.12	0.65	0.30	0.81	0.51	0.44	0.15	0.41	0.25	0.89	0.64	17
Experiment	Mind#2–EoDP	53	0.50	0.13	0.52	0.12	0.73	0.61	0.34	0.12	0.35	0.11	0.61	0.50	18
& Exercise	Mind#3–EoDP	45	0.51	0.12	0.52	0.24	0.76	0.52	0.31	0.08	0.28	0.19	0.55	0.37	19
	Mind#1–Erasmus	32	0.41	0.14	0.41	0.00	0.69	0.69	0.18	0.06	0.18	0.00	0.32	0.32	20
	Mind#2–Erasmus	53	0.39	0.11	0.41	0.10	0.59	0.49	0.24	0.09	0.24	0.06	0.55	0.49	21
	Mind#3-Erasmus	45	0.43	0.10	0.43	0.27	0.69	0.41	0.35	0.13	0.32	0.16	0.62	0.46	22
Grouped by	Mind#1–PS	32	0.51	0.13	**0.50**	0.27	0.75	0.48	0.28	0.11	**0.26**	0.12	0.49	0.37	23
Experiment	Mind#2–PS	52	0.44	0.12	**0.44**	0.10	0.73	0.63	0.26	0.08	**0.26**	0.06	0.42	0.36	24
& Treatment	Mind#3–PS	44	0.46	0.12	**0.45**	0.24	0.76	0.52	0.28	0.08	**0.26**	0.16	0.55	0.39	25
	Mind#1–MF	32	0.53	0.21	**0.58**	0.00	0.81	0.81	0.34	0.21	**0.29**	0.00	0.89	0.89	26
	Mind#2–MF	54	0.45	0.14	**0.49**	0.11	0.70	0.59	0.33	0.14	**0.34**	0.07	0.61	0.54	27
	Mind#3–MF	46	0.48	0.11	**0.48**	0.24	0.76	0.52	0.38	0.11	**0.36**	0.19	0.62	0.43	28
column		1	2	3	4	5	6	7	8	9	10	11	12	13	
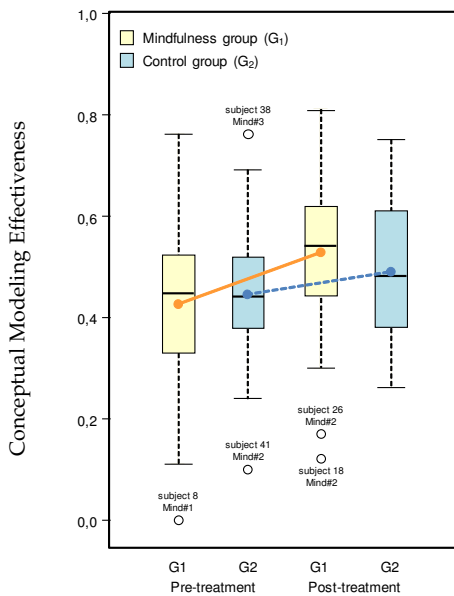
Figure 7. Box Plot of Conceptual Modeling Effectiveness in Pooled Data
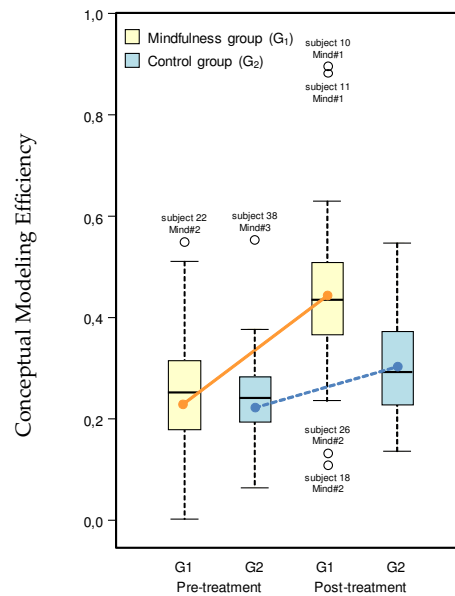

Figure 8. Box Plot of Conceptual Modeling Efficiency in Pooled Data

Table 8
2 (*group*)×2 (*time*) Mixed–model ANOVA of
Conceptual Modeling Effectiveness in Pooled Data

Source of variation	Sum of Squares	DF	Mean Square	F	Sig.	η_p^2
time	0.353	1	0.353	21.514	0.000	0.144
group*time	0.071	1	0.071	4.341	0.039	0.033
Error	2.098	128	0.016			

Table 11
2 (*group*)×2 (*time*) Mixed–model ANOVA of
Conceptual Modeling Efficiency in Pooled Data

Source of variation	Sum of Squares	DF	Mean Square	F	Sig.	η_p^2
time	0.946	1	0.946	104.862	0.000	0.450
group*time	0.272	1	0.272	30.125	0.000	0.191
Error	1.154	128	0.009			

Table 9
2 (*group*)×2 (*time*)×2 (*ord*) Mixed–model ANOVA of
Conceptual Modeling Effectiveness in Pooled Data

Source of variation	Sum of Squares	DF	Mean Square	F	Sig.	η_p^2
time	0.086	1	0.086	8.240	0.005	0.061
group*time	0.059	1	0.059	5.648	0.019	0.043
ord*time	0.786	1	0.786	75.685	0.000	0.375
ord*group*time	0.003	1	0.003	0.256	0.614	0.002
Error	1.308	126	0.010			

Table 12
2 (*group*)×2 (*time*)×2 (*ord*) Mixed–model ANOVA of
Conceptual Modeling Efficiency in Pooled Data (with *ord*)

Source of variation	Sum of Squares	DF	Mean Square	F	Sig.	η_p^2
time	0.615	1	0.615	82.263	0.000	0.395
group*time	0.269	1	0.269	36.008	0.000	0.222
ord*time	0.208	1	0.208	27.857	0.000	0.181
ord*group*time	0.005	1	0.005	0.617	0.434	0.005
Error	0.943	126	0.007			

Table 10
2 (*group*)×2 (*time*) Mixed–model ANCOVA of
Conceptual Modeling Effectiveness in Pooled Data
with *ord* as a covariate

Source of variation	Sum of Squares	DF	Mean Square	F	Sig.	η_p^2
time	0.085	1	0.085	8.259	0.005	0.061
ord*time	0.788	1	0.788	76.322	0.000	0.375
group*time	0.074	1	0.074	7.126	0.009	0.053
Error	1.310	127	0.010			

Table 13
2 (*group*)×2 (*time*) Mixed–model ANCOVA of
Conceptual Modeling Efficiency in Pooled Data
with *ord* as a covariate

Source of variation	Sum of Squares	DF	Mean Square	F	Sig.	η_p^2
time	0.617	1	0.617	82.670	0.000	0.394
ord*time	0.207	1	0.207	27.797	0.000	0.180
group*time	0.274	1	0.274	36.748	0.000	0.224
Error	0.947	127	0.010			

ord*group*time interaction (F(1,126) = 0.256, $p = 0.614$), is not significant. These results led us to think that the observed improvement in conceptual modeling effectiveness might be independent of any combination of the order in which the exercises were performed. To discard the effect of such combination, a Bayesian analysis was carried out comparing models with and without the triple interaction, obtaining a value of $3.37 \pm 3.75\%$ for the Bayes factor B_{10}, which constitutes a *substantial*[8] evidence for supporting the model without the triple interaction over the others, i.e. the observed effects of mindfulness are likely to be independent of the order in which the conceptual modeling exercises were performed.

Finally, in order to remove the effect of no–interest of the order, a 2×2 mixed–model ANCOVA with *ord* as a covariate [59] was conducted and its results are shown in Table 10. Note that when the effect of the order is controlled, not only the results are consistent with the previous ones, but also that the effect of the treatment is bigger than in Tables 8 and 9, although *small* according to [55]. Note also that the

8. According to [58], values of B_{10} between 1 and 3.2 constitute *anecdotal* evidence, between 3.2 and 10 *substantial* evidence, between 10 and 100 *strong* evidence, and values above 100 constitute *decisive* evidence.

effect of mindfulness is not only significant as in Tables 8 and 9, but *highly* significant, as described by the *group*time* interaction (F(1,127) = 7.126, $p < 0.01$).

With respect to DQ_4, the previous results confirm our initial interpretation of the difference of results of MIND#3 with respect to MIND#1–2, i.e. that the ERASMUS exercise was more difficult to solve for the subjects than the EODP exercise, as commented in Section 4.5. Nevertheless, the same results confirm that, despite the observed difference in the solving difficulty of the exercises, the effect of mindfulness on conceptual modeling effectiveness is small but statistically significant.

Tests for Conceptual Modeling Efficiency. Regarding efficiency, all the obtained results of the statistical tests are consistent with the previous results of the individual experiments, confirming a *large* and highly significant effect of mindfulness, as described by the *group*time* interactions in Tables 11, 12, and 13, where $p < 0.01$ and $\eta_p^2 > 0.014$ in all cases.

Concerning the effect of the order, the results were consistent with those obtained for effectiveness, i.e. it has a highly significant effect but its interaction with the effect of the treatment is not significant, as described by the *ord*time* (F(1,126) = 27.857, $p < 0.01$) and the *ord*group*time* interac-

Table 14
Forest Plot of the Change from Pre–Treatment to Post–Treatment in Conceptual Modeling Effectiveness in Aggregated Data

	Experimental group			Control group			Aggregated data			
Study	Sample	Mean	SD	Sample	Mean	SD	SMD and 95% CI	SMD	95% CI	Weight
MIND#1	16	0.278	0.139	16	0.195	0.133		0.612	[-0.097, 1.320]	24.3%
MIND#2	27	0.141	0.170	26	0.067	0.142		0.475	[-0.071, 1.021]	41.2%
MIND#3	23	-0.053	0.113	22	-0.102	0.092		0.482	[-0.111, 1.075]	34.5%
Fixed effect model	66			64				0.511	[0.161, 0.860]	100.0%

Heterogeneity: $I^2 = 0\%$, $\tau^2 = 0$, $p = 0.950$

Test for overall effect: $Z = 2.86$, $p = 0.004$

Table 15
Forest Plot of the Change from Pre–Treatment to Post–Treatment in Conceptual Modeling Efficiency in Aggregated Data

	Experimental group			Control group			Aggregated data			
Study	Sample	Mean	SD	Sample	Mean	SD	SMD and 95% CI	SMD	95% CI	Weight
MIND#1	16	0.331	0.126	16	0.192	0.064		1.391	[0.619, 2.163]	25.4%
MIND#2	27	0.155	0.161	26	0.048	0.094		0.810	[0.250, 1.370]	48.3%
MIND#3	23	0.119	0.071	22	-0.034	0.059		2.346	[1.587, 3.106]	26.3%
Fixed effect model	66			64				1.361	[0.972, 1.751]	100.0%

Heterogeneity: $I^2 = 80\%$, $\tau^2 = 0.511$, $p = 0.006$

Test for overall effect: $Z = 6.85$, $p = 0.000$

tions (F(1,126) = 0.617, $p = 0.434$) in Table 12. As for effectiveness, a Bayesian analysis was also conducted and its results supported the model without the triple interaction with *substantial* evidence ($B_{10} = 3.05 \pm 3.82\%$). The subsequent ANCOVA analysis with *ord* as covariate confirmed the previous results with a *large* highly significant effect of mindfulness (F(1,127) = 36.748, $p < 0.01$).

5.2 Aggregated Data Meta–Analysis

The results of the aggregated data meta–analysis for the change from pre–treatment to post–treatment are shown in Tables 14 and 15 as *forest plots* [60] for both response variables. Each line shows the sample size, mean, and standard deviation for the experimental and control groups of each study, together with a plot diagram in which each line depicts the 95%–confidence interval of the effect size estimate, measured as the *standardized mean difference* (SMD), also known as Cohen's d [61]. The weight of each study is represented as a proportional square centered on the mean effect size and the numerical counterparts of the plot diagram are shown in the last three columns. In the fourth line, the estimate of the overall effect size is represented by a dashed vertical line and its confidence interval as a diamond, considering 0 in the horizontal axis as the line of no effect.

Regarding effectiveness, the results of each individual experiment for the effect of mindfulness are non–negligible but not statistically significant, since the intervals contain the line of no effect. The confidence intervals for each experiment are therefore consistent with the results in Section 4. Nevertheless, when considering the overall results, the

positive effect of mindfulness on effectiveness is *medium*[9] and statistically highly significant because it does not contain the line of no effect and $p < 0.01$. Specifically, the overall effect (row 4) is consistent with the results in the pooled data meta–analysis, where the results become significant when considering the increased sample size.

In the case of efficiency, the outcomes of the aggregated data meta–analysis are also consistent with the previous results, showing also a highly significant ($p < 0.01$), *very large* positive effect of mindfulness.

6 THREATS TO VALIDITY

Wohlin et al. [49] provide a thorough compilation of threats to the validity of empirical studies in Software Engineering. In this section, such threats are analyzed and the actions performed to mitigate them are described not only for each individual experiment but also for the whole series of experiments where appropriate.

6.1 Conclusion Validity

The conclusion validity is concerned with the statistical relationship between the treatment and the outcome. In MIND#1, the main threat to conclusion validity was the small size of the sample, which was also one of the main reasons for the subsequent replications. Nevertheless, although small, the sample size was acceptable for the statistical tests applied, as described in [62], and all the assumptions of each test were verified before their application. In both

---

9. The thresholds when the effect size is estimated using Cohen's d are 0.20 for *small*, 0.50 for *medium*, 0.80 for *large* and 1.20 for *very large*.

replications, the sample size was significantly larger than in MIND#1, thus addressing this threat.

The application of pooled data meta–analysis has contributed also to neutralize this threat since the data of the series of experiments were treated as a single, larger sample. Moreover, the assumptions of the applied statistical tests were always verified in both pooled and aggregated meta–analyses.

6.2 Internal Validity

Internal validity is concerned with the treatment—and no other uncontrolled variable—being the cause of the outcome. Concerning the experiments in the series, the identified threats according to [49] are commented below.

History. Since both groups performed the experimental tasks simultaneously, without any significant incident, this threat was neutralized for all the experiments in the series.

Maturation. In each experiment, both groups matured simultaneously with respect to their knowledge in conceptual modeling, since they all attended the same number of ISEIS lessons with the same professor and the same content between the pre– and post—treatment exercises. Therefore, this threat was also neutralized for all the experiments in the series.

Dropout. In order to reduce this threat, the students were offered a half–a–point bonus for participating in every experiment, as commented in Section 3.3.2. Considering the difference between the students who showed interest in participating in the experiments and those in the final samples plus the 6 repeater students who were discarded (see Section 3.2), the observed dropout percentages were 56.00% in MIND#1, 33.33% in MIND#2 and 33.80% in MIND#3.

The large number of participants who dropped out of the study might suggest that the bonus offered for participating was not motivating enough for our students. Another possible cause of the large dropout rates might be the work overload experienced by our students, which makes some of them abandon the ISEIS subject itself until the September call or even until the next academic year. These overloaded students would have participated at the beginning of the experiment but certainly they would have not finished the experimental tasks, thus increasing the dropout rate. With respect to the different moments of drop–out among the three experiments, we think it is related to the specific assignment deadlines of concurrent subjects, which varies among academic years.

Selection & assignment bias. In both replications, the selection and assignment of subjects to groups were random, which usually neutralizes this threat. In MIND#1, the subjects were assigned to groups according to their preferences, so the authors performed a crosscheck by computing a one–way ANOVA test on the measures of the pre–treatment exercise response variables (see Figure 9). The results of the test revealed that there was no evidence of significant differences between groups prior to treatment. Additionally, giving participants a bonus for their participation could lead to self–selection bias. However, since the bonus was
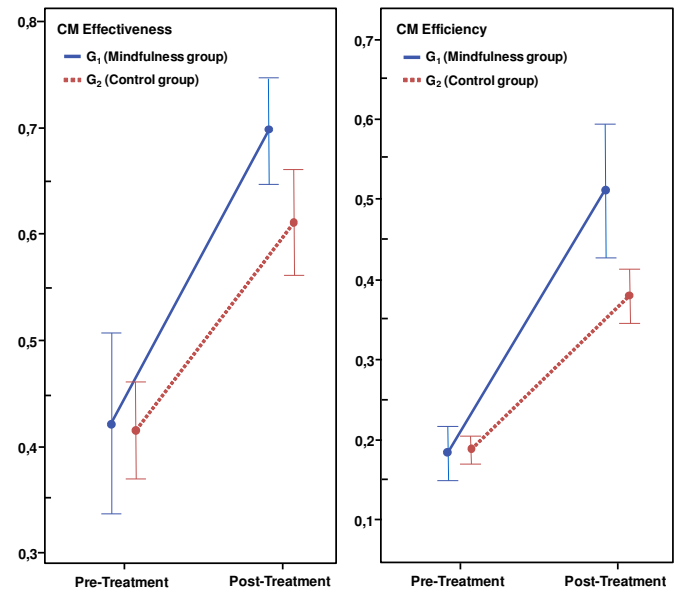

Figure 9. Profile Plots and 95% Confidence Intervals of Response Variables in MIND#1 (from [17])

given to all participants in both treatment and control groups, such possible bias affected them equally and did not constitute a threat to the internal validity of the study.

Instrumentation. In order to avoid interaction of both treatments, a potential placebo side–effect on response variables was neutralized in both replications, i.e. the public speaking workshop took place after performing the post–treatment exercise. The outcomes of the baseline experiment, when compared to those of the replications, suggest that the public speaking workshop was actually a placebo in MIND#1, as questioned in DQ$_2$ (see Section 4). Another instrumentation threat to the internal validity was the scoring of the conceptual modeling exercises being incorrect or obtained applying different criteria to different subjects. To neutralize this threat, all the exercises were blindly scored by the same experimenter.

With respect to the whole series of experiments, the main threat to internal validity was including the dataset of MIND#1 in the pooled and aggregated data meta–analyses, since it was a quasi–experiment because of the non–random assignment of subjects. However, we decided to include it for three reasons. First, the results of the pre–treatment exercise showed that both groups started from very similar situations on average, as can be seen in the profile plots of MIND#1 in Figure 9. Second, this similarity was statistically verified, as commented above. Third, the potentially higher motivation of the experimental group was discarded after checking that the dropout was also similar in both groups, as shown in Table 2.

Another relevant threat to the internal validity was the unexpected difference in the solving difficulty of the conceptual modeling exercises observed when their order was swapped in MIND#3. Considering the results of the pooled data meta–analysis in Section 5.1, the difference between the exercises is confirmed, but it is also that there is no

evidence that the effect of mindfulness depends on the specific exercise to be solved by the subjects.

Finally, as commented in Section 3.7.1, all the subjects had the same professors and the same content was taught to all of them at the same pace in order to avoid any difference in their conceptual modeling training.

6.3 Construct Validity

This validity is concerned with the relation between theory and observation. Since all the experiments included two different treatments and exercises, the *mono–operation* bias was reduced because of the cause construct being completely represented. The *mono–method* bias was also reduced by considering two different response variables[10].

6.4 External Validity

Four threats which could limit the generalization of the outcomes were identified. On one hand, the size of the interview transcripts might not be representative of real industrial problems, but they were appropriate for the available time for the pre– and post–treatment exercises. However, we think that the intellectual processes applied during conceptual modeling—potentially improved by the practice of mindfulness—are basically the same regardless of the size of the problem at hand.

On the other hand, since the experimental tasks did not require high levels of industrial experience, using Software Engineering students as subjects instead of professionals could be considered as appropriate [63]. Moreover, students are the next generation of professionals, so they are close to the population under study [64], [65].

The third threat to external validity was the fact that all the subjects were Software Engineering students at the University of Seville and had received similar training. Therefore, if the study was replicated at another institution, the results may have been different due to different training. Only an external replication can mitigate this threat.

Finally, the external validity of the study could be threatened by the bias generated by giving students a bonus, making the results applicable only to people with certain interests. However, if students are not interested in obtaining a substantial bonus on their grades easily, they would hardly be interested in participating in the experiment in a totally altruistic manner. Given that the use of an economic incentive was unfeasible, and would have constituted another form of self–selection anyway, we consider this threat to external validity intractable given the circumstances.

7 RELATED WORK

To the best of the authors' knowledge, apart from this article and our previous works [17], [18], the only empirical study relating mindfulness and performance on a

Software Engineering task is a controlled experiment on the perceived effectiveness of *stand–up* meetings held after 3–minute mindfulness sessions [46]. This study was carried out in three Dutch companies that had been using the Scrum agile methodology for at least three years. The subjects were assigned to experimental, placebo[11] and null–treatment groups. The data obtained using questionnaires after the stand–up meetings revealed that there was a positive impact on the perceived effectiveness, decision–making and listening in the experimental group with respect to the other groups.

Although we have found only one directly related work, a selection of the 65 studies recently compiled in a meta–analysis [25] about the potential benefits of mindfulness in some cognitive abilities such as attention, learning, memory or intelligence are summarized in Table 16, commented below and compared with our series of experiments. Notice that all the studies that were carried out in academic environments (marked with an asterisk in their reference in Table 16), took place at Social or Health Science Schools like Education, Experimental Psychology or Medicine. None of them took place at a School of Technology like ours.

Regarding the experimental design, most of the studies in Table 16 used a *pre–post* design, as it was our case. The *simple* design is usually adopted when subjects are experienced meditators because there is no possibility of measuring anything before the beginning of the treatment.

With respect to the response variables in the studies in Table 16, they are usually measured using standardized questionnaires, although the high variability makes difficult to perform meta–analyses because of the incomparability of measures, a serious issue to promote more rigorous meditation research, as highlighted in [25], [72].

In most of the studies in Table 16, the subject assignment to groups was random except when organizational constraints made it impossible or subjects were long–term meditators. For instance, [74] reported an experiment carried out with 20 experienced meditators and 20 matched controls without any meditation experience.

Regarding the alternative treatment administered to the control group (last column in Table 16), the meta–analysis in [25] classified studies as those in which a mindfulness group is compared with a *conventional control* (CC) group, for example a waiting list, and those in which the mindfulness group is compared with an *active control* (AC) group, i.e. attentional, relaxing or resting treatment. Sometimes, the alternative to the mindfulness group is an *inactive control* (IC) group, in which subjects are not treated. The meta–analysis in [25] also revealed that more differences between groups were detected when using CC/IC rather than AC. This finding seems reasonable because of the potential effect of the alternative treatment in AC groups. In disciplines like Psychology, it usual to compare mindfulness with other attentional techniques in order to know which one is better and in which cases, which is not our case.

Concerning the way of practicing mindfulness during the treatment, some of the experimenters opted for once–a–week longer sessions [75] or for off–site sessions supported

---

10. According to [49], *mono–operation* bias occurs if the experiment includes a single independent variable, case, subject or treatment, because the experiment may under–represent the construct and thus not give the full picture of the theory. *Mono–method* bias occurs when using a single type of measures or observations because this involves a risk that if this measure or observation gives a measurement bias, then the experiment will be misleading. By involving different types of measures and observations they can be cross–checked against each other.

11. The placebo groups listened to *Tango* by Igor Stravinsky instead of practicing mindfulness.

Table 16
Empirical Studies about the Effects of Mindfulness Training in Cognitive Abilities (based on [25]).
References with an asterisk correspond to studies carried out in academic environments.
CC stands for *conventional control*, AC for *alternative control* and IC for *inactive control*.

Ref.	No. of studies	Exp. Design	Data Gathering	Response variables	Random Assignment	Control condition
[66]*	4	2 *simple* 2 *pre–post*	Mindfulness & personality scales, creativity task	Creativity	Yes	CC
[14]	1	*simple*	*Dundee Stress State Questionnaire* (DSSQ)	Sustained attention	Yes?	CC
[67]	1	*pre–post*	Perceptual and memory tasks	Metacognitive ability	Yes	CC
[15]*	1	*pre–post*	*Graduate Records Examination* (GRE)	Working memory, reading comprehension, mind–wandering	Yes	CC
[68]	1	*pre–post*	*ad–hoc* association, originality, fluency & flexibility tasks	Creativity	No	AC
[69], [70]*	2	*pre–post*	*Torrance Tests of Creative Thinking* (TTCT)	Creativity	Yes	AC
[71]	2	*pre–post*	*Einstellung* water jar task	Cognitive rigidity	No & Yes	CC
[72]*	1	*pre–post*	*Dual attention to response task* (DART), *Spatial and temporal attention network* (STAN) & Stroop color-–word task	Working memory, attention and cognitive flexibility	Yes	IC & AC
[73]	1	*simple*	*Continuous Performance Task* (CPT), Ruff 2 & 7 Selective Attention Test & *California Verbal Learning Test* (CVLT)	Memory and attention	No	IC
[74]	1	*simple*	*Subjective Time Questionnaire* (STQ) & *Attention Network Test* (ANT)	Sense of time, attention	No	IC
This work*	3	*pre–post*	*Ad–hoc* measures	Conceptual modeling quality & productivity	1 No & 2 Yes	1 AC & 2 CC

by an audio guide [32], [76]. The findings in [25] have also revealed that the length of mindfulness training is correlated positively with the strength of its effects, and that the training should last at least 4 weeks to be effective. In our case, considering the limitations described in Section 4.2, face–to–face short (10–12 minutes) daily sessions seemed the most appropriate for facilitating student attendance, and 4–6 weeks appeared to be long enough for the training to be effective, as confirmed by our results.

8 LESSONS LEARNED

In this section, some of the lessons learned during the 3–year iterative process of running the series of controlled experiments are commented.

8.1 Subject Assignment and Control Condition

One of the lessons learned is that since subjects' motivation does not seem to be a relevant factor, as concluded from the outcomes of the series of experiments described in Section 4, random assignment to groups should be used whenever possible because of its well–known advantages for statistical analysis [49], [52], [77].

Another lesson is that the subjects in the control group should feel they are treated in the same way than the subjects in the experimental group, in order to avoid any *reactivity* effect [77]. For the same reason, it is important to have control conditions as motivating as the experimental conditions but without any effect on the outcomes. If such control conditions were not available, an alternative would be using a waiting list and administering the same or similar

treatment to the control group, but once all the measures had been taken, i.e. having a formal *null* treatment condition although all subjects receive equally motivating treatments, as we did in the series of experiments.

8.2 Experimental Design

As shown in Table 16, most mindfulness studies use *simple* or *pre–post* experimental designs. Nevertheless, there is a relevant difference between the *pre–post* studies in Table 16 and our series of experiments. While those studies use the same questionnaire in the pre–post measures, we cannot use the same conceptual modeling exercise because of the *testing* effect [49], i.e. the subjects might become familiar with the exercise and avoid previous mistakes in the second measure. As a general rule, we can say that those Software Engineering tasks whose complexity can be measured in a concise and objective way, are the best candidates for a controlled experiment using mindfulness. If such a task is found, search for two exercises with similar complexity and use a *pre–post* design. If this is not possible, consider randomizing the tasks among participants in the same group to deal with potential differences in solving difficulty, or use a *simple* design and measure response variables only once after the administration of the treatment.

One of the advantages of the *pre–post* design is that it allows the analysis of the evolution of the task performance over time. One of its drawbacks is that if the task itself has a significant effect on the outcomes and it has also a significant interaction with the treatment, the results might be inconclusive. On the other hand, in both *pre–post* and *simple* experimental designs, the variables that introduce nuisance

should be identified and controlled by means of *blocking* [51] when possible. For example, a different level of subjects' experience with the task under study. In this situation, we recommend using a questionnaire before the beginning of the conduction of the experiment, which allows creating the blocks to overcome the effects of the nuisances.

8.3 Software Engineering Task Choice

Apart from the complexity–related aspects commented in the previous section, a good Software Engineering task for a mindfulness experiment is a task for which attention, mental clarity, and problem–solving capabilities were helpful.

Another relevant aspect is the workload required for scoring the deliverables generated by the task, which could be quite time–consuming. A clear recommendation about scoring is having a reference solution to compare with and clear criteria for assigning numerical values. When possible, the scoring should be performed blindly by the same person, in order to avoid different applications of the scoring criteria. In the case of conceptual modeling, it is also recommended to have a list of synonyms for each of the concepts in the problem domain, as we had in our case.

Ideally, automated score mechanisms such as multiple option tests would alleviate the scoring workload, but this is not always possible to achieve, especially for creative tasks such as developing conceptual models.

8.4 Mindfulness for Software Engineering Students

There are different approaches for integrating mindfulness into higher education, as commented in Section 2.2. In our case, considering the results of the series of experiments and the positive feedback from the participating students, we think that workshops consisting of short sessions as those described in Section 3.3.4, during a period of at least 4 weeks, are not only a good entry level for Software Engineering students in the practice of mindfulness, but also a way for measuring its effects in different Software Engineering education aspects.

During the workshops, additional resources such as the *Insight Timer*[12] or *Headspace*[13] applications can also be recommended to the students in order to continue the practice once the workshops are over, helping them to achieve a mindful state into their daily life.

Apart from the workshops, the first author has also obtained positive feedback from practicing mindfulness some minutes during her lessons, which can be seen not only as a complementary activity to the workshops themselves, but also as a dissemination action of mindfulness among students.

9 CONCLUSIONS AND FUTURE WORK

In this article, we have presented a 3–year series of experiments to evaluate the effect of mindfulness on the conceptual modeling performance of Software Engineering students (see Table 17 for a summary of results). The series is composed of three experiments, the baseline experiment

---

12. https://insighttimer.com/
13. https://www.headspace.com/

---

and two internal replications carried out at the University of Seville during the 2013–2016 academic years. During the planning of the replications, some aspects of the baseline experiment were modified taking into account not only the experimental results but also the lessons learned from each experiment. The understanding of this evolution and the resulting foundations may be useful for researchers in the area willing to carry out further replications [21].

In addition to the description of the followed process and a narrative synthesis of the results, a joint analysis consisting of a pooled and an aggregated data meta–analyses of the whole series of experiments is also presented in this article. All analyses have yielded consistent results, i.e. students who practice mindfulness get better results on the task (effectiveness) and they are more productive, i.e. they arrive at such solutions more quickly (efficiency). The fact that the practice of mindfulness significantly impacts not only conceptual modeling productivity but also quality is a new finding of the joint analysis. This finding has been obtained thanks to the increased sample size when processing the aggregated data, which has allowed us to uncover subtle effects that could not be observed in the isolated experiments.

As future work, we want to study whether the benefits of mindfulness are also applicable to other tasks in Software Engineering in which being focused and having a clear mind is crucial, e.g. coding, debugging, formal technical reviews, etc. Another interesting path to explore is cooperating with psychologists to determine which are the most relevant mental abilities for some Software Engineering tasks and how those abilities, such as abstraction [78], might be improved by the practice of mindfulness.

On the other hand, some case studies in real companies are underway in order to confirm our results with professionals. In the long term, the ultimate aim of our empirical studies is to define the best way to introduce the practice of mindfulness both in academia and industry and develop a mindfulness–based productivity and stress–reduction program as a means to achieve that goal.

LABORATORY PACKAGE

The laboratory package is available at the EXEMPLAR platform [79] in https://exemplar.us.es/demo/BernardezMindfulnessTSE. Apart from the experimental material described in Section 3.4, the lab–pack also includes i) descriptions of all the experiments in SEDL (*Scientific Experiments Description Language*), the domain specific language for experiment description of the EXEMPLAR platform [80, Ch. 6]; ii) raw data files in CSV format; and iii) statistical analysis and graph–generating scripts in R.

Table 17
Summary of Outcomes of the Series of Experiments including Meta–analyses.
The data for MIND#1–3 and the pooled data meta–analysis correspond to 2 (*group*) × 2 (*time*) mixed–model ANOVAs.

Outcome	MIND#1	MIND#2	MIND#3	Pooled data Meta–analysis	Aggregated Meta–analysis
CM Effectiveness Improvement	Not significant ($F(1,30) = 2.713$, $p = 0.110$)	Not significant ($F(1,51) = 2.908$, $p = 0.094$)	Not significant ($F(1,43) = 2.614$, $p = 0.113$)	Significant ($F(1,128) = 4.341$, $p = 0.039$)	Highly significant ($p = 0.004$)
CM Effectiveness Effect Size	Medium ($\eta_p^2 = 0.086$)	Small ($\eta_p^2 = 0.054$)	Small ($\eta_p^2 = 0.057$)	Small ($\eta_p^2 = 0.033$)	Medium ($d = 0.511$)
CM Efficiency Improvement	Highly significant ($F(1,30) = 17.001$, $p = 0.000$)	Highly significant ($F(1,51) = 8.698$, $p = 0.005$)	Highly significant ($F(1,43) = 61.602$, $p = 0.000$)	Highly significant ($F(1,128) = 30.125$, $p = 0.000$)	Highly significant ($p = 0.000$)
CM Efficiency Effect Size	Large ($\eta_p^2 = 0.370$)	Large ($\eta_p^2 = 0.146$)	Large ($\eta_p^2 = 0.590$)	Large ($\eta_p^2 = 0.191$)	Very large ($d = 1.361$)
Sample size	32	53	45	130	130

REFERENCES

[1] T. DeMarco and T. Lister, *Peopleware: productive projects and teams*. Addison-Wesley, 2013.

[2] S. Hastie and S. Wojewoda. (2015) Standish Group 2015 Chaos Report – Q&A with Jennifer Lynch. [Online]. Available: https://bit.ly/1Y6nWAv

[3] D. Graziotin, X. Wang, and P. Abrahamsson, "Software developers, moods, emotions, and performance," *IEEE SOFTWARE 1405.4422*, 2014.

[4] M. Csikszentmihalyi, *Flow: The Psychology of Optimal Experience*. Harper Perennial, 1991.

[5] A. N. Meyer, L. E. Barton, G. C. Murphy, T. Zimmermann, and T. Fritz, "The work life of developers: Activities, switches and perceived productivity," *IEEE Transactions on Software Engineering*, vol. 43, no. 12, pp. 1178–1193, 2017.

[6] J. Kabat-Zinn, *Full Catastrophe Living: Using the Wisdom of Your Body and Mind to Face Stress, Pain, and Illness*. Delta, 1990.

[7] ——, *Wherever You Go, There You Are: Mindfulness Meditation in Everyday Life*. Hyperion, 1994.

[8] D. M. Davis and J. A. Hayes, "What are the benefits of mindfulness? a practice review of psychotherapy–related research." *Psychotherapy*, vol. 48, no. 2, p. 198, 2011.

[9] S. Lazar. (2011) How meditation can reshape our brains: Sara Lazar at TEDxCambridge 2011. [Online]. Available: https://youtu.be/m8rRzTtP7Tc

[10] K. Schaufenbuel, "Bringing mindfulness to the workplace," The Kenan–Flagler Business School, Executive Development, University of North Carolina, Tech. Rep., 2014. [Online]. Available: https://unc.live/2W5W0QR

[11] H. Agnew. (2014) 'Mindfulness' gives stressed–out bankers something to think about. [Online]. Available: https://on.ft.com/2RAmVFT

[12] J. T. Ramsburg and R. J. Youmans, "Meditation in the higher–education classroom: Meditation training improves student knowledge retention during lectures," *Mindfulness*, vol. 5, no. 4, pp. 431–441, 2014.

[13] S. L. Shapiro, K. W. Brown, and J. A. Astin, "Toward the integration of meditation into higher education: A review of research," *Teachers College Record*, vol. 113, no. 3, pp. 493–528, 2011.

[14] E. Cardeña, J. O. Sjöstedt, and D. Marcusson-Clavertz, "Sustained attention and motivation in zen meditators and non-meditators," *Mindfulness*, vol. 6, no. 5, pp. 1082–1087, 2015.

[15] M. D. Mrazek, M. S. Franklin, D. T. Phillips, B. Baird, and J. W. Schooler, "Mindfulness training improves working memory capacity and GRE performance while reducing mind wandering," *Psychological Science*, vol. 24, no. 5, pp. 776–781, May 2013.

[16] D. Vago and S. David, "Self–awareness, self–regulation, and self–transcendence (S–ART): a framework for understanding the neurobiological mechanisms of mindfulness," *Frontiers in Human Neuroscience*, vol. 6, p. 296, 2012.

[17] B. Bernárdez, A. Durán, J. A. Parejo, and A. Ruiz-Cortés, "A controlled experiment to evaluate the effects of mindfulness in software engineering," in *Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 2014, pp. 17–27.

[18] ——, "An experimental replication on the effect of the practice of mindfulness in conceptual modeling performance," *Journal of Systems and Software*, vol. 136, pp. 153–172, 2018.

[19] V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård, and M. V. Zelkowitz, "The empirical investigation of perspective–based reading," *Empirical Software Engineering*, vol. 1, no. 2, pp. 133–164, 1996.

[20] F. J. Shull, J. C. Carver, S. Vegas, and N. Juristo, "The role of replications in empirical software engineering," *Empirical Software Engineering*, vol. 13, no. 2, pp. 211–218, 2008.

[21] M. Cruz, B. Bernárdez, A. Durán, J. A. Galindo, and A. Ruiz–Cortés, "Replication of studies in empirical software engineering: A systematic mapping study, from 2013 to 2018," *IEEE Access*, vol. 8, pp. 26 773–26 791, 2020.

[22] A. Santos, O. S. Gomez, and N. Juristo, "Analyzing families of experiments in SE: a systematic mapping study," *IEEE Transactions on Software Engineering (Preprints)*, 2018.

[23] A. Puddicombe, *Get some headspace*. Hodder and Stoughton, 2011.

[24] V. Simón, *Aware and Awake*. Descleé De Brouwer, 2013.

[25] P. Sedlmeier, C. Loße, and L. C. Quasten, "Psychological effects of meditation for healthy practitioners: an update," *Mindfulness*, vol. 9, no. 2, pp. 371–387, Apr 2018.

[26] D. Goleman and R. J. Davidson, *Altered Traits: Science Reveals how Meditation Changes Your Mind, Brain, and Body*. Penguin, 2017.

[27] D. Vago. (2017) Self–transformation through mindfulness: David Vago at TEDxCambridge 2017. [Online]. Available: https://youtu.be/1nP5oedmzkM

[28] J. Kabat-Zinn, "Mindfulness–based stress reduction MBSR," *Constructivism in the Human Sciences*, 2003.

[29] P. Grossman, L. Niemann, S. Schmidt, and H. Walach, "Mindfulness–based stress reduction and health benefits: A meta–analysis," *Journal of Psychosomatic Research*, vol. 57, no. 1, pp. 35–43, 2004.

[30] S. L. Shapiro, J. A. Astin, S. R. Bishop, and M. Cordova, "Mindfulness-based stress reduction for health care professionals: results from a randomized trial." *International Journal of Stress Management*, vol. 12, no. 2, pp. 164–176, 2005.

[31] D. K. Reibel, J. M. Greeson, G. C. Brainard, and S. Rosenzweig, "Mindfulness–based stress reduction and health–related quality of life in a heterogeneous patient population," *General Hospital Psychiatry*, vol. 23, no. 4, pp. 183–192, 2001.

[32] P. A. Poulin, C. S. Mackenzie, G. Soloway, and E. Karayolas, "Mindfulness training as an evidenced-based approach to reducing stress and promoting well-being among human services professionals," *International Journal of Health Promotion and Education*, vol. 46, no. 2, pp. 72–80, 2008.

[33] D. P. Barbezat and M. Bush, *Contemplative practices in higher education: Powerful methods to transform teaching and learning.* John Wiley & Sons, 2013.

[34] Center for Wellness and Health Promotion — Harvard University Health Services. (2019) Mindfulness & Meditation. [Online]. Available: https://bit.ly/2TbUafa

[35] University of California Berkeley School of Law. (2019) Mindfulness at Berkeley Law. [Online]. Available: https://bit.ly/2RIHdx6

[36] The University of Texas at Austin — Counseling and Mental Health Center. (2019) MindBody Labs. [Online]. Available: https://bit.ly/2ba5bgC

[37] University of Cambridge. (2019) Mindfulness at Cam. [Online]. Available: https://bit.ly/2W4kiuy

[38] N. Barnes, P. Hattan, D. S. Black, and Z. Schuman-Olivier, "An examination of mindfulness–based programs in us medical schools," *Mindfulness*, vol. 8, no. 2, pp. 489–494, Apr 2017.

[39] M. B. Schure, J. Christopher, and S. Christopher, "Mind–body medicine and the art of self-care: Teaching mindfulness to counseling students through yoga, meditation, and qigong," *Journal of Counseling and Development*, vol. 86, no. 1, pp. 47–56, Winter 2008.

[40] N. Shachtman, *Enlightenment engineers: meditation and mindfulness in Silicon Valley*, WIRED Magazine, 2013.

[41] Search Inside Yourself Leadership Institute. (2019) Search inside yourself program impact report. [Online]. Available: https://bit.ly/2LUMNdb

[42] C.-M. Tan, *Search Inside Yourself.* Addison–Wesley, 2012.

[43] C. Kelly. (2012) O.k., Google, take a deep breath. [Online]. Available: https://nyti.ms/2QjiY79

[44] S. Matook and K. Kautz, "Mindfulness and agile software development," in *19th Australasian Conference on Information Systems (ACIS) Proceedings*, December 2008, pp. 638–647.

[45] R. Vidgen and X. Wang, "Coevolving systems and the organization of agile software development," *Information Systems Research*, vol. 20, no. 3, pp. 355–376, September 2009.

[46] P. den Heijer, W. Koole, and C. J. Stettina, "Don't forget to breathe: A controlled trial of mindfulness practices in agile project teams," in *International Conference on Agile Software Development.* Springer, 2017, pp. 103–118.

[47] A. Jedlitschka, M. Ciolkowski, and D. Pfahl, *Guide to Advanced Empirical Software Engineering.* Springer–Verlag, 2008, ch. Reporting experiments in Software Engineering, pp. 201–228.

[48] J. C. Carver, "Towards reporting guidelines for experimental replications: a proposal," in *1st International Workshop on Replication in Empirical Software Engineering*, 2010.

[49] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering: an introduction.* Springer Publishing Company, Incorporated, 2012.

[50] M. Genero, A. Fernández-Saez, J. Nelson, G. Poels, and M. Piattini, "A systematic literature review on the quality of UML models," *Journal of Database Management*, vol. 22, no. 3, pp. 46–70, 2012.

[51] N. Juristo and A. M. Moreno, *Basics of Software Engineering Experimentation.* Kluwer Academic Publishers, 2001.

[52] D. T. Campbell and S. Julian, *Experimental and Quasi–Experimental Designs for Research.* United States: Wadsworth, 1963.

[53] E. L. Fancher, "Comparison of methods of analysis for pretest and posttest data," Master's thesis, University of Georgia, 2013.

[54] J. A. Gliner, G. A. Morgan, and R. J. Harmon, "Pretest–posttest comparison group designs: Analysis and interpretation," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 42, pp. 500–503, 2003.

[55] J. T. Richardson, "Eta squared and partial eta squared as measures of effect size in educational research," *Educational Research Review*, vol. 6, no. 2, pp. 135–147, 2011.

[56] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, no. 5, p. 365, 2013.

[57] G. V. Glass, "Primary, secondary, and meta-analysis of research," *Educational researcher*, vol. 5, no. 10, pp. 3–8, 1976.

[58] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the american statistical association*, vol. 90, no. 430, pp. 773–795, 1995.

[59] A. Rutherford, *Introducing ANOVA and ANCOVA: a GLM approach.* Sage, 2001.

[60] S. Lewis and M. Clarke, "Forest plots: trying to see the wood and the trees," *BMJ: British Medical Journal*, vol. 322, no. 7300, p. 1479, 2001.

[61] J. Cohen, *Statistical power analysis for the behavioral sciences.* Academic press, 2013.

[62] M. Jørgensen, T. Dybå, K. Liestøl, and D. I. Sjøberg, "Incorrect results in software engineering experiments: How to improve research practices," *Journal of Systems and Software*, vol. 116, pp. 133–145, 2016.

[63] D. Falessi, N. Juristo, C. Wohlin, B. Turhan, J. Münch, A. Jedlitschka, and M. Oivo, "Empirical software engineering experts on the use of students and professionals in experiments," *Empirical Software Engineering*, vol. 23, no. 1, pp. 452–489, Feb 2018.

[64] B. A. Kitchenham, S. L. Pfleeger, D. Hoaglin., K. E. Emam, and J. Rosenberg, "Preliminary Guidelines for Empirical Research in Software Engineering," *IEEE Transactions on Software Engineering*, vol. 28, no. 8, pp. 721–734, August 2002.

[65] A. A. Porter, L. G. Votta, and V. R. Basili, "Building Knowledge through Families of Experiments," *IEEE Transactions on Software Engineering*, vol. 25, no. 4, pp. 456–473, July 1999.

[66] M. Baas, B. Nevicka, and F. S. T. Velden, "Specific mindfulness skills differentially predict creative performance," *Personality and Social Psychology Bulletin*, vol. 40, no. 9, pp. 1092–1106, 2014.

[67] B. Baird, M. D. Mrazek, D. T. Phillips, and J. W. Schooler, "Domain-specific enhancement of metacognitive ability following meditation training." *Journal of Experimental Psychology: General*, vol. 143, no. 5, p. 1972, 2014.

[68] L. S. Colzato, A. Szapora, and B. Hommel, "Meditate to create: the impact of focused-attention and open-monitoring training on convergent and divergent thinking," *Frontiers in psychology*, vol. 3, p. 116, 2012.

[69] X. Ding, Y.-Y. Tang, R. Tang, and M. I. Posner, "Improving creativity performance by short-term meditation," *Behavioral and Brain Functions*, vol. 10, no. 1, p. 9, 2014.

[70] X. Ding, Y.-Y. Tang, Y. Deng, R. Tang, and M. I. Posner, "Mood and personality predict improvement in creativity due to meditation training," *Learning and Individual Differences*, vol. 37, pp. 217–221, 2015.

[71] J. Greenberg, K. Reiner, and N. Meiran, "Mind the trap: mindfulness practice reduces cognitive rigidity," *PloS one*, vol. 7, no. 5, p. e36206, 2012.

[72] C. G. Jensen, S. Vangkilde, V. Frokjaer, and S. G. Hasselbalch, "Mindfulness training affects attention—or is it attentional effort?" *Journal of Experimental Psychology: General*, vol. 141, no. 1, p. 106, 2012.

[73] E. L. Lykins, R. A. Baer, and L. R. Gottlob, "Performance-based tests of attention and memory in long-term mindfulness meditators and demographically matched nonmeditators," *Cognitive therapy and research*, vol. 36, no. 1, pp. 103–114, 2012.

[74] E. Schötz, S. Otten, M. Wittmann, S. Schmidt, N. Kohls, and K. Meissner, "Time perception, mindfulness and attentional capacities in transcendental meditators and matched controls," *Personality and Individual Differences*, vol. 93, pp. 16–21, 2016.

[75] S. L. Shapiro, G. E. Schwartz, and G. Bonner, "Effects of mindfulness-based stress reduction on medical and premedical students," *Journal of behavioral medicine*, vol. 21, no. 6, pp. 581–599, 1998.

[76] A. P. Jha, E. A. Stanley, A. Kiyonaga, L. Wong, and L. Gelfand, "Examining the protective effects of mindfulness training on working

memory capacity and affective experience." *Emotion*, vol. 10, no. 1, p. 54, 2010.

[77] D. Navarro. (2018) Learning statistics with R: A tutorial for psychology students and other beginners (version 0.6). [Online]. Available: https://learningstatisticswithr.com/

[78] J. Kramer, "Is abstraction the key to computing?" *Commun. ACM*, vol. 50, no. 4, pp. 36–42, Apr. 2007.

[79] J. A. Parejo, S. Segura, P. Fernandez, and A. Ruiz-Cortés, "Exemplar: An experimental information repository for software engineering research," in *Actas de las XIX Jornadas de Ingeniería del Software y Bases de Datos (JISBD'2014)*, 2014, pp. 155–159.

[80] J. A. Parejo, "Moses: A software ecosystem for metaheuristic optimization," Ph.D. dissertation, University of Sevilla, 2013.

Antonio Ruiz–Cortés is Professor of Computer Science at the University of Seville and elected member of Academia Europæ. He leads the Unit of Excellence of the I3US Institute, and the ISA research group. His current research focuses on service–oriented computing, business process management, experimental software engineering, testing, and software product lines, being the recipient of the MIP Award of SPLC'2017 and VaMoS'2020. He is an associate editor of Springer Computing.

Beatriz Bernárdez is an assistant professor at the University of Seville. Her current research focuses on empirical software engineering, requirements engineering and the application of mindfulness in software engineering. She has the *Search Inside Yourself* Certification of the personal growth program developed by Google. She has served as a reviewer for the IEEE TSE and EMSE Journal and has collaborated in the organization of international conferences such as ESEM'2007, SPLC'2017, and ICSE'2021.

Amador Durán is an associate professor of Software Engineering at the University of Seville. His current research focuses on requirements engineering, empirical software engineering, business process modeling, metamorphic testing, software variability, and formal methods. He is the author of the requirements management tool REM, used by universities and companies in several countries. He also serves regularly as a reviewer for international journals and conferences.

José A. Parejo received his PhD in 2013 from the University of Seville. During the review process of this article, he became a tenured associate professor of the Department of Computer Languages and Systems at the same university. His research interests include empirical software engineering, metaheuristic optimization and simulation, RESTful web services, and the combination of all of them. He is the author of the EXEMPLAR platform, a repository for experiment replication.

Natalia Juristo is a full professor of Software Engineering at the Universidad Politécnica de Madrid and holds a Finland distinguish professor (FiDiPro) research grant since 2013. Her main research interests include experimental software engineering, requirements, and testing. She co–authored the book *Basics of Software Engineering Experimentation*. She is a member of the editorial boards of IEEE TSE and EMSE Journal. She is a senior member of the IEEE.