An Experimental Replication on the Effect of the Practice of Mindfulness in Conceptual Modeling Performance ☆

Beatriz Bernárdez[a,*], Amador Durán[a], José A. Parejo[a], Antonio Ruiz–Cortés[a]

[a]*Department of Computer Languages and Systems, University of Seville*
E.T.S.I. Informática, Av. Reina Mercedes s/n, 41012, Seville, Spain

The final publication is available at Elsevier via `http://dx.doi.org/10.1016/j.jss.2016.06.104`

Abstract

Context: Mindfulness is a meditation technique aimed to increase clearness of mind and awareness. In the 2013–2014 academic year, an experiment was carried out to test whether the practice of mindfulness during 4 weeks improved or not the conceptual modeling performance using UML class diagrams of 32 second–year students of Software Engineering at the University of Seville.
Objective: An internal replication with some changes in the original design was performed in the first semester of the 2014–2015 academic year in order to confirm the insights provided by the original study and increase the confidence in its conclusions. The sample were 53 students with the same profile than in the original study.
Method: Half the students (27 subjects) practiced mindfulness during 6 weeks, while the other half (26 subjects), i.e. the *control group*, received no treatment during that time. All the students developed two conceptual models using UML class diagrams from a transcript of an interview, one before and another after the 6 weeks of mindfulness sessions, and the results were compared in terms of conceptual modeling effectiveness and efficiency.
Results: The results of both experiments were similar, showing that the practice of mindfulness significantly improves conceptual modeling efficiency. Regarding conceptual modeling effectiveness, an improvement is observed in practice, but the analysis shows that such improvement is not statistically significant. After a reanalysis of data, consistent results have also been obtained.
Conclusion: After a replication that leads to the same conclusions as the original study, the adequacy of the original experiment is confirmed and the credibility of its results is increased. Thus, we can state that the practice of mindfulness can improve the efficiency of Software Engineering students in the development of conceptual models, although further experimentation is needed in order to confirm the results in other contexts and other Software Engineering activities different from conceptual modeling.

Keywords: Mindfulness, Replication, Conceptual Modeling, Software Psychology

1. Introduction

Mindfulness is a meditation technique which has demonstrated to be useful for, among other things, educating attention and enhancing mental clarity, thus improving problem–solving capabilities, as described by Davis and Hayes (2011), Tan (2012), and Mrazek et al. (2013), among others. After experimenting the benefits of mindfulness at personal and professional levels for some years, we considered that the students in the Degree in Software Engineering at the University of Seville could also benefit from the practice of mindfulness, especially in a technique such as conceptual modeling in which concentration and clearness of mind is so important. In order to confirm our intuition, an experiment—the *original study*—was carried out during the first semester of the 2013–2014 academic year (Bernárdez et al., 2014). In that experiment, a group of students attended a mindfulness training workshop during four weeks, whereas a second group of students—the *control* group—attended a *placebo* training workshop about public speaking during the same amount of time. Two conceptual modeling exercises using UML class diagrams were performed by all the students (see Appendix A), one before and another after participating in the corresponding training workshop, and their performance were compared.

The conclusions of the original study were promising. After some weeks of practicing mindfulness, evidence suggested that students have a better performance in conceptual modeling compared to the students not practicing mindfulness; i.e. students practicing mindfulness create models of similar quality faster. However, the results in the original study regarding effectiveness—whether students practicing mindfulness produce better conceptual models or not—were not fully conclusive. Some improvement was observed on average, but the differences were not statistically significant.

Thus, we decided to replicate the experiment following a *same experiment & same objects* approach (Gómez et al., 2014)

☆This work was partially supported by the EU Commission (FEDER), the Spanish and Andalusian R&D&I program grants TAPAS (TIN2012–32273), COPAS (P12–TIC–1867) and THEOS (TIC–5906).

*Corresponding author. Tel +34954553870
Email addresses: `beat@us.es` (Beatriz Bernárdez), `amador@us.es` (Amador Durán), `japarejo@us.es` (José A. Parejo), `aruiz@us.es` (Antonio Ruiz–Cortés)

with a twofold purpose: to check the experiment results in order to increase the validity and reliability of the observed outcomes, i.e. the main goal of replications according to Juristo and Gómez (2012); and to overcome some limitations of the original experimental design.

As described by De Magalhães et al. (2014), the publications about replications in SE either i) present one or more replications of an original study, or ii) contribute some knowledge on replication, i.e. process, guidelines, lessons learned, taxonomies, etc. This article corresponds mainly to the first of De Magalhães et al. categories since it presents an internal replication of an original study previously developed by the authors (Bernárdez et al., 2014). To a lesser extent, it also contributes to the second category by providing some lessons learned during the replication (see Section 7.1).

Considering two of the main problems reported by Da Silva et al. (2014) with respect to replication presentation, i.e. the lack of a widely accepted guideline for reporting an experiment replication in Software Engineering (SE), as Carver (2010) comments; and the unavailability of *lab–packages*, that leads also to an increased difficulty for external replications, this work has been organized based on the proposal by Jedlitschka et al. (2008), following some of the recommendations by Carver (2010), and the corresponding lab–pack is available at `https://exemplar.us.es/demo/BernardezJSS2016`.

Specifically, the rest of the article is organized as follows: in Section 2, the practice of mindfulness is briefly described; in Section 3, a summary of the original study is presented; in Section 4, the replication is thoroughly described; in Section 5, the outcomes of both experiments are compared; in Section 6, related work is commented; finally, in Section 7, the conclusions, lessons learned about replications and the future work are presented.

2. The Practice of Mindfulness

The term *mindfulness*—the translation into English of the Pali word *sati*, a Buddhist concept meaning awareness, attention, and remembering (Simón, 2013)—refers to a practice in which a person or a group of people draw away to a quiet place for meditating during at least ten minutes. During meditation, the intention of the mindfulness practitioner is keeping her mind calmed and focused only on breathing (the usual meditation support because of its unavoidability), discarding any other thoughts that could come to mind. The usual steps for a mindfulness session, based on the recommendations of Puddicombe (2011) and Simón (2013), are summarized in Table 1.

The goal 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 developing the ability to keep focused through acknowledging and abandoning thoughts without identifying ourselves with them, mindfulness helps us to perceive our environment clearly and to solve problems more efficiently by reducing mental wandering while performing tasks.

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.
3	Close your eyes and scan your body, relaxing each body part one at a time.
4	Take three slow, deep breaths.
5	Begin to breathe normally, but focusing on your breathing.
6	If thoughts come to you, simply acknowledge them, set them aside, and return your attention to your breath.
7	Enjoy the rare chance to let your mind simply *be*.
8	When you are ready to end your practice, bring your conscious attention back to your surroundings and open your eyes slowly.

2.1. Neurological effects of mindfulness

At a neurological level, the effects of mindfulness are explained by some changes in brain activity, mainly in the prefrontal cortex, which is the main area involved in problem solving, as described by Seligman (2012). A hyperactivity of the prefrontal cortex has the undesired effects of rumination and wandering that, paradoxically, prevent us from solving problems properly and having a clear vision of reality, as commented by Simón (2013). This hyperactivity is one of the consequences of our current relationship with technology, i.e. the ubiquity of Internet–connected devices and the continuous interruptions they generate from social networks, email systems, etc. making very difficult to focus on only one task at a time Gordhamer (2013). Some neuroscientists like Brefczynski-Lewis et al. (2007), Lutz et al. (2009), and Brewer et al. (2011) have demonstrated that a continued practice of mindfulness reduces prefrontal cortex hyperactivity while increases the activity of other areas of the brain which are active when concrete tasks are performed.

2.2. Psychological and social benefits of mindfulness

In 1979, Jon Kabat–Zinn founded the *Stress Reduction Clinic* at the University of Massachusetts Medical School and started to apply mindfulness as a therapeutic treatment in the *Mindfulness–Based Stress Reduction* (MBSR) program[1] (Kabat-Zinn, 2003). Other mindfulness–based therapeutic programs have been also successfully applied to individuals prone to anxiety and other chronic diseases, as reported by Grossman et al. (2004), Shapiro et al. (2005) and Germer et al. (2013). For example, in Riebel et al. (2001), neuro–psychologists studied the effects of mindfulness in 136 heterogeneous patients showing that, after two months of daily 20–minute practice, a significant percentage experienced better personal well–being in terms of mental clarity, equanimity, wisdom and self–compassion based on standard health surveys (questionnaires).

[1] `http://www.umassmed.edu/cfm/stress-reduction/`

The benefits of the practice of mindfulness in students have also been reported. For example, Schure et al. (2008) present a qualitative study examining the influence of mindfulness in a 15–week course with graduate students. Participants reported an increase of their mental clarity, organization, awareness, and acceptance of emotions and personal issues. Mrazek et al. (2013) describe a controlled experiment based on *Graduate Record Examinations* (GRE) assessing verbal, quantitative and analytical skills to measure reading comprehension, concentration, level of mind wandering, and working memory capacity. The outcomes showed a great improvement in the group of people that attended 15 mindfulness sessions.

For a detailed literature review about the main benefits of mindfulness at a personal level, i.e. self–control, self–regard, equanimity, self–awareness, self–insight, intuition, regulation emotion, and well—being in general, see the work by Davis and Hayes (2011).

With respect to social relations, the main benefits of mindfulness are related to empathy, assertiveness, emotion regulation, decreased reactivity, increased response flexibility, counseling skills, and emotional intelligence in general, as reported by Davis and Hayes (2011). Benefits of mindfulness in labor relations, especially in stressful working areas like health or teaching, have also been reported by Poulin et al. (2008), showing better levels of emotional exhaustion, life satisfaction, and teaching self–efficacy (in the case of teachers) in those subjects practicing mindfulness.

2.3. Mindfulness in software engineering

In some software industries in Silicon Valley, the practice of mindfulness is fostered arguing improvements in employee relationships, such as reacting less emotionally, better memory and executive functions, and increased ability to concentrate on fast–changing stimuli, as reported by Shachtman (2013). Particularly in Google, engineer Chade–Men Tan (2012) is developing a mindfulness–based program for understanding co–workers' motivations, enhancing their creativity and productivity, and developing emotional intelligence. Matook and Kautz (2008) and Vidgen and Wang (2009) recommend the practice of mindfulness for software developers using agile methodologies, in order to create a good atmosphere in work groups, in the daily stand–up meetings, in the review and retrospective meetings, and in the interactions with customers and users, etc. (Sutherland (2014)).

Psycho–social aspects are critical factors in SE in general, but they have a special influence in the Requirements Engineering (RE) phase, when interaction with customers and users is more critical for the project success than in any other phase, as described by Davis (1995). It is essential to put oneself in the shoes of customers and users in order to understand their position and the needs to be satisfied by the software system to be developed, as prescribed by Shneiderman (1980) in his classical book *Software Psychology*. More often than not, software developers are not experts in the problem domain at the inception phase of a project. Therefore, software engineers—or more specifically, *requirements engineers*—should develop the skills of understanding the problem domain as it actually is, being open–minded, avoiding excessive simplification, and focusing all their attention on eliciting users needs (Capretz, 2003; Sammon et al., 2014).

For all these reasons, we think it should be considered that software engineers in general, and requirements engineers in particular, should have some notions of mindfulness and practice it. The benefits of mindfulness at a personal level[2] have been widely reported and acknowledged by the Psychology community (Young, 2012), but the ultimate goal is to improve the SE process and software quality. Specifically, our research is focused on RE, and in conceptual modeling in particular, because is the most *social* phase of SE, because of our 20–year experience in the field, and because we teach RE–related subjects, thus making experimental studies with our students feasible.

3. Original Study

In this section, the original study (Bernárdez et al., 2014) which, as commented in the introductory section, took place during the first semester of the 2013–2014 academic year, is briefly described following an adaptation of the proposal by Carver (2010).

3.1. Research questions

The research question that was the basis for the experimental design in the original study was the following:

RQ1 *Has the practice of mindfulness some effect on the performance of students in conceptual modeling?*

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

RQ1.1 *Has the practice of mindfulness some effect on the effectiveness of students in conceptual modeling?*

RQ1.2 *Has the practice of mindfulness some effect on the efficiency of students in conceptual modeling?*

Alternatively, using the *Goal–Question–Metric* (GQM) template recommended by Wohlin et al. (2012), the experiment carried out in the original study can be specified as:

Analyze the practice of mindfulness

for the purpose of evaluating its effects

with respect to the performance of students in conceptual modeling

from the point of view of the experimenters

in the context of second–year students in the Degree in Software Engineering at the University of Seville.

[2]For a very illustrative *infographic* about the personal benefits of mindfulness, visit `http://www.informationisbeautiful.net/visualizations/what-is-meditation-mindfulness-good-for/`.

3.2. Participants

As shown in Figure 1, after the introductory presentation and the recruitment, 75 out of 87 second–year students which were enrolled in the *Introduction to Software Engineering and Information Systems* (ISEIS) annual course, showed some interest in participating in the original study. Depending on their preferences expressed in an interest questionnaire, the participants were divided into two groups: the mindfulness group (G_1, 38 subjects) and the control group attending the *placebo* public speaking workshop (G_2, 37 subjects). As a result, the original study was therefore a *quasi–experiment*[3], due to the absence of random assignment. Despite of the disadvantages of this assignment mechanism, as described by Gliner et al. (2009) and Juristo and Moreno (2001), it was chosen because we considered individuals' motivation essential to avoid mortality and to ensure the treatment (i.e. mindfulness or public speaking) was actually applied.

Of the 75 initial subjects, the sample was finally composed by the 35 subjects satisfying the selection criteria, i.e. having developed both conceptual models using UML class diagrams (and not any other modeling notation), and having attended at least 11 out of every 16 sessions of the corresponding workshop. After discarding 3 outliers, the filtered sample was constituted by 32 second–year students (30 men, 2 women), 16 in each group. They were offered half–a–point bonus in their first–semester ISEIS grade[4] for taking part in the experiment in order to increase their motivation.

Considering that the assignment of subjects was not random, the differences between G_1 and G_2 were examined before continuing with the experiment. Following Campbell and Julian (1963), a one–way analysis of variance, i.e. a one–way ANOVA, was conducted on each of the outcome measures in order to check the *similarity of groups* before the treatment was started. Since the null hypotheses were not rejected (the *p*–value of effectiveness was 0,896 and the *p*–value of efficiency was 0,816), there was no evidence of significant differences between the groups.

3.3. Independent variables

In the original study, we considered that the independent variables or *factors* that were likely to have an impact on the results were the following:

- Training Workshop (**TRWK**): this factor represents the training workshop in which the students participated. It has two levels, *mindfulness* and *public speaking*. Since they are the same than those performed in the replication, a detailed description of the *mindfulness* and *public speaking* workshops are provided in sections 4.4.2 and 4.4.3 respectively.

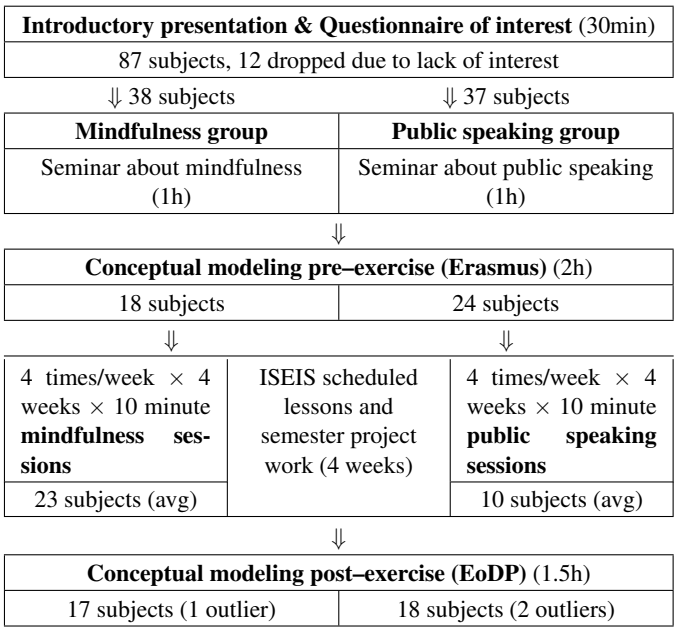

Figure 1: Schedule and number of subjects in the original study

- Conceptual Modeling Exercise (**CMEX**): this factor has two levels, *pre–exercise* and *post–exercise*, which correspond respectively to the conceptual modeling exercises performed by the students before and after participating in the training workshops.

3.4. Dependent variables

As described in Section 3.1, the dependent variables in the original study were conceptual modeling *effectiveness* and *efficiency*. In general terms, effectiveness can be defined as the degree to which objectives are achieved. In the original study, conceptual modeling effectiveness was defined as the achieved *semantic quality* (Genero et al., 2012) of the conceptual models developed by the subjects, i.e. how similar they were to a *reference* conceptual model developed by the experimenters. On the other hand, conceptual modeling efficiency was interpreted as the semantic quality achieved per unit of time.

In order to provide values for these two dependent variables an indirect measure, *semantic quality* (SEM_Q), was used[5]. The semantic quality measures the completeness and correctness (Genero et al., 2012) of the models developed by the experimental subjects with respect to a *reference* model that properly represent the relevant problem domain concepts described in the interview transcripts used as the inputs of the conceptual modeling exercises (see Section 3.7 and Appendix A for details). Assuming UML class diagrams as the conceptual modeling language, model elements were considered as *correctly identified* if there existed semantically equivalent elements in the reference conceptual model. The expression for computing SEM_Q is the following:

$$\text{SEM}_Q = \text{CLASS}_{\text{OK}} - \frac{\text{CLASS}_{\text{KO}}}{2} + \text{ASSOC}_{\text{OK}} + \text{ATTR}_{\text{OK}}$$

[3] A *quasi–experiment* is a controlled experiment in which the assignment of treatments to subjects is not random. See Wohlin et al. (2012) for details.

[4] In Spain, grades are usually in the range $[0, 10]$, considering a course as passed when the obtained grade is greater than or equal to 5. Annual courses such as ISEIS has two *partial* grades, one for the first semester and another one for the second semester.

[5] In the original presentation of the original study (Bernárdez et al., 2014), this measure was named SEMEX, after *semantic expressiveness*.

in which CLASS_{OK}, ASSOC_{OK} and ATTR_{OK} are the number of correctly identified classes, associations and attributes respectively, and CLASS_{KO} is the number of classes incorrectly identified, a correction factor introduced to penalize spurious model elements. Having defined SEM_Q, conceptual modeling effectiveness and efficiency can be specified as follows:

Effectiveness: the percentage of semantic quality achieved by a subject measured in a ratio scale:

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

where CLASS_R, ASSOC_R, ATTR_R are respectively the number of classes, associations, and attributes in the reference conceptual model previously agreed by the experimenters, i.e. their sum is the semantic quality of the reference model.

Efficiency: following the recommendations by Kitchenham et al. (2002), efficiency is defined as the semantic quality achieved per unit of time (minutes):

$$\text{EFFICIENCY} = \frac{\text{SEM}_Q}{\text{TIME}}$$

3.5. Context variables

We identified some context variables or *parameters*, i.e. other independent variables that were controlled at a fixed level during the experiment (Wohlin et al., 2012). As recommended by Juristo and Moreno (2001), these parameters and how they were controlled are defined below in order to facilitate the experiment replication.

- The background of the students in conceptual modeling: students with prior knowledge and practice in conceptual modeling would have performed better in the conceptual modeling exercises than the rest of the students. In order to avoid this situation, we tried to have a sample as homogeneous as possible, discarding all the subjects who were either a repeater student or had previous experience in conceptual modeling.

- The ISEIS scheduled lessons taught to the students: in order to avoid any difference in the content and methodology of the ISEIS scheduled lessons taught to the students—which included an introduction to conceptual modeling using UML class diagrams and some related exercises—all the students had the same professor and the same content was taught to all of them at the same pace.

- The complexity of the conceptual modeling exercises and the order in which they were performed by the students: in order to properly compare the results of the conceptual modeling exercises before and after the training workshop sessions, they had to have a similar complexity and a similar level of familiarity of the students with their problem domains.

  Table 2: Structural measures of interviews and reference conceptual models

Structural measures	Pre-exercise	Post-exercise
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

  On one hand, had the two conceptual modeling exercises been very different in complexity, the results of both exercises would have not been comparable, i.e. the students would have scored a much higher score in a simple conceptual modeling exercise than in a complex one. In order to control this context variable, both exercises were chosen with a similar complexity (see Table 2).

  On the other hand, considering the limited time the subjects had to develop the conceptual modeling exercises, unfamiliar problem domains could have had a very strong impact in the outcomes. In order to control this context variable, the problem domains of the two exercises were chosen taking into account their familiarity to the students, i.e. the pre–exercise was about Erasmus grants, whereas the post–exercise was about the management of End–of–Degree projects (EoDP).

  Considering that both exercises were similar in complexity and familiarity to the students, they order in which they were performed was not considered as relevant and was therefore chosen randomly.

- The number of sessions on each training workshop: in order to observe the effect of the mindfulness practice in the students, a period of four weeks with four 10 minute sessions per week was initially considered as enough, although the possibility of increasing the number of sessions in future replications according to the observed outcomes was always an option.

3.6. Design

Since all the ISEIS students had to attend the scheduled lessons and work on their semester projects during their participation in the training workshops, they were all supposed to increase their performance in conceptual modeling. The goal of the original study was therefore to know whether the aforementioned increase was bigger in the students practicing mindfulness when compared to those students in the control group practicing public speaking.

Taking into account that CMEX is a *within–subjects* factor and TRWK is a *between–subjects* factor, a 2×2 mixed factorial design (Campbell and Julian, 1963) was chosen for the original study. This is a common experimental design not only in the fields of Psychology and Medicine—when the evolution after a certain amount of time of patients under a given therapeutic

treatment needs to be studied—but also in some studies related to mindfulness such as those performed by Poulin et al. (2008), Schure et al. (2008) and Mrazek et al. (2013). In this kind of experimental design, each subject is assigned to one single treatment (e.g. mindfulness), usually including a *placebo* treatment (e.g. public speaking), and two repeated measures on the response variables (e.g. conceptual modeling effectiveness and efficiency), are taken before (e.g. pre–exercise) and after (post–exercise) the application of the treatment under study in order to evaluate its effects.

Considering the chosen experimental design, the resulting tasks for each group are shown in Figure 1. Each column displays the progression of each group, including the number of students who performed each task. Those tasks performed by both groups are represented as one single row occupying both columns. Since the number of students in the final task, i.e. the *sample size*, was different in both groups, the resultant data set was therefore unbalanced. A detailed description of the tasks is provided in section 4.4, since they are the same than those used in the replication.

3.7. Artifacts

The artifacts used in the original study, which are available in the lab–pack at `https://exemplar.us.es/demo/BernardezJSS2016`, are the following in order of appearance in the experiment schedule in Figure 1:

- The slides of the introductory presentation for the recruitment of students.
- The questionnaire about the interest on participating in the experiment, the choice of the training workshop (mindfulness or public speaking), and a commitment to attend the training workshop sessions of choice.
- The slides of the introductory seminars for each of the training workshops (mindfulness and public speaking).
- The two conceptual modeling exercises and their corresponding reference conceptual models. Both exercises had the same dynamics, i.e. the students had to develop a conceptual model based on a transcription of an interview between a customer and a requirements engineer they were provided with in a sheet of paper (see Appendix A). Both exercises were aligned with the goals and material covered in the rest of the course, as recommended by Carver et al. (2003).

3.8. Summary of results

The main result of the original study was that the practice of mindfulness made ISEIS students capable of achieving similar results in conceptual modeling that the students in the control group, but in less time, i.e. they became more *efficient*.

With respect to the effect on conceptual modeling effectiveness, although not negligible, was not statistically significant. Its average value after the training workshops was higher than before for both groups, as depicted in the profile plot in Figure 2. This improvement is probably due to students' knowledge of conceptual modeling had been improved due to the ISEIS lessons and semester project. It was also noticeable the steeper slope of the line showing the improvement of the mindfulness group. As expected, the performed mixed–model ANOVA analysis (see Table 3) revealed a significant effect for CMEX at the $\alpha = 0.01$ level, i.e. both groups had a statistically better effectiveness in the post–exercise than in the pre–exercise. However, the interaction between CMEX and TRWK was not significant at the $\alpha = 0.01$ level, i.e. although the mean of effectiveness varied significantly for both exercises, the effect of the treatment (TRWK) was not necessarily linked to these differences.

With respect to conceptual modeling efficiency, its average value after the training workshops was also higher than before for both groups, as depicted in the profile plot in Figure 3. In this case, the performed mixed–model ANOVA analysis revealed not only a significant effect at the $\alpha = 0.01$ level for CMEX, but also for the interaction between CMEX and TRWK. This meant that, although the mean efficiency varied significantly for both exercises in both groups, the practice of mindfulness is linked to a higher improvement in the efficiency.

4. Experimental replication

In this section, the replication of the original study is presented following some of the recommendations by Carver (2010) and using the guidelines proposed by Jedlitschka et al. (2008).

4.1. Motivation for conducting the replication

As commented in the introductory section, the replication had a twofold purpose. On one hand, the validation of the results of the original study presented in Section 3, in order to increase the validity and reliability of the observed outcomes following a *same experiment & same objects* approach (Gómez et al., 2014). On the other hand, to overcome some limitations of the experimental design of the original study. This twofold purpose led to the following motivations for conducting the replication:

- Confirming our intuition about the benefits of the practice of mindfulness by obtaining statistically significant results not only on conceptual modeling efficiency (as observed in the original study), but also on conceptual modeling effectiveness.
- Mitigating the selection and assignment bias threat and avoiding the limitations on statistical analysis caused by the non–random assignment of students to groups in the original study.
- Avoiding any disturbing factor generated by the placebo treatment (public speaking) on the experiment outcomes, according to the feedback obtained during the presentation of the original study at the *International Symposium on Empirical Software Engineering and Measurement* (ESEM) held in Torino in September 2014.

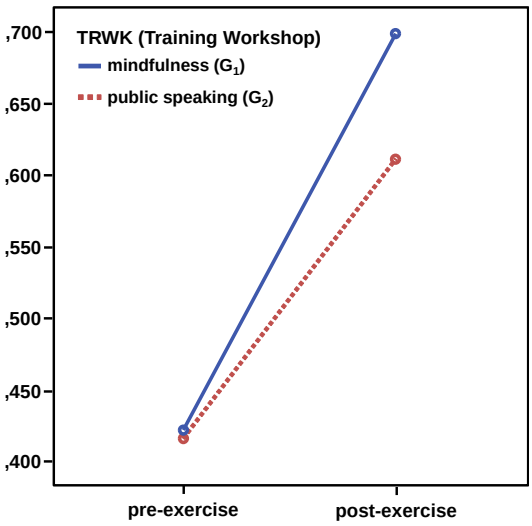

Figure 2: Profile plot of mean of effectiveness in the original study

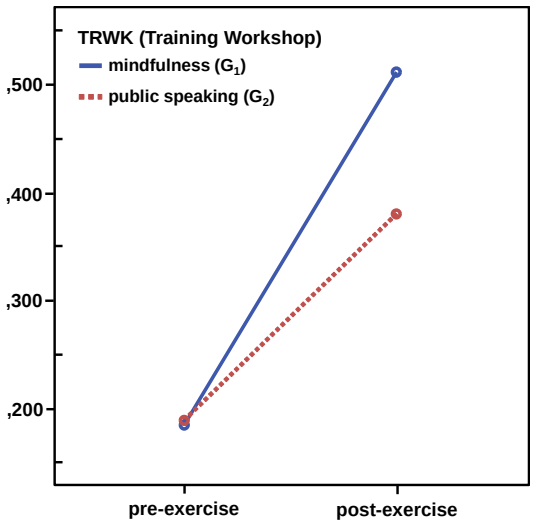

Figure 3: Profile plot of mean of efficiency in the original study

Table 3: Mixed–model ANOVA of conceptual modeling effectiveness in the original study

Source of variation	Type III Sum of Squares	Degrees of Freedom	Mean Square	F–ratio	Significance	η_p^2
CMEX	0.860	1	0.860	90.253	0.000	0.757
CMEX * TRWK	0.026	1	0.026	2.713	0.110	0.086
Error(CMEX)	0.276	29	0.010			

Table 4: Mixed–model ANOVA of conceptual modeling efficiency in the original study

Source of variation	Type III Sum of Squares	Degrees of Freedom	Mean Square	F–ratio	Significance	η_p^2
CMEX	1.047	1	1.047	247.515	0.000	0.895
CMEX * TRWK	0.072	1	0.072	17.001	0.000	0.370
Error(CMEX)	0.123	29	0.004			

Table 5: Summary of changes in the experimental replication

Adjustment	Motivation	Altered dimension	Type of threat
Increasing the number and duration of mindfulness sessions	To achieve a statistically significant improvement of conceptual modeling effectiveness after mindfulness practice	Operationalization	External validity
Random assignment of subjects to training workshops	To mitigate the selection and assignment bias threat and statistical analysis limitations	Protocol	Internal validity
Public speaking workshop postponed after post–exercise	To mitigate the potential placebo disturbing factor on experiment outcomes	Operationalization	Internal validity

4.2. Changes to the original experiment

According to the exposed motivations, the changes carried out in the replication are summarized in Table 5 and described below. In Table 6, a schema of the different replication aspects and the section in which they are defined following the organization proposed by Jedlitschka et al. (2008) are shown. For those aspects which are the same as in the original study, the reader can consult the corresponding definition in Section 3.

4.2.1. First adjustment: more mindfulness sessions

In order to make more evident the benefits of mindfulness and eventually achieve a statistically significant improvement in conceptual modeling effectiveness, we decided to augment the number and duration of mindfulness sessions in the replication. In the original study, the mindfulness training workshop took four weeks with four 10–minute sessions per week. In the replication, the mindfulness workshop took six weeks with

Table 6: Aspects of the experimental replication

Aspect	Description
Goals	Identical to original study
Participants	Section 4.3
Experimental material	Identical to original study
Tasks	Section 4.4
Variables & Parameters	TRWK levels changed to mindfulness and *null*
Design	Identical to original study, but using random assignment
Hypotheses	Identical to original study, together with the new hypotheses in Section 4.5.2
Execution	See figure 4 and its comments below
Analysis, Evaluation & Threats	Section 4.6
Experimenters	Identical to original study, i.e. internal replication

four 12–minute sessions per week. Consequently, the public speaking training workshop was also enlarged with respect to the original study with new tasks to be completed by the students as homework after the face–to–face sessions.

4.2.2. Second adjustment: random assignment

During the original study we were afraid the students left the training workshops due to lack of motivation. In order to mitigate this risk—apart from getting an extra bonus in their qualifications for participating in the experiment—they were allowed to choose which workshop to attend according to their preferences, thus minimizing potential abandon. This choice, i.e. not having a *random* assignment of subjects to groups, affected the experimental design and how resulting data could be analyzed (see Sections 4.5 and 4.6). Nevertheless, during the original study we observed that the students were highly motivated independently of the workshop they were attending, and that some of them even asked for participating in both workshops. This observation led us to use random assignment in the replication.

4.2.3. Third adjustment: null treatment in control group

The third change introduced in the replication was motivated by the feedback obtained during the presentation of the original study at the ESEM'2104 conference (Bernárdez et al., 2014). After the presentation, some questions were posed about the potential effects of the public speaking workshop in the experiment outcomes, despite of our original intention of using it as a *placebo*. Considering this feedback, in the replication the public speaking workshop was held after the experiment (see Figure 4 in Section 4.4). As a consequence, the group control had a *null* treatment and therefore the levels of the independent variable TRWK were changed from *mindfulness* and *public speaking* in the original study to *mindfulness* and *null* in the replication (see Table 6).

Table 7: Participant flow through each stage of the experiment replication

	Interest	Pre–exercise	Attend. (mean)	Post–exercise	Sample
G_1 (mindfulness)	84	40	24,28	28	27
G_2 (null treatment)		42	14,12	29	26

We considered to keep the public speaking workshop in the replication in order to generate a *fairness* atmosphere among the students, avoiding complains about one group obtaining a bonus without having to attend any workshop, and making them feel equally important in the ongoing research independently from the workshop they attended.

4.3. Participants

The experiment replication was carried out during the first semester of the 2014–2015 academic year. Up to 95 students attended the presentation of the ongoing research and a brief introduction to mindfulness. The interest questionnaire described in Section 3.7 was provided to the students, who filled it out manually. Only 11 students showed no interest in participating and were therefore kept aside from the experiment, which was started with 84 students interested in the research, as shown in the first column in Table 7.

In order to perform random assignment, all the questionnaires were marked, half with MF (mindfulness), half with PS (public speaking). Then, they were handed out face down to all the students, who took them blindly and were therefore assigned randomly to the corresponding group, i.e. G_1 for mindfulness and G_2 for null treatment (public speaking after post–exercise). However, in the interest form, students were asked about their preferred training workshop, in order to have this information available in the case of anomalous results had to be analyzed.

The attendance column in Table 7 shows the average number of students who attended the corresponding training workshop sessions.[6] The pre and post–exercise columns show the number of subjects who performed the pre and post–exercises. Finally, the sample column indicates the final number of subjects who where finally considered during analysis and evaluation of this experiment (53 students, 50 male and 3 female).

The used criteria selection was: ISEIS students enrolled for the first time and without any previous experience in mindfulness. In the case of the subjects in the mindfulness group, they had to attend at least the 75% of the sessions in order to be considered in the sample. In the case of the subjects in the control group, they all were chosen since the treatment was null during the experiment.

The difference between the sample size and the number of participants was motivated by the following deviations:

- In the G_1 group, 28 out of the 40 students who performed the pre–exercise, performed also the post–exercise. One

[6]The average attendance in G_2 corresponds to the public speaking training workshop held after the post–exercise was performed.

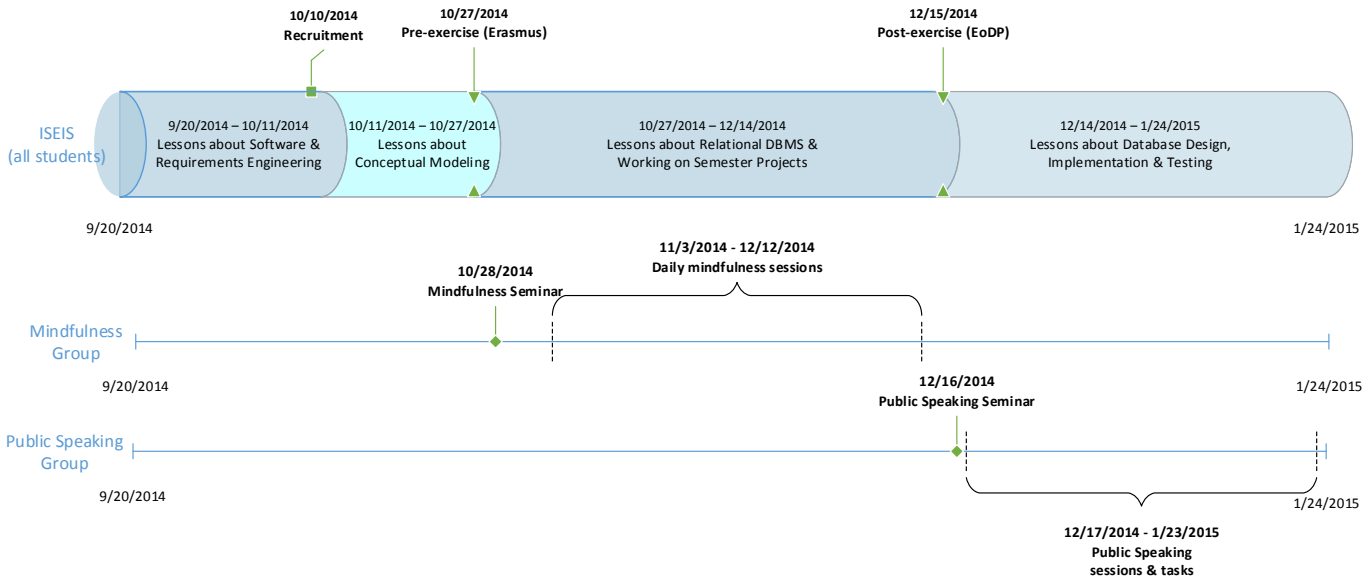

Figure 4: Detailed schedule of the experimental replication

of the students used a wrong modeling notation in the pre–exercise and was therefore excluded from the sample.

- In the G_2 group, 29 out of the 42 students who performed the pre–exercise, performed also the post–exercise. As in the G_1 group, one of the students used a wrong modeling notation in both the pre and post–exercises while other two students were identified as repeaters. Consequently, these three students were excluded from the sample.

4.4. Tasks

In the context of the ISEIS subject, the tasks that were performed during the experiment are shown in Figure 4. The progression of the scheduled ISEIS lessons are shown in the top horizontal axis, whereas the tasks corresponding to the mindfulness and public speaking training workshops are shown in the middle and bottom axis respectively. All of them are detailed below.

4.4.1. Scheduled ISEIS lessons

During the six weeks that the experiment takes, the ISEIS students do not only take the scheduled lessons corresponding to software engineering, requirements engineering, conceptual modeling and relational databases, they also work in groups on their semester project.

In the semester project, the students have to allocate themselves into groups from one to five members and look for a real organization, usually a small business or a nonprofit organization in which a relative or a friend of them works or collaborates. Once they have found an organization to work for, they have to perform a requirements elicitation process, develop a requirements specification, develop a conceptual model and transform the conceptual model into a relational database schema. All the student groups working in a semester project have an professor assigned as their adviser, and they all are supposed to advance in their practical knowledge about conceptual modeling.

4.4.2. Mindfulness training workshop

As previously commented in Section 4.2, in the mindfulness training workshop the sessions were face–to–face, 12–minute long, four days a week, and they took place during six weeks. We thought that this was the most appropriate arrangement for young students in order to facilitate their attendance. Other experimenters (Shapiro et al., 1998; Poulin et al., 2008; Chadwick et al., 2009; Jha et al., 2010) have opted for once–a–week longer sessions or for non–presential sessions supported by an audio guide.

The mindfulness sessions took place during the recess, from 10:20am to 10:40am, always following the same dynamics: the students and the experimenter responsible for conducting the mindfulness sessions[7] meet in a classroom; they all sit down, lights are turned off and curtains are drawn letting only some dim light in the room; when they all are in silence, an alarm is programmed for 12 minutes; during the first five minutes, the students are guided in their body scans (step 3 in Table 1); then, during the remaining 7 minutes, they are invited to focus solely on their breathing. Sometimes, in case some students get distracted during the meditation, the experimenter asks “where is your mind now?” in order to re–focus them on breathing. In the event a student were late, they were instructed to enter the room without making any noise and sit in one of the chairs that were intentionally left empty near the door.

[7]The experimenter who conducted all the mindfulness sessions is the first author of this article. She has taken many courses on mindfulness, studied the most relevant works on the topic and has been a practitioner for several years.

4.4.3. Public speaking training workshop

In the public speaking training workshop, some tasks were performed face–to–face and other were performed individually by the students as homework. With respect to the former, after the initial seminar, the students were given some basic guidelines on how to prepare a speech, some notions on non–verbal communication and some famous speeches were commented in public, e.g. José Saramago's acceptance speech of the Nobel Prize in Literature in 1998[8] and Barack Obama's keynote address at the 2004 Democratic National Convention[9]. A monologue performed by a volunteer student was also analyzed in public.

With respect to the non–presential tasks, the students were invited to look in the internet for a 5–minute video summarizing the keys of public speaking, for a video of a—in their opinion—very good public speaker and to prepare a script of a public presentation on a topic of their interest following the recommended guidelines. All theses non–presential tasks were managed using the *Blackboard* learning management system available at the University of Seville.

4.4.4. Pre and post exercises

The dynamics of the pre and post conceptual modeling exercises were the same: first, the experimenter responsible for conducting the exercise handed out the interview transcript and the answer sheets, both in paper format. Then, a volunteer student read the interview transcript aloud and then the students wrote the start time in their answer sheets; after analyzing the interview transcript, they developed a list of potential information requirements in which they identified the relevant concepts of the problem domain that the information system had to store information about; using the list of requirements, they developed the corresponding UML class diagram and, when they were finished, wrote the end time in their answer sheets.

In the post–exercise in the mindfulness group, a mindfulness session was conducted before the beginning of the exercise, i.e. before reading the interview transcript aloud.

4.5. Design and Hypotheses

For the sake of completeness and consistency with the analysis and reporting performed in the original study, two different experimental designs and sets of hypotheses for the replication are provided in the two subsections below.

In the first subsection, the hypotheses that were tested in the original study—and that were tested again in the replication—following the *two–factors* experimental design shown in Table 8, are stated.

In the second subsection, assuming that randomization on the assignment of individuals to groups allows to neglect the bias introduced by the CMEX factor, a different set of hypotheses (distinguished by an asterisk as superscript) are stated considering only TRWK as a independent variable and not taking into account CMEX, following the *one–factor* experimental design shown in Table 9.

[8] Available at `https://www.youtube.com/watch?v=lWpM5A5lBMI`

[9] Available at `https://www.youtube.com/watch?v=eWynt87PaJ0`

Table 8: Two factors, two levels — 2×2 mixed factorial design

		CMEX	
		pre–exercise	post–exercise
TRWK	mindfulness	G_1	G_1
	null treatment	G_2	G_2

Table 9: Two levels — simple *between* subjects design

TRWK	Group
mindfulness	G_1 (post–exercise only)
null treatment	G_2 (post–exercise only)

4.5.1. Hypotheses for two factors

Considering two factors (CMEX and TRWK), both of them with two levels, two groups of hypotheses—one for each dependent variable—were tested in the replication.

Effectiveness hypotheses.

$H_{0,1}$ **:** there is no difference in the conceptual modeling effectiveness of subjects in the pre and post–exercises, i.e. neither the teaching of ISEIS nor the mindfulness sessions have a significant effect on the conceptual modeling effectiveness of subjects. // $H_{1,1} : \neg H_{0,1}$

$H_{0,2}$ **:** there is no difference in the conceptual modeling effectiveness between the subjects who have practiced mindfulness and those subjects in the control group, i.e. the experimenters assume that the differences observed between the conceptual modeling effectiveness for the pre and post–exercises are due to the teaching of ISEIS exclusively. // $H_{1,2} : \neg H_{0,2}$

Efficiency hypotheses.

$H_{0,3}$ **:** there is no difference in the conceptual modeling efficiency of subjects in the pre and post–exercises, i.e. neither the teaching of ISEIS nor the mindfulness sessions have a significant effect on the conceptual modeling efficiency of subjects. // $H_{1,3} : \neg H_{0,3}$

$H_{0,4}$ **:** there is no difference in the conceptual modeling efficiency between the subjects who have practiced mindfulness and those subjects in the control group, i.e. the experimenters assume that the differences observed between the conceptual modeling efficiency for the pre and post–exercises are due to the teaching of ISEIS exclusively. // $H_{1,4} : \neg H_{0,4}$

4.5.2. Hypotheses for one factor

Considering only one factor (TRWK) with two levels, two simple groups of hypotheses, one for each dependent variable, were complementary tested in the experiment replication.

$H^*_{0,1}$ **:** there is no difference in the conceptual modeling effectiveness of subjects who have practiced mindfulness and those subjects in the control group. // $H^*_{1,1} : \neg H^*_{0,1}$

$H^*_{0,2}$ **:** there is no difference in the conceptual modeling efficiency of subjects who have practiced mindfulness and those in the control group. // $H^*_{1,2} : \neg H^*_{0,2}$

4.6. Analysis and evaluation

In this section, apart from the descriptive statistics of the experimental replication, the two sets of hypotheses stated in the previous section are analyzed.

4.6.1. Descriptive statistics

A summary of the descriptive statistics of the conceptual modeling effectiveness and efficiency of the pre and post exercises in each group in the experiment replication is displayed in Table 10. Figures 5 and 6 depict the distribution of the same dependent variables as box plots under the different experimental conditions of the replication. We carefully scrutinized the conceptual models, times and data of subjects 26, 18, and 41, but since no anomalies or causes for exclusion were found, they were considered as genuine outliers.

According to Finney et al. (1998), the probability of a random group assignment to produce a decompensated distribution is very low. Nevertheless, the differences between G_1 and G_2 were examined before starting the mindfulness sessions as a double–check. A one–way ANOVA analysis was conducted on each of the outcome measures in the same way as described in Section 3.2 for the original study. Since the null hypotheses were not rejected (p–values for effectiveness and efficiency were 0.359 and 0.381 respectively), there was no evidence of significant differences between groups.

The means of conceptual modeling effectiveness and efficiency obtained for each level of the independent variables are shown on Figures 7 and 8. As in the original study, the mean for effectiveness after treatment is higher than before for both groups, although in this case the line showing the improvement of the mindfulness group has a steeper slope than the line of the control group. In the post–exercise, the difference in medians for effectiveness is 0.09 with an standard deviation of 0.129 for the control group and 0.135 for the mindfulness group. In the post–exercise, the difference in medians for efficiency is 0.141 with an standard deviation of 0.082 for the control group and 0.122 for the mindfulness group. Regarding correctly identified elements (classes, attributes and associations), subjects from the mindfulness group can identify up to a 10% of more on average in the post–exercise (2.2 correct elements more). Furthermore, subjects from the mindfulness group can identify up to 0.128 elements more per minute (7 elements more per hour) than their counterparts in the control group on average.

4.6.2. Analysis based on two factors with two treatments

In order to determine whether parametric or non–parametric tests could be used, two different tests were applied to the obtained data. First, a Shapiro–Wilk normality test (see Table 11) was performed in order to check whether the obtained data followed or not a normal distribution. Then, a Levene test was also performed to check for the *homoscedasticity*, i.e. the homogeneity of variances (see Table 12).

The results of the Shapiro–Wilk test in Table 11, showed that, for all the obtained data except the corresponding to the conceptual modeling effectiveness in the post–exercise of group G_1, there was no evidence to reject the normality hypothesis at the level of $\alpha = 0.05$. On the other hand, the results of the Levene test showed that there was not evidence to reject the homoscedasticity hypothesis.

Although the data corresponding to conceptual modeling effectiveness the post–exercise of G_1 could not be considered as normal, a mixed–model ANOVA with TRWK as a *between–subjects* factor and CMEX as a *within–subjects* factor was applied because of its robustness in such situations, as described by Glass et al. (1972).

The analysis summarized in Table 13, corresponding to the results obtained by means of a mixed–model ANOVA for conceptual modeling effectiveness, revealed a significant effect at the $\alpha = 0.01$ level for CMEX, i.e. both groups had a statistically better effectiveness after the treatment, measured by the percentage of semantic expressiveness achieved. However, the interaction between CMEX and TRWK was not significant at the $\alpha = 0.01$ level. This means that, although the mean of conceptual modeling effectiveness varied significantly from one exercise to another, the effect of TRWK is not necessarily linked to these differences. Therefore, the null hypothesis $H_{0,1}$ is rejected at the $\alpha = 0.01$ significance level, but there is not enough evidence as to reject $H_{0,2}$ at such significance level.

The analysis summarized in Table 14, corresponding to the results obtained for conceptual modeling efficiency, revealed a significant effect at the $\alpha = 0.01$ level for CMEX, i.e. both groups had a statistically better efficiency after the treatment. Furthermore, the interaction between CMEX and TRWK produced also a statistically significant difference at the $\alpha = 0.01$ level. This means that the mean of effectiveness varied significantly from one exercise to another, and that the effect of the TRWK is linked to these differences. Therefore, the null hypotheses $H_{0,3}$ and $H_{0,4}$ are rejected at the $\alpha = 0.01$ significance level.

4.6.3. Analysis based on one factor with two treatments

The use of random assignment of individuals to groups allowed to carry out a much simpler analysis, assuming that the effects of the SE learning and the order in which the pre and post–exercises were performed, impacted equally to all subjects. Thus, a simple two–group comparison is applied for both dependent variables using the data obtained in the conceptual modeling post–exercise.

Since the groups were unbalanced (see Table 7), the two–sample t-test by Welch (1947) was applied instead of a classic t–student test. In the case of conceptual modeling effectiveness, the p–value was 0.1981. As a consequence, there was not evidence enough as to reject the null hypothesis ($H^*_{0,1}$). A 95 percent confidence interval of the difference in the means between groups was also generated, which ranged from -0.025

Table 10: Descriptive statistics of the experiment replication

	Conceptual modeling effectiveness				Conceptual modeling efficiency			
	null treatment		mindfulness		null treatment		mindfulness	
	pre-exercise	post-exercise	pre-exercise	post-exercise	pre-exercise	post-exercise	pre-exercise	post-exercise
n	26	26	27	27	26	26	27	27
mean	0,405	0,472	0,378	0,519	0,231	0,279	0,253	0,408
sd	0,103	0,129	0,110	0,135	0,064	0,082	0,107	0,122
median	0,435	0,460	0,370	0,550	0,240	0,273	0,243	0,424
min	0,100	0,260	0,110	0,120	0,063	0,134	0,070	0,109
max	0,590	0,730	0,530	0,700	0,330	0,425	0,550	0,614
range	0,490	0,470	0,420	0,580	0,268	0,291	0,480	0,505

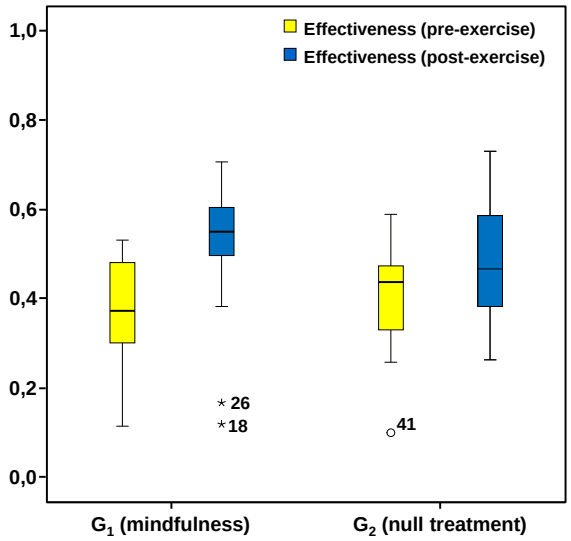

Figure 5: Box plot of conceptual modeling effectiveness

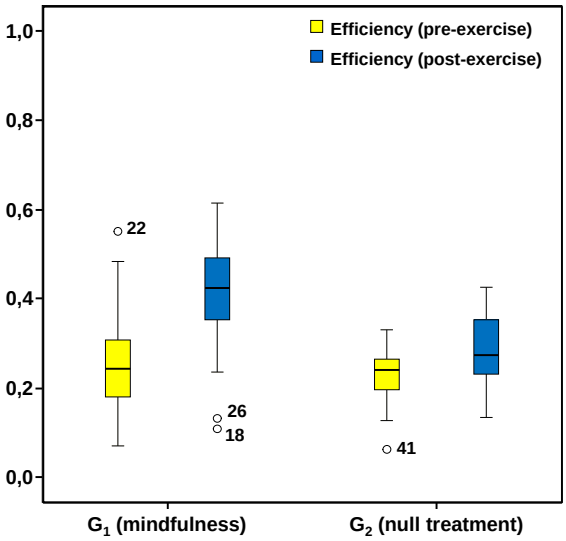

Figure 6: Box plot of conceptual modeling efficiency

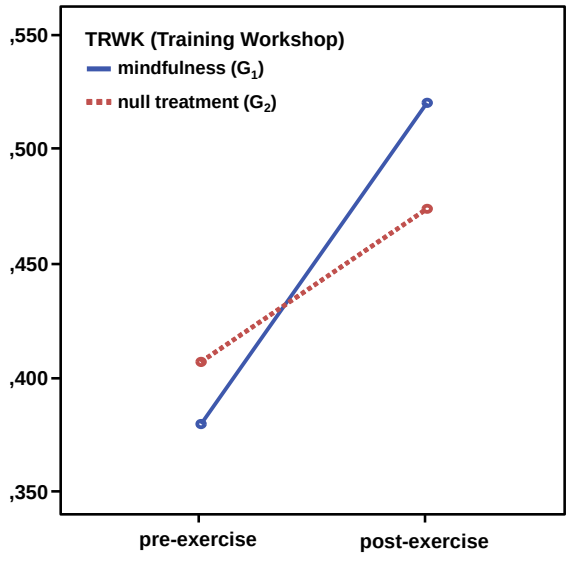

Figure 7: Profile plot of conceptual modeling effectiveness

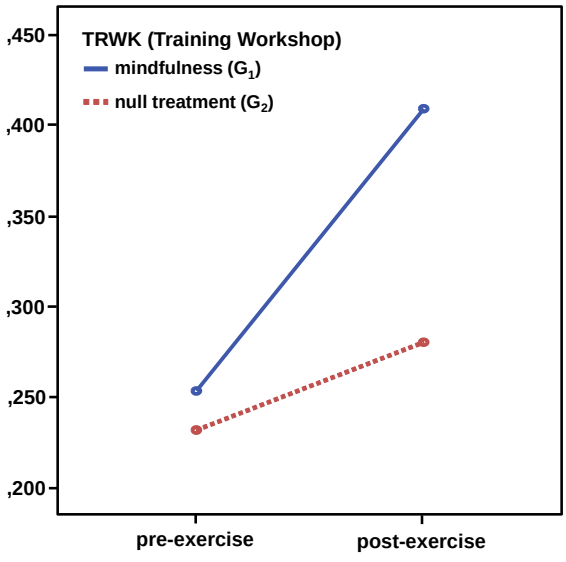

Figure 8: Profile plot of conceptual modeling efficiency

Table 11: Shapiro–Wilk normality test results

Dependent variable	Group	Shapiro–Wilk significance level
Effectiveness	G_1 (pre–exercise)	0.112
	G_2 (pre–exercise)	0.121
	G_1 (post–exercise)	0.001
	G_2 (post–exercise)	0.500
Efficiency	G_1 (pre–exercise)	0.124
	G_2 (pre–exercise)	0.446
	G_1 (post–exercise)	0.248
	G_2 (post–exercise)	0.421

Table 12: Levene test results

Dependent variable	F value	Levene significance level
Effectiveness	0.4849	0.6935
Efficiency	2.0816	0.1073

Table 13: Mixed–model ANOVA of effectiveness in the experiment replication

Source of variation	Type III Sum of Squares	Degrees of Freedom	Mean Square	F–ratio	Significance	η_p^2
CMEX	0.286	1	0.286	23.146	0.000	0.312
CMEX * TRWK	0.036	1	0.036	2.908	0.094	0.054
Error(CMEX)	0.629	51	0.012			

Table 14: Mixed–model ANOVA of efficiency in the experiment replication

Source of variation	Type III Sum of Squares	Degrees of Freedom	Mean Square	F–ratio	Significance	η_p^2
CMEX	0.275	1	0.275	31.469	0.000	0.382
CMEX * TRWK	0.076	1	0.076	8.689	0.005	0.146
Error(CMEX)	0.446	51	0.009			

to 0.12, which confirms the absence of evidence as to reject the null hypothesis, since it contains the value 0. The *effect size* is a quantitative measure of the strength of the impact of the treatment, that provide information about the magnitude and direction of the difference between the two groups. Cohen's d statistic is a popular effect size estimator for the difference between the means of two samples. The value of Cohen's d statistic for the effectiveness was 0.358, which is in the middle of the cut–offs introduced by Cohen (2013) himself for small effects (0.25) and medium effects (0.5). Thus, we deduced that although there was no evidence enough such as to claim that there is a statistically significant difference in conceptual modeling effectiveness, it seems to be an actual difference in practice.

Regarding conceptual modeling efficiency, the two–samples t–test provided a p–value of 0.0004. Thus, it was concluded that the difference between the efficiency of the groups is statistically significant, and the null hypothesis ($H^*_{0,2}$) was rejected in favor of the alternative hypothesis. Regarding the effect size, Cohen's d statistic value was 1.233, which denoted a large effect size that translated in an important speed improvement in practice. Actually, this means that 88% of the students in the mindfulness group would be above the mean of the control group (Cohen's U3, as defined in (Cohen, 2013)). Furthermore, there is a 80% chance that a student picked at random from the mindfulness group would have a higher efficiency than a student picked at random from the control group, i.e. the probability of superiority is 80% .

4.7. Threats to validity on the experimental replication

Wohlin et al. (2012) provides a thorough compilation of threats to the validity of empirical studies. In this section, the threats related to Wohlin et al.'s *conclusion*, *internal*, *construct*, and *external* validities are analyzed and the actions performed to mitigate them are described. Regarding to internal validity, only *multiple groups* and *social* threats are considered, since *simple group* threats refer to situations in which there is no control group and that is not the case of our work.

4.7.1. Threats to conclusion validity

The conclusion validity is concerned with the statistical relationship between the treatment and the outcome. The main threats of this category are analyzed below.

Low statistical power & violated assumptions of statistical tests. In the original study, the main threat to conclusion validity was the small size of the sample. Nevertheless, although

small, the sample size was acceptable for the statistical tests applied, as described by Juristo and Moreno (2001), and all the assumptions for each statistical test were verified before their application. In the replication, the sample size was significantly larger, all the assumptions for each test were verified before their application, thus neutralizing these two threats.

Fishing & the error rate. Fishing refers the impact of researchers looking for a specific outcome of the experiment. In the replication, subjects were assigned to groups randomly, and only those subjects who missed one of the conceptual modeling exercises were excluded from the sample. The assessment of the exercises was blinded, without knowing the subject's group. Moreover, since no other outlier removal or data post processing was applied, the fishing threat could be considered as neutralized. With regard to the error rate threat, it did not affect the conclusions of this study because there were no multiple comparison statistical tests, and the simple comparison tests were applied only once per hypothesis.

Reliability of measures. In order to avoid discrepancies and inaccuracies, the assessment of the exercises was performed by the same person based on consensual reference solutions (see appendix A), thus neutralizing this threat.

Reliability of treatment implementation. In order to ensure the reliability of the treatment implementation, we opted for face–to–face, 12–minute long sessions led always by the same person, thus making easy to check that subjects were actually receiving the treatment (see Section 4.4.2 for details).

Random irrelevancies in experimental setting. This threat relates to elements outside the experimental setting which may disturb the results. Regarding to ISEIS sessions, the attendance of students was similar in both groups and the sessions were taught by the same professor. Additionally, no disturbances or interruptions were observed during the conceptual modeling exercises.

Random heterogeneity of subjects. This threat takes into account the risk of the variation due to individual differences being larger than the variation due to the treatment. A statistical test using the data from the pre–exercise was performed in order to ensure that the groups were similar regarding the outcome variables, both in terms of their means and their variances (see Section 4.6.1 for details).

4.7.2. Threats to internal validity – multiple groups

In order to analyze multiple groups threats, Wohlin et al. suggest to review whether the experimental group, i.e the mindfulness group, and the control group may be affected differently by the single group threats which are analyzed below.

History. Since both groups performed the experiment tasks simultaneously, without any significant incident, this threat was neutralized.

Maturation. With respect to their knowledge in SE, both groups maturated simultaneously since they all attended the same number of ISEIS sessions, with the same professor and content, between the pre and post–exercises. Therefore, this threat was also neutralized.

Testing. This threat refers to the effect on the outcomes when performing a test twice, due to the knowledge about the test gained by subjects in the first test. Indeed, due to the chosen experimental design, subjects performed two conceptual modeling exercises and maturated and improved their modeling skills due to the ISEIS lessons. Nevertheless, the effect of the treatment is analyzed taking into account such influence by studying the interaction CMEX *TRWK (see ANOVA for mixed–design on tables 13 and 14). Furthermore, such testing effect should affect both groups evenly.

Instrumentation. In order to avoid interaction of both treatments, i.e. a potential placebo side–effect on response variables in the experiment, the public speaking sessions were taught after performing the post–exercise in the replication (see Section 4.2.3) thus neutralizing this thread.

Statistical regression. This threat occurs when subjects are assigned to groups based on previous studies. This threat had no impact in the replication because groups were randomly assigned.

Selection. This threat is related to a selection of subjects producing non–equivalent groups. The three usual reasons for this threat are i) non–random assignment; ii) sample size too small; and iii) higher motivation in volunteer subjects than in the whole population. The two first reasons were neutralized in the replication because of the random assignment (see Section 4.2.2) and because of the bigger size of the sample than in the original study (see Section 4.3). With respect to the third reason, both groups can be considered as being equally motivated due to random assignment.

Mortality. In order to reduce this threat, the students were offered a half–a–point bonus for participating in the experiment (see Section 3.2). Furthermore, no subjects abandoned the mindfulness sessions due to tedium or lack of interest.

Interactions with selection. This type of threat is due to different behavior in different groups. In the replication, the selection and assignment of individuals to groups was random, which usually neutralizes this threat. Furthermore, the authors performed an additional crosscheck by computing a one–way ANOVA test on the measures of the pre–exercise dependent variables. The results of the test revealed that there were not evidence of significant differences between groups prior to treatment.

4.7.3. Threats to internal validity – social

This group of threats refers to the impact of mental perceptions, psychology and social interactions of the subjects on the outcomes of the experiment.

Diffusion or imitation of treatments. This threat refers to situations in which the control group learns about the treatment being applied to the experimental group. Although it is possible but not likely that some subjects in the control group learned and practiced mindfulness on their own, we do not think that such practice took place as regularly as in the experimental group. Additionally, they did not practice mindfulness just before the post–exercise as the experimental group did (see Section 4.4.4).

Compensatory equalization of treatments. There are two situations in which this threat could affect an experiment. One is whether a control group is given a compensation as a substitute for not getting any treatments; depending on the compensation, this might affect the outcome of the experiment. In our case, subjects were compensated equally in both groups, thus this threat was neutralized. The other situation takes place when the control group is treated with a placebo as a compensation for not receiving the actual treatment and the placebo treatment may have an impact on the experiment outcomes. In such case, the placebo treatment would not be an actual placebo, but an alternative treatment. As commented in Section 4.2.3, this threat was one of the main criticisms posed after the presentation of the original study at the ESEM'2014 conference. In order to neutralize this threat in the replication, we applied the placebo treatment after performing the second conceptual modeling exercise, ensuring that the control group received a null treatment. Thus, this threat was fully neutralized.

Compensatory rivalry & resentful demoralization. This threat refers to situations where the motivation of subjects is different due to their perception of the treatments, thinking that one treatment is better than another. In the replication, although the treatment of the control group was null, the subjects received an appealing treatment (public speaking) after the second measurement. Moreover, the professors gave the same importance to both treatments without promoting one over the other.

4.7.4. Construct validity

The construct validity is concerned with the relation between theory and observation, i.e. whether variables correctly represent the theoretical constructs or not. Wohlin et al. (2012) split it into two categories: *design* and *social* threats.

Regarding to design threats, in order to avoid *inadequate preoperational explication of constructs*, we defined beforehand how to perform the conceptual modeling exercises, and a consensual reference solution was agreed for each exercise. Furthermore, all the conceptual modeling exercises were assessed by the same person, in order to avoid inconsistencies in the assessment. Besides, the *mono–operation bias* was reduced by the inclusion of two different treatments and two exercises in the experiment. Similarly, the *mono–method bias* was also reduced by considering two different dependent variables, i.e. conceptual modeling effectiveness and efficiency.

Regarding *confounding constructs and levels of constructs*, apart from the treatments, we did not measure any variable using levels of presence or absence of a construct, avoiding as a consequence such confusion. The *interaction of different treatments* did not affect this study since each subject belonged to a single group, and no individual received both trainings. *Interaction of testing and treatment* did not affect this experiment either since the exercises of conceptual modeling were very different from the training workshops, i.e. the conceptual modeling exercises itself cannot make the subjects more sensitive or receptive to the training workshops effects. Finally, *restricted generalizability across constructs* states that treatment could affect another relevant construct negatively. No negative effects of the practice of mindfulness have been reported in the literature, thus we consider that this threat does not affect our study.

Regarding *social threats to construct validity*, the experimenters tried to maintain a distance and an aseptic attitude with the subjects, both in the training workshops and in the conceptual modeling exercises. We tried not to influence the decisions of the subjects and revealed as minimum information as possible about the experiment, explicitly avoiding *hypothesis guessing*. We avoided also the use of the term *experiment*, using *research* instead, in order to avoid that subjects could feel observed. Furthermore, we did not revealed the variables that will be measured to the subjects.

4.7.5. External validity

The greater the external validity, the more the results of an empirical study can be generalized to current SE practice. The two identified threats that could limit such generalization are analyzed below.

Interaction of setting and treatment. Regarding to the materials used, the size of the interview transcripts might not be representative of industrial problems, but it was appropriate for the available time for the pre and post–exercises. However, we think that the intellectual processes applied during conceptual modeling—potentially improved by mindfulness—are basically the same regardless of the size of the problem at hand.

Interaction of selection and treatment. Regarding to experimental subjects, since the tasks performed during the pre and post–exercises did not require high levels of industrial experience, using students as subjects instead of SE professionals could be considered as appropriate (Porter et al., 1999). Moreover, students are the next generation of professionals, so they are close to the population under study (Kitchenham et al., 2002).

Finally, on the replication, both workshops and the conceptual modeling exercises occurred without any incidents, so it appears that the *interaction of history and treatment* had no effect during the execution of the experiment.

5. Comparison of results to original

The results of both experiments are similar for the analysis based on ANOVA for mixed design, and such results are in concordance with those provided by the complementary analysis carried out in the replication. According to Lindsay and Ehrenberg (1993), since the results of the two studies match, we have

confirmed that the result is robust to all the changes performed in the replication.

Regarding the absolute values of conceptual modeling effectiveness and efficiency, it is noteworthy that the scores of both dependent variables in the replication (academic year 2014–2015) are lower than in the original study (academic year 2013–2014). In order to determine whether the dissimilarity between academic years were significant, two two–sample t-tests on both dependent variables for the difference between academic years were performed. In the case of conceptual modeling effectiveness, the p–value generated by the test was 0.0002, i.e. the difference between the effectiveness of the different academic years is statistically significant. Regarding efficiency, the same test provided a p–value of 0.0004, i.e. the difference between the efficiency of the different academic years is also statistically significant. Therefore, it seems that the conceptual modeling level of the sample in the 2013–2014 academic year is better than those of the sample in 2014–2015.

Searching for the possible causes of this dissimilarity, the qualifications of the ISEIS students of both academic years for the first–semester exam were analyzed. We found that the average grade for ISEIS students in 2013–2014 was 6.2, and the percentage of students who passed the exam was 81%. Conversely, the average score in 2014–2015 was 5.2, and the percentage of students who passed the exam was 73%. Furthermore, we checked again that the contents and pace of the ISEIS lessons was the same for all groups in both years and that did not change from one year to another. Therefore, there is not a clear cause for this dissimilarity apart from the intrinsic subjects' variability on each academic year.

Consequently, we can state that the results of the treatment are similar in the original study and in the replication under different circumstances, i.e. (i) with samples of students with very different performance, as described above; (ii) when performing the public speaking training workshop concurrently with the mindfulness training workshop and when performing it later; and (iii) whether the students who practiced mindfulness did it on their preference or were randomly assigned.

If each dependent variable is analyzed, the results in conceptual modeling efficiency are resounding in both experiments. Regarding effectiveness, it is conceivable that extending the mindfulness practice from 4 to 6 weeks would have improved it, making students not only develop conceptual models faster, but also better. Figure 9, shows the average of the differences between the exercises for conceptual modeling effectiveness per group and year. It is easy to appreciate that the ratio of the differences between the mindfulness group and the control group is much bigger for the replication (2014–2015). We interpret such increase as an effect of extending the mindfulness practice treatment from 4 to 6 weeks. However, such increase is not enough such as to make the differences in conceptual modeling effectiveness statistically significant.

Finally, the data from the original study and from the replication cannot be analyzed together due to the differences in the experimental design. However, if a third experiment with a similar design to the replication presented in this article carried out, a combined analysis or some meta–analysis could be performed. Such an additional replication would allow to change the parameter corresponding to the order in which the conceptual modeling exercises are performed, evaluating its impact on the results (see Section 7.2).

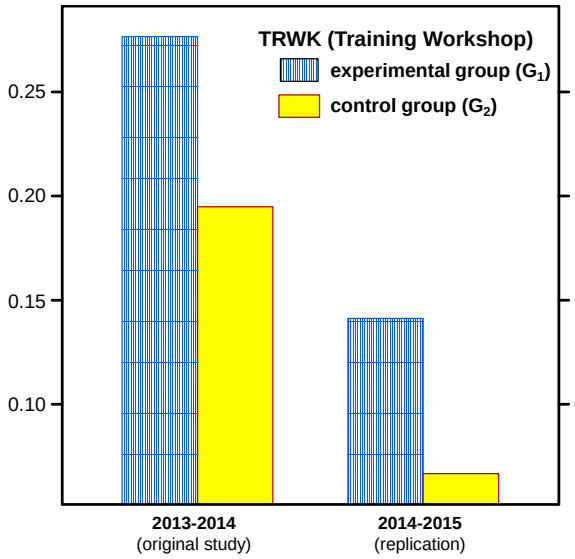

Figure 9: Bar–chart of the average differences of conceptual modeling effectiveness between pre and post–exercises per group and academic year

6. Related work

In this article, two research fields overlap when reviewing related works: replications in empirical SE, specifically in RE, and replications related to mindfulness in Psychology and Medicine.

With respect to the former field, there are several studies—some of them cited in previous sections—and the main challenges are well known (Da Silva et al., 2014). Nevertheless, some groups of experiments related to requirements and conceptual model inspection with the same design as ours can be highlighted, such as the ones by Basili et al. (1996) and Miller et al. (1998), Laitenberger et al. (2000) or some of those featured in the compendium by Regnell et al. (2000).

Focusing on the area of replications and UML modeling (class diagrams and other diagrams), the main families of experiments are set out in the surveys by Moody (2005), Budgen et al. (2011) and Genero et al. (2012). Most of these studies analyze how the use or characteristics of the model impacts on other properties of any software artifact. For instance, Kuzniarz et al. (2004); Lange et al. (2006); Genero et al. (2007) study their impact on understandability, maintainability and defects density of the model. Alternatively, Marchesi (1998); Chen et al. (2004); Scanniello et al. (2014) study its relation with the properties of the source code, such as its size, understandability or modifiability. This work is linked to that approach since mindfulness is proposed as a means for improving the conceptual modeling performance of subjects. It is clear from the works cited above that improving that performance impacts on the software process or product.

The experimental design used in the research presented in this article and based on repeated–measures is common in SE, but usually with the aim to reduce threats to internal validity due to individual differences or when the number of participants is not so large (Scanniello et al., 2014). However, we have used it with the purpose of studying the evolution of people over time under a specific type of treatment, as in Psychology or Medicine.

To the best of our knowledge, within the field of experimental replications related to mindfulness, the following are the studies that have been conducted in a replication:

Shapiro et al. (2005). This work presents a third study in a randomized controlled study that implemented a 2 (experimental vs. wait–list control group) x 2 (baseline, post–treatment) study design. The three studies examine whether psychological distress, stress and job burnout decrease after participants were involved in an 8–week MBSR program (eight 2–hour sessions, 1 session per week).

The main difference of this replication compared with previous studies lies with the participants, in that it is the first applied to health care professionals rather than students (Shapiro et al., 1998; Jain et al., 2004).

Of the 18 subjects allocated to the mindfulness group, 8 did not complete the treatment (44%). Reasons included, health issues, family problems or insufficient time, though subjects expressed their interest in the program. One possible solution to avoid the problem of dropping out is to make sessions shorter and located in the subjects' usual place of work or study, and furthermore to take place daily to compensate for the reduced duration. Though, understandably, this may not always be possible.

Regarding the results, significant differences were observed in scores in perceived stress and self–compassion, but not in satisfaction with life and burnout scale. It is interesting to note that results were not compared with those obtained in previous studies.

Bondolfi et al. (2010). This work presents a replication in order to evaluate whether MBCT (*Mindfulness Based Cognitive Therapy*) reduces the risk of depressive relapse when compared with TAU (*Treatment As Usual*). The experiment design is a two levels–simple between subjects design (there is no pre and post–treatment).

The mindfulness sessions are group sessions with eight weekly 2–hour training sessions, and at least 4 MBCT sessions were considered as the minimum.

The main difference of this replication with previous studies is that it took place in Switzerland and the previous in Canada. The adjustments made in this replication arise from the adaptation of medical protocols in the Swiss context. As regards to the results, in the previous experiments differences were observed between the groups, in that the ratio of depressive relapse in the MBCT group was significantly less (40% in the original study and 36% in the first replication) than that of the TAU group (66% in the original study and 78% in the first replication).

In the second replication, however, results differed. A decrease in the ratio of the MBCT group in relation to the TAU group was observed, but it is not statistically significant. Therefore in the replication, the ratio of depressive relapse in the TAU group is 36% and in the MBCT group 33%. The high survival rate in the TAU condition, compared to the two former studies is due, according to the authors, to the excellent quality of the Swiss healthcare system.

The two mindfulness replications summarized above differ from their corresponding original studies in the types of the experimental subjects. In our replication, this does not change specifically, as they are still ISEIS students. In future replications, this could be changed in order to verify whether the results are achieved in a different population.

Regarding the experimental design of the two mindfulness–related replications commented above, the design of the former is the same as the shown in Table 8, and the design of the latter is the same as the one in Table 9.

Finally, we observe that the number of replications in this field are low when compared to the number of original experiments, some of which are cited in Section 2.

7. Conclusions and future works

It is widely known that not only the improvement in skills of software engineers, but also working on certain aspects of their personality impacts on software quality, as described by Acuña et al. (2009) and Kosti et al. (2014). The hypothesis at the start of this study was that the practice of mindfulness would have a positive effect on the conceptual modeling efficiency and effectiveness of SE students.

According to the reviewed literature, we expect that not only SE students, but other types of students (Mrazek et al., 2013) and professionals (Poulin et al., 2008) could benefit from mindfulness practice. Our work is focused in SE students and conceptual modeling because it is our professional field, and because our students, although with reasonably good programming skills, usually find conceptual modeling difficult to apply. We think that some of the well–known benefits of mindfulness, i.e. mental clarity, reading comprehension, concentration, and so forth, could be a relevant support for this task.

Although it is out of the scope of this work, we also think that mindfulness could improve some social skills which are very useful for requirements engineers such as extroversion or teamwork capabilities (Capretz, 2003; Tan, 2012).

In this study, we have presented an experimental replication conducted within the context of second–year students of Software Engineering at the University of Seville. The original study was presented at ESEM'2014 and, given the feedback received, in the interest in verifying the experimental results a replication was undertaken in which minor changes were introduced. Both experiments revealed the same outcome: students who practiced mindfulness obtained similar results in less time than the others, i.e. a significant improvement in conceptual modeling efficiency was observed. However, the observed improvement in effectiveness, although important, was not statistically significant. The results have repeated using random

Table 15: Summary of empirical replications about the effects of the practice of mindfulness

Reference	No. of previous experiments	Effect of mindfulness on	Type of Replication	Changes on replication	Consistent results
Shapiro et al. (2005)	2	Stress	Differentiated	Professional subjects instead of students	Partial
Bondolfi et al. (2010)	2	Depressive relapse	External / independent	Swiss hospitals instead of Canadian hospitals	No
This work	1	Conceptual modeling effectiveness and efficiency	Internal & Differentiated	Random assignment instead of selection	Yes

assignation and avoiding the influence of the public speaking workshop on the scores of the dependent variables in the post–exercise, including also the differences observed between the samples of academic years 2013–2014 with 2014–2015.

The carried out replication has made us more confident in the results of the original study and it encourages us to think that the improvement in efficiency would be quite interesting for SE professionals and organizations that would have their productivity improved.

7.1. Lessons learned

After the conclusion of the experiment replication described in this article, we have identified some *lessons learned* that could be helpful for other researchers in further replications. They are described in this section.

7.1.1. What to tell about the original study and the replication

When the aim of an article is presenting an experiment replication, many doubts arise about what should be told when describing the original study and what should be left for the description of the replication, considering that most information is the same for both experiments. Following the proposal by Carver (2010), we decided to provide just enough details about the original study so that the reader could follow the article but without making the description of the replication reiterative. We have decided to thoroughly describe and justify the adjustments in the replication, referencing the conference paper in which the original study is presented (Bernárdez et al., 2014) for the readers interested in further details.

7.1.2. Be prepared for human nature

Unexpected situations are common when dealing with human subjects in an experiment. For example, in our case one student did not write the end time of his conceptual modeling exercise on the answer sheet. Since the exercises were left by the students on a pile as soon as they finished, we could approximate the end time of this exercise using the answer sheets left before and after on the pile. Another situation, as commented in Section 4.3, was the use of a wrong modeling notation—BPMN—by some of the experimental subjects. This could have been avoided by insisting on the modeling notation the subjects had to use or by using the term "UML class diagram" in the task description instead of the more general term "conceptual modeling".

Recording this kind of situations would help not only internal but also external experimental replications. It would be very interesting to add this knowledge—what happened, how it was solved, and what to do to avoid it in the future—to the lab–packs available for other researchers.

7.1.3. Summary tables promote communication

Last but not least, we have learned how a summary table with the adjustments in the replication (see Table 5 in Section 4.1) and another comparing the aspects of the replication to the original study (see Table 6 in Section 4.2) promotes communication among the authors and with other researches we consulted during our work. For the elaboration of Table 5 we have followed the recommendations by Gómez et al. (2014), whereas Table 6 is based on the proposal by Jedlitschka et al. (2008).

Additionally, we think it would be very useful to develop a template for describing replication adjustments, improving the understanding of their motivations and facilitating external replications.

7.2. Future work

The current status of our ongoing research and the imminent future work is summarized in Figure 10. Replication #2 represents a future controlled experiment similar to Replication #1, in which the context variable *order of exercises* is swapped. This variable, i.e. the effect of the order in which conceptual modeling exercises are conducted, was a randomized variable in replication #1, but in replication #2 we plan to change the order of the exercises, in order to study its impact on the outcomes. This change in the order in which conceptual modeling exercises are conducted will allow to rule out its influence on the obtained results. Additionally, a combined analysis of datasets of the two replications in a joint manner could be performed.

Further on, taking into account that the two replications conducted so far have been internal, our intention is to conduct external replications to assess that similar results can be obtained in any other location and by any other researchers, as Schmidt (2009) recommends. In this replication process, it would be interesting to change the experimental protocols to ensure that the observed results are independent of the procedure, materials, or instruments used in the experiment that arrived at the result (Juristo and Gómez, 2012). For example, one could study the effects of mindfulness on the quality and performance of the sub-

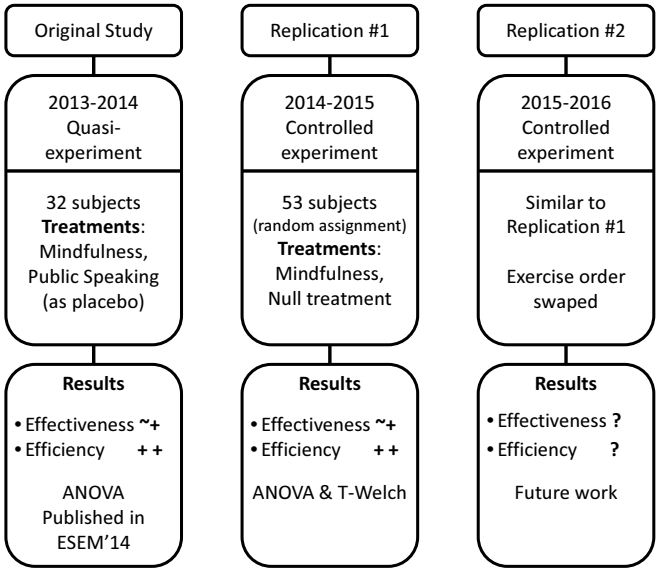

Figure 10: Ongoing research on mindfulness and SE at the University of Seville

jects during other SE activities, such as elicitation and negotiation of requirements, or resolution of technological problems. Another factor to take into account is the the previous knowledge of SE practices, which for student subjects is strongly related to the course in which the experiment takes place. In this regard, it would be advisable to perform further experiments with students in their last courses.

Finally, after conducting experiments within an academic context, we would like to perform some case studies in real software companies in order to develop a mindfulness–based personal growth program for its further adoption.

Acknowledgments

The authors would like to thank the ISEIS students who participated in the experiments for their excellent attitude. We would also like to thank Dr. N. Juristo from the University Politécnica de Madrid for encouraging us to carry out the replication of the original study and for giving priceless support on all our many requests. We also thank the staff of the Escuela Técnica Superior de Ingeniería Informática of the University of Seville for arranging anything we needed in the classrooms for the mindfulness and public speaking sessions. Moreover, we would also like to thank the reviewers and participants in ESEM'2014, and also the reviewers of this article, for their valuable feedback.

References

Acuña, S. T., Gómez, M., Juristo, N., 2009. How do personality, team processes and task characteristics relate to job satisfaction and software quality? Information and Software Technology 51 (3), 627–639.

Basili, V. R., Green, S., Laitenberger, O., Lanubile, F., Shull, F., Sørumgård, S., Zelkowitz, M. V., 1996. The empirical investigation of perspective–based reading. Empirical Software Engineering 1 (2), 133–164.

Bernárdez, B., Durán, A., Parejo, J. A., Ruiz-Cortés, A., 2014. 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, pp. 17–27.

Bondolfi, G., Jermann, F., Van der Linden, M., Gex-Fabry, M., Bizzini, L., Rouget, B. W., Myers-Arrazola, L., Gonzalez, C., Segal, Z., Aubry, J.-M., et al., 2010. Depression relapse prophylaxis with mindfulness-based cognitive therapy: replication and extension in the swiss health care system. Journal of affective disorders 122 (3), 224–231.

Brefczynski-Lewis, J. A., Lutz, A., Schaefer, H., Levinson, D., Davidson, R., 2007. Neural correlates of attentional expertise in long-term meditation practitioners. Proceedings of the national Academy of Sciences 104 (27), 11483–11488.

Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y.-Y., Weber, J., Kober, H., 2011. Meditation experience is associated with differences in default mode network activity and connectivity. Proceedings of the National Academy of Sciences 108 (50), 20254–20259.

Budgen, D., Burn, A. J., Brereton, O. P., Kitchenham, B. A., Pretorius, R., 2011. Empirical evidence about the UML: a systematic literature review. Software: Practice and Experience 41 (4), 363–392.

Campbell, D. T., Julian, S., 1963. Experimental and Quasi–Experimental Designs for Research. Wadsworth, United States.

Capretz, L. F., 2003. Personality types in software engineering. International Journal of Human–Computer Studies 58 (2), 207–214.

Carver, J., Jaccheri, L., Morasca, S., Shull, F., 2003. Issues in using students in empirical studies in software engineering education. In: Software Metrics Symposium, 2003. Proceedings. Ninth International. IEEE, pp. 239–249.

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

Chadwick, P., Hughes, S., Russell, D., Russell, I., Dagnan, D., 2009. Mindfulness groups for distressing voices and paranoia: A replication and randomized feasibility trial. Behavioural and Cognitive Psychotherapy 37 (04), 403–412.

Chen, Y., Boehm, B. W., Madachy, R., Valerdi, R., 2004. An empirical study of eservices product UML sizing metrics. In: Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on. IEEE, pp. 199–206.

Cohen, J., 2013. Statistical power analysis for the behavioral sciences. Academic press.

Da Silva, F. Q., Suassuna, M., França, A. C. C., Grubb, A. M., Gouveia, T. B., Monteiro, C. V., dos Santos, I. E., 2014. Replication of empirical studies in software engineering research: a systematic mapping study. Empirical Software Engineering 19 (3), 501–557.

Davis, A., 1995. 201 principles of software development. McGraw–Hil, New York.

Davis, D. M., Hayes, J. A., 2011. What are the benefits of mindfulness? a practice review of psychotherapy–related research. Psychotherapy 48 (2), 198.

De Magalhães, C. V., Da Silva, F. Q., Santos, R. E., 2014. Investigations about replication of empirical studies in software engineering: preliminary findings from a mapping study. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. ACM, pp. 37–48.

Finney, K., Rennolls, K., Fedorec, A., 1998. Measuring the comprehensibility of z specifications. Journal of Systems and Software 42 (1), 3–15.

Genero, M., Fernández-Saez, A., Nelson, J., Poels, G., Piattini, M., 2012. A systematic literature review on the quality of UML models. Journal of Database Management 22 (3), 46–70.

Genero, M., Manso, E., Visaggio, A., Canfora, G., Piattini, M., 2007. Building measure–based prediction models for UML class diagram maintainability. Empirical Software Engineering 12 (5), 517–549.

Germer, C. K., Siegel, R. D., Funton, P. R., 2013. Mindfulness and Psychotherapy. The Guilford Press, New York.

Glass, G., Peckham, P. D., Sanders, J. R., 1972. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research 42 (1), 237–288.

Gliner, J. A., Morgan, G. A., Leech, N. L., 2009. Research methods in applied settings: an integrated approach to design and analysis (second edition). Tylor and Francis Group.

Gómez, O. S., Juristo, N., Vegas, S., 2014. Understanding replication of exper-

iments in software engineering: A classification. Information and Software Technology 56 (8), 1033–1048.

Gordhamer, S., 2013. Wisdom 2.0: The New Movement Toward Purposeful Engagement in Business and in Life. Tylor and Francis Group.

Grossman, P., Niemann, L., Schmidt, S., Walach, H., 2004. Mindfulness–based stress reduction and health benefits: A meta–analysis. Journal of Psychosomatic Research 57 (1), 35–43.

Jain, S., Shapiro, S., Swanick, S., Bell, I., Schwartz, G., 2004. Mindfulness meditation versus relaxation training for medical, premedical, nursing, and prehealth students: Differential effects on response style and psychological distress. In: Poster presented at the Second Annual Mindfulness in Medicine and Health Care Conference, Worcester, MA. p. 1.

Jedlitschka, A., Ciolkowski, M., D, P., 2008. Guide to Advanced Empirical Software Engineering. Springer Verlag, Ch. Reporting experiments in Software Engineering, pp. 201–228, eds. Forrest Shull and Janice Singer and Dag I.K. Sjøberg.

Jha, A. P., Stanley, E. A., Kiyonaga, A., Wong, L., Gelfand, L., 2010. Examining the protective effects of mindfulness training on working memory capacity and affective experience. Emotion 10 (1), 54.

Juristo, N., Gómez, O., 2012. Replication of software engineering experiments. In: Meyer, B., Nordio, M. (Eds.), Empirical Software Engineering and Verification. Vol. 7007 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 60–88.

Juristo, N., Moreno, A. M., 2001. Basics of Software Engineering Experimentation. Kluwer Academic Publishers.

Kabat-Zinn, J., 2003. Mindfulness-based stress reduction MBSR. Constructivism in the Human Sciences.

Kitchenham, B. A., Pfleeger, S. L., Hoaglin., D., Emam, K. E., Rosenberg, J., August 2002. Preliminary Guidelines for Empirical Research in Software Engineering. IEEE Transactions on Software Engineering 28 (8), 721–734.

Kosti, M. V., Feldt, R., Angelis, L., 2014. Personality, emotional intelligence and work preferences in software engineering: An empirical study. Information and Software Technology 56 (8), 973–990.

Kuzniarz, L., Staron, M., Wohlin, C., 2004. An empirical study on using stereotypes to improve understanding of UML models. In: Program Comprehension, 2004. Proceedings. 12th IEEE International Workshop on. IEEE, pp. 14–23.

Laitenberger, O., Atkinson, C., Schlich, M., El Emam, K., 2000. An experimental comparison of reading techniques for defect detection in UML design documents. Journal of Systems and Software 53 (2), 183–204.

Lange, C. F., DuBois, B., Chaudron, M. R., Demeyer, S., 2006. An experimental investigation of UML modeling conventions. In: Model Driven Engineering Languages and Systems. Springer, pp. 27–41.

Lindsay, R. M., Ehrenberg, A. S., 1993. The design of replicated studies. The American Statistician 47 (3), 217–228.

Lutz, A., Slagter, H. A., Rawlings, N. B., Francis, A. D., Greischar, L. L., Davidson, R. J., 2009. Mental training enhances attentional stability: neural and behavioral evidence. The Journal of Neuroscience 29 (42), 13418–13427.

Marchesi, M., 1998. Ooa metrics for the unified modeling language. In: Software Maintenance and Reengineering, 1998. Proceedings of the Second Euromicro Conference on. IEEE, pp. 67–73.

Matook, S., Kautz, K., December 2008. Mindfulness and agile software development. In: 19th Australasian Conference on Information Systems (ACIS) Proceedings. Association for Information Systems/University of Canterbury, pp. 638–647.

Miller, J., Wood, M., Roper, M., 1998. Further experiences with scenarios and checklists. Empirical Software Engineering 3 (1), 37–64.

Moody, D. L., 2005. Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. Data & Knowledge Engineering 55 (3), 243–276.

Mrazek, M. D., Franklin, M. S., Phillips, D. T., Baird, B., Schooler, J. W., May 2013. Mindfulness training improves working memory capacity and GRE performance while reducing mind wandering. Psychological Science 24 (5), 776–781.

Porter, A. A., Votta, L. G., Basili, V. R., July 1999. Building Knowledge through Families of Experiments. IEEE Transactions on Software Engineering 25 (4), 456–473.

Poulin, P. A., Mackenzie, C. S., Soloway, G., Karayolas, E., 2008. 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 46 (2), 72–80.

Puddicombe, A., 2011. Get some headspace. Hodder and Stoughton, USA.

Regnell, B., Runeson, P., Thelin, T., 2000. Are the perspectives really different?–further experimentation on scenario–based reading of requirements. Empirical Software Engineering 5 (4), 331–356.

Riebel, D., Greeson, J., Brainard, G., Rosenzweig, S., July 2001. Mindfulness–based stress reduction and health–related quality of life in a heterogeneous patient population. Jefferson Myrna Brind Center of Integration Medicine Faculty 2 (1), 1–20.

Sammon, D., Nagle, T., McAvoy, J., 2014. Analysing ISD performance using narrative networks, routines and mindfulness. Information and Software Technology 56 (5).

Scanniello, G., Gravino, C., Genero, M., Cruz-Lemus, J., Tortora, G., 2014. On the impact of UML analysis models on source-code comprehensibility and modifiability. ACM Transactions on Software Engineering and Methodology (TOSEM) 23 (2), 13.

Schmidt, S., 2009. Shall we really do it again? the powerful concept of replication is neglected in the social sciences. Review of General Psychology 13 (2), 90.

Schure, M. B., Christopher, J., Christopher, S., Winter 2008. Mind–body medicine and the art of self-care: Teaching mindfulness to counseling students through yoga, meditation, and qigong. Journal of Counseling and Development 86 (1), 47–56.

Seligman, M., 2012. Flourish: A Visionary New Understanding of Happiness and Well–being. Free Press, New York.

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

Shapiro, S. L., Astin, J. A., Bishop, S. R., Cordova, M., 2005. Mindfulness-based stress reduction for health care professionals: results from a randomized trial. International Journal of Stress Management 12 (2), 164–176.

Shapiro, S. L., Schwartz, G. E., Bonner, G., 1998. Effects of mindfulness-based stress reduction on medical and premedical students. Journal of behavioral medicine 21 (6), 581–599.

Shneiderman, B., 1980. Software psychology: human factors in computer and information systems. Winthrop.

Simón, V., 2013. Aware and Awake. Descleé De Brouwer, Spain.

Sutherland, J., 2014. Scrum: The Art of Doing Twice the Work in Half the Time. Crown Business.

Tan, C.-M., 2012. Search inside yourself. Addison-Wesley Publishing Company, Massachusetts.

Vidgen, R., Wang, X., September 2009. Coevolving systems and the organization of agile software development. Information Systems Research 20 (3), 355–376.

Welch, B. L., 1947. The generalization of 'student's' problem when several different population variances are involved. Biometrika 34 (1-2), 28–35.
URL `http://biomet.oxfordjournals.org/content/34/1-2/28.short`

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

Young, S., 2012. The Science of Enlightenment. Springer Publishing Company, Incorporated.

A. Conceptual modeling exercises

This appendix contains the translation of the problem statements of the two conceptual modeling exercises into English together with their reference solutions.

A.1. Erasmus grants exercise — Problem statement

The transcript below corresponds to an interview with the Erasmus Coordinator of the Higher Technical School of Informatics Engineering (*E.T.S.I. Informática*) of the University of Seville (USE) that was held in order to identify the goals and requirements of a web application for managing the processes related to the Socrates–Erasmus program.

Question: *Well, let's start. As I told you in my previous email, the goal of this interview is to get a first idea of the processes for application and assignment of Socrates–Erasmus program grants in order to computerize as many parts of the processes as possible. I am particularly interested in knowing the information that has to be managed by the web application to be developed.*

Answer: *All right. Where do we start?*

Q: *Let's start by the time an USE student thinks about applying for an Erasmus grant.*

A: *Well, interested students usually browse and query the Erasmus destinations offered by their centres (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 the student 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. Importantly, student information can be checked by the ID in the computer system of the Office of the Vice President for Students using an XML web services interface.*

Q: *Please, tell me about qualified foreign languages.*

A: *Each destination has one or more associated languages that are a requirement for applicants. For example, most universities in countries with little widespread languages (Scandinavian countries, Finland, Poland, Romania, etc.) accept candidates who know their official language or English. Universities in other countries like France or Germany require almost always French or German at least. In order for a student to get a grant, she must either prove that is qualified for some level of a language (e.g. having passed several language courses, a TOEFL or Proficiency exam, etc.) or seat for the Foreign Language Test (FLT) that is freely organized by the Foreign Language Institute (FLI) for Erasmus applicants without qualified foreign languages.*

Q: *What does the IRS do with the EPAs?*

A: *Once the application period is closed, the IRS draws up a schedule of FLTs. In this schedule, the date, time and location of the FLTs organized by the FLI are indicated. Each applicant must take the FLTs corresponding to requested but non–qualified foreign languages. For example, if a student applies for destinations at universities in Germany and France, she has to take the German and French FLTs, assuming that she is not qualified for any of the two languages. We want both the FLT schedule and their results to be managed by the application and to be consulted via Internet.*

Q: *And once the FLTs have been taken?*

A: *Once the FLTs have been taken, the FLI sends the results to the IRS, where the lists of candidates (LoCs) for each USE centre which has offered destinations—almost all—are elaborated. The candidates are those applicants who are qualified for the requested levels of language of a destination, either because they present some evidence, either because they have passed the FLT for some language associated with a destination. Of course, the idea is that this whole process to be managed by the application and that the LoCs are published on the Internet.*

Q: *What do the faculties and schools with the LoCs?*

A: *In the USE centres, the Erasmus Commissions have to propose which applicants become holders of the grants and which become substitutes and in what order, using the criteria they consider as more appropriate. The lists of holders and substitutes (LoHSs) are sent by the Erasmus Commissions of each centre to the IRS. As I mentioned before, the idea is that the LoHSs are published on the Internet using the application.*

Q: *And what does the IRS do once it knows who are selected to be the grant holders?*

A: *Then it sends a Notification Letter (NL) by registered mail to each student selected as a holder. In the NL, several documents are included, but we can detail that in another interview. So far, we could point out that the web application should automatically send an email to the holders.*

Q: *OK, let's say for the moment that once a student receives the NL, she begins to "enjoy the grant." We will go into detail in the following interview. Thank you for your cooperation.*

A: *You're welcome. Happy to help.*

A.2. Erasmus grants exercise — Reference solution

The reference conceptual model corresponding to the solution of the Erasmus grants exercise is shown in the UML class diagram in Figure A.11.

A.3. End–of–Degree project exercise — Problem statement

The goal of the system to be developed is to computerize the management of End–of–Degree Projects (EoDPs) of the Department of Computer Languages and Systems (DoCLS). Below is an interview with one of the professors in the EoDP Commission of the DoCLS:

Question: *What is the main goal of the system to be developed?*

Answer: *Basically, improve the management of EoDPs. Apart from publishing on the web all our EoDP offers so the students*

can be aware of them, we also want to provide IT support for the EoDP–related workflow.
Q: *So, on the one hand publishing EoDP offers, and on the other hand computerize the workflow. About EoDP offers, what information do you want to be published on the web about them?*
A: *In a EoDP offer we publish its title, a description, to what degree (Software Engineering, Computer Engineering or Information Technology) applies and the professors who offer the EoDP.*
Q: *Professors? I thought that an EoDP was offered and mentored by a single professor.*
A: *This is usually the case, but an EoDP can have up to two mentors. Even one of them could be an external expert not being a professor, although there must always be a professor acting as a mentor. I almost forgot, it is also important to know which students are assigned to a certain offer and if an offer supports more applications.*
Q: *Well, I think that has to do with the EoDP workflow and you have not told me about it yet. Now, once an offer is published with the information you have told me, what happens then?*
A: *Well, if a student is interested in working in a EoDP, she contacts the offering professor and, if they reach an agreement, the professor fills out an EoDP Admission Application (AA) and sends it via email to the Subject Coordinator (SC) of EoDPs. If the SC approves the AA, the mentor professor is notified by email and the EoDP formally begins.*
Q: *What information is contained in the AAs?*
A: *An AA contains data about the student applying for the EoDP, (Tax ID number, name, degree, e–mail and telephone), data about the EoDP (title, objectives and technologies to be used), mentor (or mentors) of the EoDP and any comment to be sent to the SC. When the SC approves the AA, the admission date is added to the AA. In case the AA is denied, the date of denial and the justification for the denial are added to the AA.*
Q: *Can the SC deny an AA?*
A: *Yes, the SC can do it if she thinks the EoDP is not appropriate, although this happens very rarely.*
Q: *How can an EoDP be inappropriate if it has been previously offered? What happens in case of denial?*
A: *Well, first EoDP offers are published under the responsibility of each professor, they do not pass any filtering by the SC, so that some EoDPs which are inappropriate from the point of view of the SC could be offered. On the other hand, an AA that does not correspond to any EoDP can be submitted, for example if the student makes an EoDP proposal to a professor, or if a professor offers a student an EoDP not previously offered (or a variant of an EoDP). In any case, if the SC denies an AA, she always provides the reasons, so the professor and the student can re–elaborate the EoDP and submit another AA.*
Q: *Do you find interesting that the new system include the need for the SC to authorize AAs before they are published?*
A: *I hadn't really thought about it, but now that you mention it I find it interesting. Yes, I think that the SC should authorize the publication of EoDP offers.*
Q: *Although there is really no correspondence between EoDP offers and AAs, isn't there?*
A: *Well, there is a correspondence to some extent, but there may be AAs without a previous offer.*
Q: *So, should the new system allow to set relationships between offers and AAs or should we consider offers as merely informative and therefore outside the workflow?*
A: *I think it's best that when a professor enters a AA into the system, she is allowed to use an EoDP offer as a reference (to copy the data into the AA, mainly) and, if desired, to set an association between the offer and the AA, but without being compulsory.*
Q: *How long should an EoDP offer be applicable?*
A: *Until the offering professor or the SC decide to cancel it. Although it would be interesting to add an expiration date, or even better an expiration academic year. Something like "applicable until 2016/17". If an offer is not cancelled, the system should stop showing it after the beginning of its expiration academic year. It would also be interesting to associate offers a maximum number of AAs, so if an offer reaches it, it appears as "assigned" but continues to be displayed on the web while it is not "expired".*
Q: *I've heard that students must publicly defend the EoDP, how is this process carried out?*
A: *The SC elaborates an EoDP Public Defense (PD) calendar in which each EoDP is assigned a committee composed of two professors, indicating the date and time of each defense. This calendar is published to let all stakeholders be aware of it, so it is important that the system stores it too, even that helps the SC in its elaboration.*
Q: *Is any email sent to the students and the members of the committee for citing them for the PD?*
A: *A broadcast email is sent for all professors, but it would be nice if the new system would send an mail to each student notifying the time of her PD, and to teachers as well.*
Q: *What happens during a PD?*
A: *Well, once a student has finished her PD, the committee issues a PD Report providing comments, a grade and a justification of the grade. Days later, the SC publishes the grades as the Provisional Minutes, which if no claims are presented, become the Final Minutes.*
Q: *Should the new system store data about PD reports and minutes?*
A: *Only PD reports. The provisional and final minutes are managed by the Grading System of the University of Seville, which falls outside the responsibilities of the system to be developed.*
Q: *I understand, I think I have enough information to start working. If you agree, in the next meeting we will discuss issues related with time frames for each step of the workflow and exceptional situations such as a student not attending to her PD or not being enrolled in the EoDP subject.*
A: *All right, I will collect information about it to answer your questions as best as possible.*

A.4. End–of–Degree project exercise — Reference solution

The reference conceptual model corresponding to the solution of the End–of–Degree projects exercise is shown in the UML class diagram in Figure A.12.

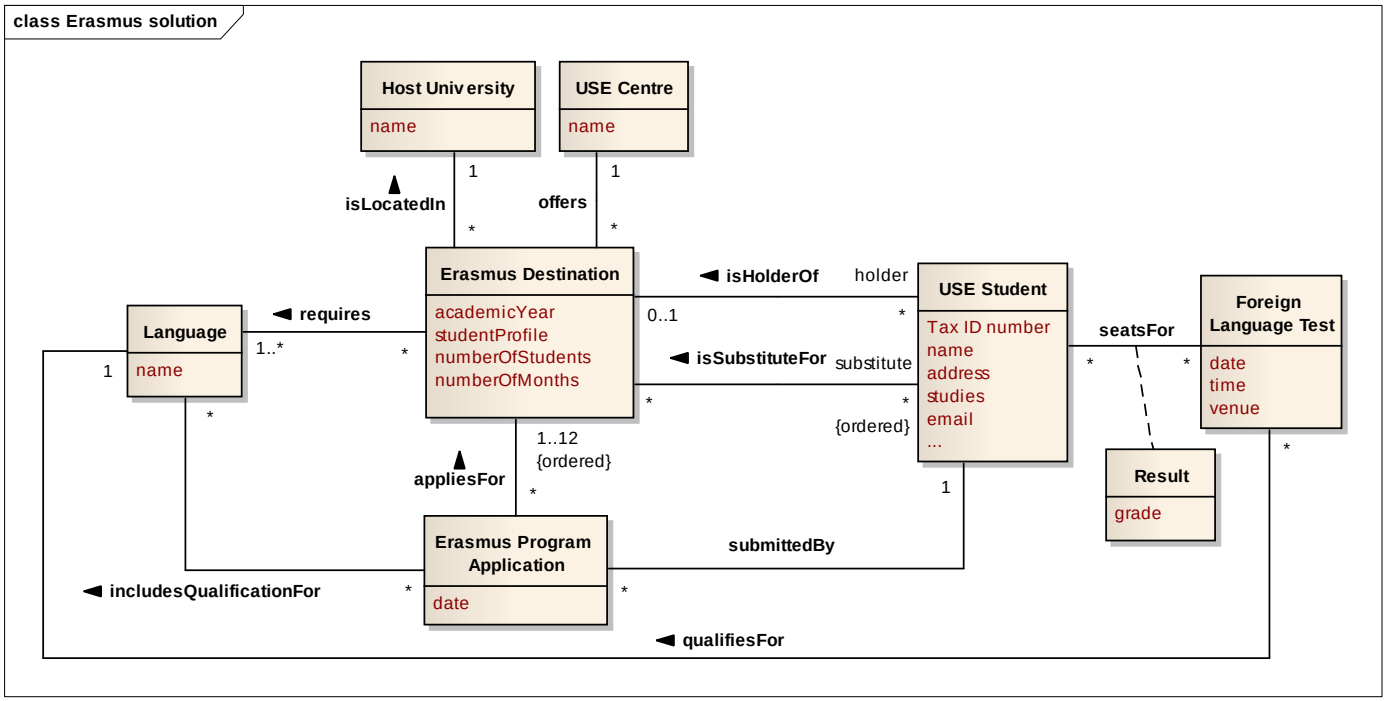

Figure A.11: Erasmus grants exercise reference solution

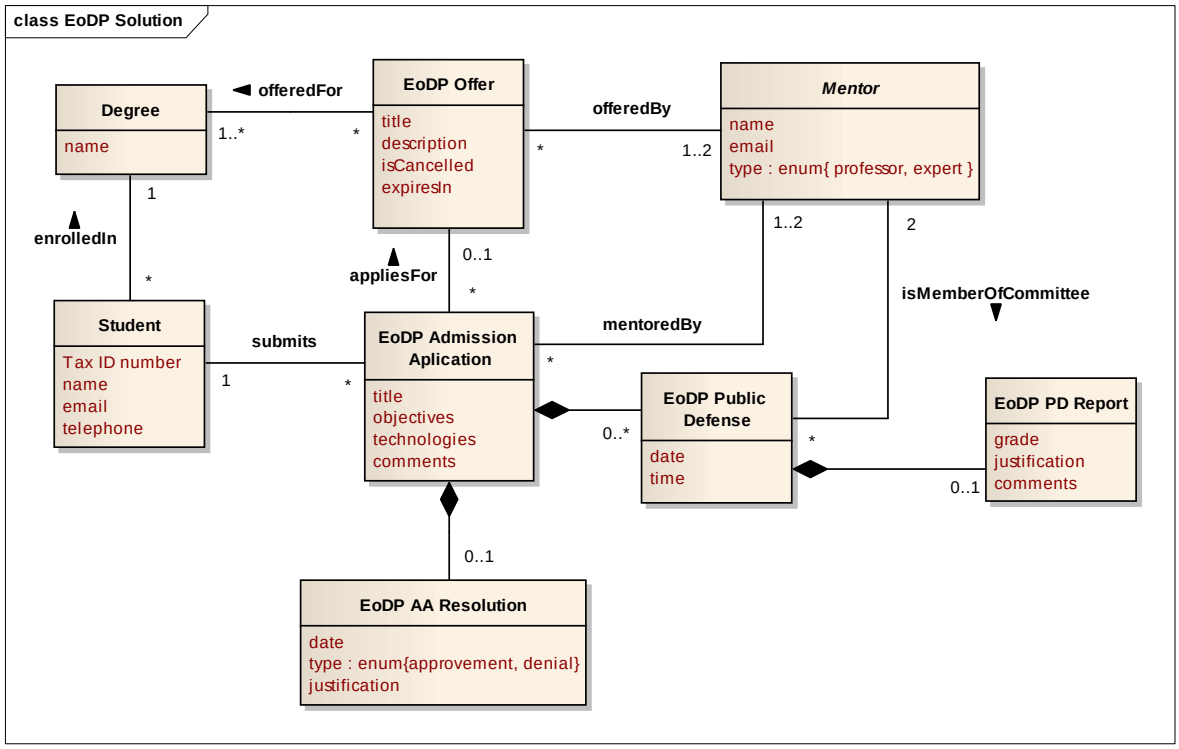

Figure A.12: End–of–Degree Project exercise reference solution