# A EMPIRICAL EVALUATION – EXTENDED VERSION

## A.1 Experiment design and execution

This section presents the empirical experiment that we designed and carried out to validate the usability of our UML profile and the proposed methodology. We followed the basic methodology for conducting usability studies [49]—which is derived from the classical approach for conducting controlled experiments—as well as the Empirical Standards for Software Engineering Research [48]. As recommended in [49], instead of formulating an hypothesis, this experiment aims at answering our research questions.

The experiment consisted of an off-site and asynchronous screening test (Session 0) to evaluate the subjects prior knowledge and adequacy to participate in the experiment, and an on-site exercise with three parts (Sessions 1-3) and a duration of 3 hours. The protocol that we defined and applied is the following.

(1) *Session 0:* We provided two documents introducing the basic notions of uncertainty and the use of MagicDraw to the participants. After reading these documents, they were asked to fill in questionnaire Q0, in which they are asked about their level of knowledge about UML class and object diagrams, UML profiles, MagicDraw, uncertainty in general and uncertainty in domain and/or software models. The purpose of the introductory material is to ensure that all participants of the experiment have a minimum knowledge. A requirement for participation is to be familiar with UML and MagicDraw. Knowledge on uncertainty beyond that provided on the materials is not a requirement. Any mismatch between the prior knowledge and expected knowledge of the potential participants or failure to read the pre-experiment documents leads to the exclusion of the participants from the experiment.

(2) *Session 1 [50 min]:* In this session, we first provided an introduction to uncertainty, Subjective Logic and `SBoolean`, followed by a brief tutorial on UML profiles in MagicDraw. Both presentations were followed by a set of questions. Then, participants performed the first exercise, an example in practice of our Belief Uncertainty UML Profile. It uses the example of the Smart House (Section 2.1). Then, we ask them to fill in Questionnaire Q1. The goal of this questionnaire is to check whether participants are able to understand how opinions are represented in UML models, and their semantics. After the authors submitted the questionnaire, we solved and clarified the issues that led to errors to avoid problems in the next sessions.

(3) *Session 2 [60 min]:* We give users the second exercise. This is an individual hands-on exercise that describes a scenario in which each participant is supposed to be an historian seeking to find the *Ark of the Covenant.* Each participant is assigned randomly a role (out of three different roles) and they have to decorate a given model with their opinions. This model contains 6 instances, 8 attributes and 4 links. During the development of this exercise, we recorded the participants' screens. At the end, we asked them to fill in Questionnaire Q2 where we asked them about the ease of use and expressiveness of our Belief Uncertainty Profile to represent opinions.

(4) *Session 3 [70 min]:* In this session, groups of people are randomly created in such a way that each group contains three participants, each one with a different role. In the first part of this session, participants are asked to add their individual opinions on the fact that the Ark of the Covenant is located under Mount Nebo (i.e., their opinions on the existence of one instance in a model that we provided). Then, we ask them to work collaboratively and reach an agreement on whether they should dig in Mount Nebo or not—taking into account that there are consequences for both a good and a bad decision. Then, participants fill Questionnaire Q3A individually where they are asked about their experience (i.e., whether they were able to make a decision, their satisfaction with the decision, the difficulties they faced and the

| Number of participants with degree | | | |
|---|---|---|---|
| PhD | MSc | BSc | None |
| 4 (29%) | 5 (36%) | 4 (29%) | 1 (7%) |

(a) Participants' education.

| | Prior knowledge about | | | | |
|---|---|---|---|---|---|
| | UML diagrams | UML profiles | MagicDraw | Uncertainty | Uncertainty on models |
| Median | 4 | 2.5 | 3.5 | 2 | 1 |
| Mode | 4 | 4 | 4 | 3 | 0 |

(b) Participants' prior knowledge.

Fig. 14. Participants' profiles.

method they followed). We start the second part of this session with a presentation about the fusion operators followed by a round of questions. Then, we ask participants to use the fusion operators to make a decision. Note that we explain our iterative methodology to participants and ask them to put it in practice. During this session, we record the participants' screens as well as the audio of their discussions. Finally, we ask participants to individually respond to a questionnaire with two parts. The first one (Q3B) intended to report on their experience using the fusion operators. The second one (Q4) with questions about the whole experiment such as the perceived usefulness and easy of use of `SBoolean`, the Belief Uncertainty profile and the fusion operators, their feedback about our tool and process and suggestions for future improvements.

Seventeen people accepted our invitation to participate in the experiment and qualified for it. Although Nielsen and other authors maintain that five users are enough for usability testing [40, 61], other authors suggest the rule of 16 ± 4 participants [2]. We run a pilot with 3 participants and the experiment with 14 participants, with which we tried to cover both situations.

We carried out the pilot on November 9th and used it to refine our protocol, materials, instructions, proposed exercises and questionnaires. Then, we discarded its results. To accommodate to the participants agendas, we carried out the experiment on three different dates: November 11th in Málaga with 6 participants, November 18th in Barcelona with 5 participants, and November 26th in Málaga with 3 participants. The first author of the paper was always present and ensured that all the sessions were equally executed. The language used was either Spanish or English according to the participants' preferences.

## A.2 Results

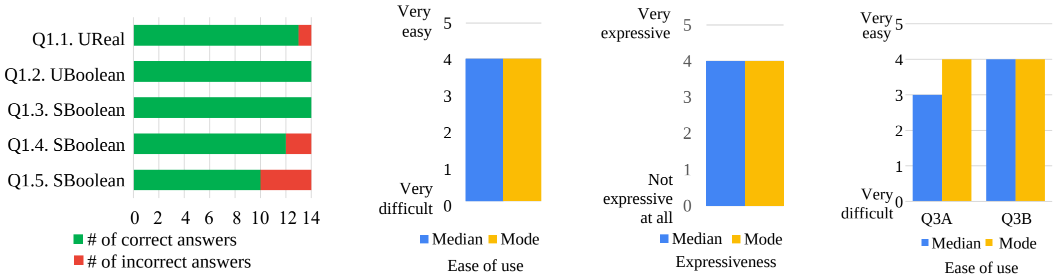
After validating the collected data, we confirmed that all the responses were valid, no outliers were detected and no entry had to be discarded.

Tables 14a and 14b present the degrees that our participants hold and their prior knowledge, respectively. When asking about their prior knowledge, we used a Likert scale where 0 is "Not at all" and 5 is "High". All the participants work in academia, either at the Open University of Catalonia or the University of Málaga.

The questionnaire Q1 was composed by 5 multiple-choice questions with 4 possible answers and only 1 correct. As presented in Figure 15a, we can observe that, out of the 14 participants, 13 responded correctly to the question about the semantics of an `UReal` attribute, 14 to the question about an `UBoolean` attribute, 14 the question about an `SBoolean` opinion, 12 one question about the degree of belief of an opinion, and 10 one question about the degree of uncertainty. Despite the limited previous knowledge of our participants about uncertainty on domain models, we can conclude that, a brief presentation was enough for most participants to understand our proposal (syntax and semantics).

In the questionnaire Q2, we asked our participants to rate "how difficult using the UML profile they think it is" using a Likert scale where 0 was "Very easy" and 5 "Very difficult". The responses

|  | Q2.1 Difficulty | Expressiveness |
|---|---|---|
| Median | 4 | 4 |
| Mode | 4 | 4 |



(a) Q1.1-Q1.5 (b) Q2.1 (c) Q2.2 (d) Q3A-B

Fig. 15. Results of questionnaires.

in Figure 15b show a median and mode of 1, meaning that they found our profile easy to use. We also asked them to rate "how expressive they think the UML Profile is for representing opinions in UML" using a Likert scale where 0 is "Not expressive at all" and 5 "Very expressive". In Figure 15c, we can observe that both the median and mode are 4, meaning that participants found our proposal expressive. The final screenshot that they uploaded as well as their screen recordings show that all of them were able to successfully complete the exercise.

When participants responded to questionnaire Q3A, they had made decisions on whether to dig or not without knowing about the fusion operators. All of them were able to reach a consensus and, using a Likert scale ranging from 1 "easy" to 4 "difficult", we asked them how difficult it was to make the decision. The median is 2 and the mode 1. Then, we asked them about the process they followed. Among others, the most popular processes that they reported were: "each member made a binary decision and we voted", "I was unsure so I let the others decide", "we discussed and each one tried to convinced the others". Interestingly, there was one group that calculated the average of each of the components of an SBoolean—one of the members said: "we calculated the average and concluded that we should dig with an SBoolean(0.6, 0, 0.4, 0.5)". This makes us realize that, somehow, these participants were trying to find an operation to merge their opinions.

In questionnaire Q3A (i.e., after using the fusion operators to make a decision), we also asked them to rate how easy or difficult making the decision was. This time both the median and mode are 1 (easy). The main difficulty that the participants found was how to choose the most appropriate fusion operator, which we already know is not an easy problem [28]. On the other hand, our participants found straightforward the use of our plugin to fuse opinions (one even said "we only need to click a button").

In Figure 16, we report on the responses of the participants to the Likert-scale final questions. We can observe that they are satisfied with our proposal. Furthermore, it is worth mentioning that the analysis of the participants' screen recordings and voice confirm that they successfully applied our methodology and our tool supported all the process (e.g., they defined their opinions, iteratively refined them and used the fusion operators to reach a consensus—the only part that was not part of this experiment was the creation of instance models due to the lack of time). Finally, in the open-ended questions, all the participants except one described our approach with positive adjectives such as interesting, novel, easy-to-use, intuitive, etc. One of them also pointed to the fact that all the concepts are a bit confusing at the beginning but once you get used to them, it is very useful. Only one participant complained about the fusion operators saying that he "considers the decision wiser when fully [...] debated orally and humanly weighted".

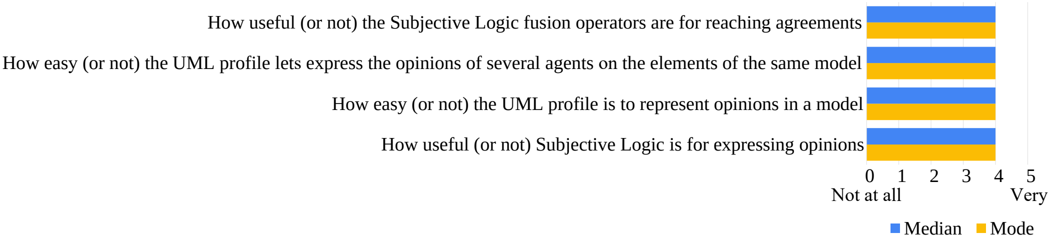


Fig. 16. Final Questions

## A.3 Lessons Learned

Most of the participants found our proposal useful, expressive and usable. We observed how the learning curve is steeper at the beginning, which is reflected in some in some incorrect answers to the questions of questionnaire Q1. However, as the experiment progressed and participants became familiar with all the concepts of our proposal, they were able to use it faster and correctly.

Participants expressed that selecting the most appropriate fusion operator was the hardest part. However, the results show that all of them were able to success in this task. We believe that, with some training, good documentation and/or examples of use, this aspect does not prevent the application of our proposal. Nevertheless, we plan to investigate how we can better assist users in this step in the future.

The fact that participants saw explicitly the opinions of the other members before and during their discussions towards reaching a consensus made them realise that their opinions could be refined. This shows that our iterative methodology was faithfully put in practice.