An Empirical Evaluation of the Use of Models to Improve the Understanding of Safety Compliance Needs

Jose Luis de la Vara[1], Beatriz Marín[2], Clara Ayora[3], and Giovanni Giachetti[4]

[1]Universidad de Castilla-La Mancha, Spain
joseluis.delavara@uclm.es

[2]Universidad Diego Portales, Chile
beatriz.marin@mail.udp.cl

[3] Tree Technology, Spain
claraayora@gmail.com

[4]Universidad Tecnológica de Chile INACAP, Chile
ggiachetti@inacap.cl

Abstract.
Context: Critical systems in application domains such as automotive, railway, aerospace, and healthcare are required to comply with safety standards. The understanding of the safety compliance needs specified in these standards can be difficult from their text. A possible solution is to use models.
Objective: We aim to evaluate the use of models to understand safety compliance needs.
Method: We have studied the effectiveness, efficiency, and perceived benefits in understanding these needs, with models and with the text of safety standards, by means of an experiment. The standards considered are DO-178C and EN 50128. We use SPEM-like diagrams to graphically represent the models.
Results: The mean effectiveness of 20 undergraduate students in understanding the needs and the mean efficiency were higher with models (22% and 38%, respectively), and the difference is statistically significant (p-value ≤ 0.02). Most of the students agreed upon the ease of understanding the structure of safety compliance needs with models when compared to the text, but on average, the students were undecided about whether the models are easy to understand or easier to understand than the text.
Conclusions: The results allow us to claim that the use of models can improve the understanding of safety compliance needs. Nonetheless, there seems to be room for improvement in relation to the perceived benefits. It must be noted that our conclusions may differ if the subjects were experienced practitioners.

Keywords: Experiment, understanding, comprehension, model, safety-critical system, safety standard.

1. Introduction

Safety-critical systems, including software-intensive ones, are those that can harm people, property, or the environment when some failure occurs [1], e.g. an aircraft, a car, a pacemaker, or a train. As a way to ensure that these systems do not pose unreasonable risk, they are required to comply with safety standards [2]. Examples of these standards include the generic IEC 61508 standard, DO-178C for avionics, ISO 26262 for automotive, and EN 50128 for railway. Third parties such as certification authorities and safety assessors usually evaluate these systems for safety certification.

Safety standards state specific criteria that must be fulfilled for compliance demonstration. These criteria include activities to execute, data to manage, requirements to fulfil, and information about when the elements should be considered. The criteria can be referred to as safety compliance needs [3]. System suppliers must show that safety compliance needs are met so that a system is allowed to operate. They need to understand the criteria and to follow them considering the requirements, data, activities, and additional information elements, but these tasks can be difficult.

Safety standards are textual documents that can consist of hundreds of pages and can define thousands of criteria for compliance [3]. Practitioners have acknowledged issues in understanding safety compliance needs and in applying the standards [4,5]; e.g. their text usually contains ambiguous and inconsistent fragments. Certification risks can arise as a consequence of these issues because a system supplier might misinterpret or miss some needs and thus develop a non-compliant system.

It is common that new approaches are proposed to facilitate safety compliance and certification [6,7]. Several authors, e.g. [8], advocate that the use of models can help practitioners understand safety compliance needs. However, there is scant or weak evidence available about the improvement in the understanding of safety compliance needs when using models. Prior work does not provide conclusive results [9], only provides insights from pilot studies [10], or is not based on actual model usage but on experts' perceptions [3,8]. Furthermore, there is a general lack of experiments on safety certification [2], i.e. of empirical studies that have compared the use different techniques for a same task in a controlled setting.

In order to fil the existing gaps, we have conducted an experiment to analyse the understanding of safety compliance needs with models, studying the effectiveness, efficiency, and perceived benefits. This aids in determining whether safety compliance needs could be more suitably represented in practice with models.

Twenty undergraduate students worked on the identification of safety compliance needs in EN 50128 and in DO-178C using models in the form of diagrams based on SPEM [11] and text of the standards. SPEM supports the specification of the activities, artefacts, roles, and techniques of a process, among other characteristics. In addition, the students expressed their opinion about their understanding of safety compliance needs with models and with the text of safety standards.

The mean effectiveness of understanding safety compliance needs was higher when using models of standards than when using their text (22% higher) as well as the mean efficiency

(38% higher). Furthermore, the difference is statistically significant (p-value ≤ 0.02). Most subjects found benefits in the use of models, especially to understand the relationships between the concepts of a standard. However, on average the subjects did not agree that the models were easy to understand or that the models were easier to understand than the text of standards.

We argue that this paper is the first study that reports a statistically significant improvement in the understanding of safety compliance needs when using models. We have previously run a pilot experiment [10] and an experiment [9], but the results were not conclusive enough and adjustments have been necessary in the experiment design to reduce and mitigate threats to validity, e.g. the notation used to represent safety standards (see Section 3).

The remainder of the paper is organised as follows. Section 2 presents the background of the paper. Section 3 reports the experiment process, and Section 4 the results. Finally, Section 5 summarises our main conclusions.

2. Background

We divide the background of the paper into safety standards, model-based specification of safety compliance needs, and related work.

2.1 Safety Standards

Safety standards correspond to industry-agreed best practices to guarantee that a system does not pose unreasonable risk, e.g. to guarantee that a system failure could not cause severe injury or death. As safety cannot be shown, the standards implicitly define how sufficient confidence in acceptable system safety can be developed. This includes practices for technical risk reduction, trust in the work conducted, and compliance management. Safety and compliance are usually assessed by third parties. Such an assessment can lead to safety certification, as a formal recognition of a system's acceptable safety for a given application and in a given context. For software systems, the standards deal with the necessary practices to suitably manage the safety requirements allocated to software. Two examples of these standards are DO-178C and EN 50128.

DO-178C (Software Considerations in Airborne Systems and Equipment Certification) [12] is the main safety standard for avionics and aerospace software. It provides guidance (1) to produce software for airborne products that performs its intended function with a level of confidence in safety, and (2) to determine in a suitable way that the software aspects comply with airworthiness requirements. DO-178C defines five software levels, Level A (highest) to Level E (lowest), that map to how catastrophic the consequences of a failure could be and establish the rigour necessary to demonstrate compliance. The practices in DO-178C also include:

- Objectives for the processes of the software lifecycle.
- Activities to satisfy the objectives.
- Descriptions of the software lifecycle data to provide as evidence of objectives' satisfaction.
- Variations according to the software level.
- Additional considerations for certain situations, e.g. software reuse.

EN 50128 (Railway applications - Communications, signalling and processing systems - Software for railway control and protection systems) [13] is the main safety standard for railway software. It provides requirements with which the development, deployment, and maintenance of safety-related software must comply. It also defines requirements concerning organisational structure, the relationship between organisations, and division of responsibility involved in the software lifecycle, including requirements on the qualification and expertise of personnel. EN 50128 defines five levels of software safety integrity, Level 0 (lowest) to Level 4 (highest), based on the risk resulting from software failure. The practices in EN 50128 also include:

- Techniques and measures for the five levels of software safety integrity.
- Descriptions of the techniques.
- The process of specifying the safety functions allocated to software.
- Means to develop, confirm, and manage these functions.
- Software lifecycle documentation.

2.2 Model-Based Specification of Safety Compliance Needs

In recent years, different authors have argued that the understanding of safety compliance needs, the management of safety compliance information, and thus safety certification can be facilitated by using models. This proposal is in line with the common use of graphical representations for certain safety assurance artefacts [14], e.g. safety analysis results and safety cases. Model-based solutions have been proposed to specify safety compliance needs for concrete safety standards or parts of them, e.g. for IEC 61508 [8] and for testing with DO-178B [15], and for specific compliance needs, e.g. for process- [16] and artefact-related [17] needs. Standards have also been published for model-based system assurance and certification [18]. Practitioners have reported the use of models for safety certification purposes [4,5].

We use a holistic generic metamodel for specification of safety compliance needs [3] for the experiment. An excerpt is shown in Figure 1. A concrete usage example of this metamodel (i.e. a model) employed in the experiment is presented below in Figure 2, Section 3. The different types of safety compliance needs can be specified with the metamodel, i.e. information about processes, artefacts, requirements, and their applicability for safety assurance. The metamodel has been validated with data from real projects, with practitioners, and with safety standards from several application domains [3]. All the parts of the standards were considered for validation. Both researchers and practitioners have successfully used the metamodel to represent parts of safety standards, e.g. in the scope of industrial case studies on system assurance and certification [19,20].

The metamodel supports the specification of safety compliance needs by means of:

- Reference requirements, which correspond to conditions to fulfil, e.g. consistency of software requirements and software testability;
- Reference activities, which represent units of behaviour to execute, e.g. software requirements process and software unit design;
- Reference roles, which are types of agents to be involved, e.g. developer and verifier;
- Reference artefacts, which are units of data to manage, e.g. software requirements data and software integration verification report;

- Reference techniques, which represent specific ways to execute a reference activity or to create a reference artefact, e.g. software modelling and model checking;
- Reference artefact relationships, which are relationships to record between two reference artefacts, e.g. the relationship 'includes', for software requirements data that includes high-level requirements, and the relationship 'based on', for software integration verification report based on software integration test specification, and;
- Reference artefact attributes, which correspond to characteristics of a reference artefact, e.g. the priority of a software requirement and the result of a test case.

All these classes are specialisations of Reference element. Relationships between the classes can also be specified, as well as with applicability information, e.g. about the use of a reference technique (recommended or mandatory) depending on how critical a function is (DAL A or SIL 3). Reference artefact, activity, technique, and role further specialise Constrained reference assurable element.

These elements are called "Reference" because they do not correspond to the actual assets managed in an assurance project, such as the specific artefacts created, but to the elements that safety standards indicate that a project will need to manage for compliance.

Further information about the metamodel can be found in [3], including its full version and more details about its elements, its usage, and its development and validation processes.

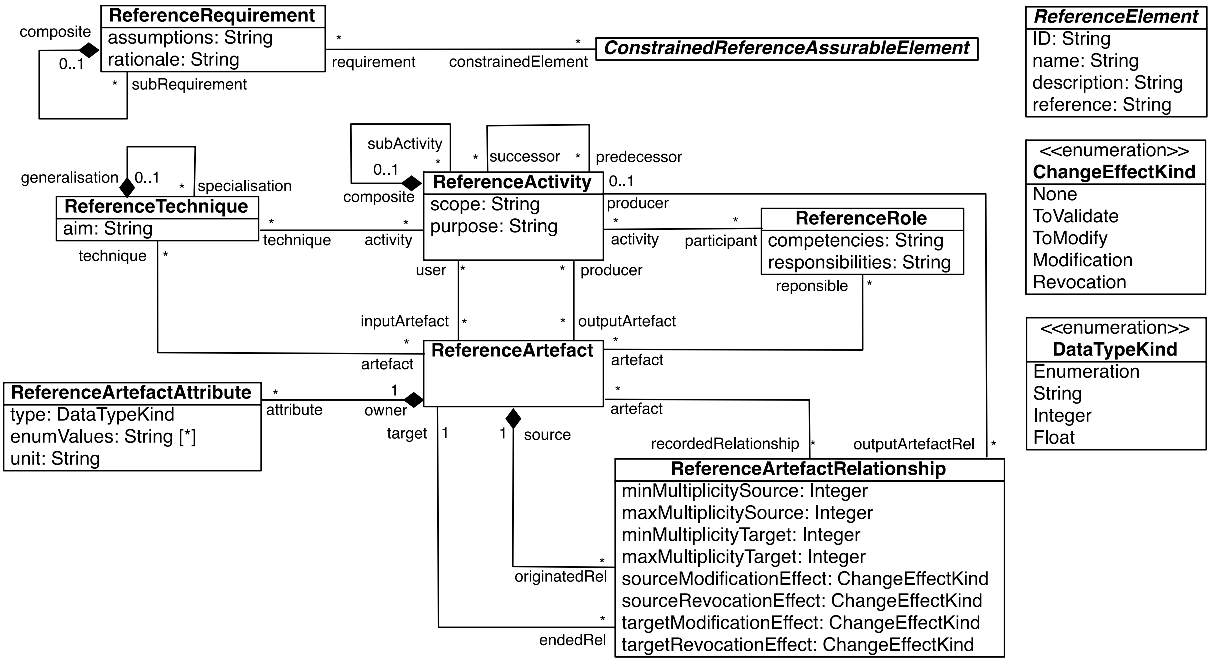


Figure 1. Metamodel to specify safety compliance needs (excerpt) [3]

2.3 Related Work

We started our experimental work on the use of models to improve the understanding of safety compliance needs with a **pilot experiment** [10]. The main purpose of this experiment was to validate our first design and to derive hypotheses. We analysed the results of understanding safety compliance needs with 15 undergraduate students and found both evidence and counterevidence that the use of models improves understanding. The overall conclusion was that the extent to which models help in understanding safety compliance needs seemed to be lower than what researchers expected.

Afterwards, we conducted **an experiment** [9] with 16 undergraduate students to compare the understanding of safety compliance needs with models specified as UML object diagrams and with text of safety standards. The mean effectiveness was higher with models (17%) and the mean efficiency too (15%). However, the differences lacked statistical significance. Despite the benefits found in the use of models, on average the subjects were undecided about the ease of understanding. A limitation in this experiment was model representation with UML object diagrams because all the objects are graphically represented in the same way, thus objects of different classes can be more difficult to distinguish. We have addressed this limitation in the experiment reported in this paper by changing the notation.

Table 1 summarises the different characteristics of the experiments that we have conducted. The content of the models and the text of safety standards used are the same for all the experiments. The subjects correspond to students of the same course but in different years. No subject has participated in several experiments. All the changes have aimed to address possible issues in the experiment design, e.g. adjustments of the questions for which we found that some misinterpretation might have occurred. The largest change has arguably been from the experiment reported in this paper to the previous one. The material has been revised to change the notation of the models and the questionnaire about perceived benefits.

Except our previous work, we are only aware of publications that have analysed the improvement in the understanding of safety compliance needs according to **experts' opinion**. On an IEC 61508 model [8] and with a sample of 12 practitioners, most of the subjects regarded the model as easy to understand and expressed their interest in using the model to understand the standard. At a training session on the holistic generic metamodel [3], four practitioners acknowledged benefits in the use of models to understand safety compliance needs. Among the different aspects of safety standards, the understanding of the concepts of the standards and the understanding of the relationships between the concepts were regarded as largely improved. These publications have contributed to claiming that the understanding of safety compliance needs could be improved with models. Nonetheless, further evidence is necessary. It is also required that the evidence is gathered from experiments that compare the use of models for understanding of safety compliance needs and the use of the text of standards to be able to state a cause-effect relationship.

The number of **experiments on safety certification-targeted activities** is arguably low [1,2,21]. Among the experiments on approaches that exploit models, Briand et al. [22] analysed the use of a traceability approach based on SysML for safety inspections. The results show that decision correctness increases when using the approach. Textual use cases and system diagrams have been compared in relation to their support for identification of different types of hazards by non-experts [23]. The results show that, in most of the cases, hazard identification can be as good or better with the use cases. The authors concluded that both the representation and how the information is brought into focus matter so that no hazard is missed. Abdulkhaleq and Wagner [24] compared three safety analysis techniques: fault tree analysis, failure models and effects analysis, and systems-theoretic accident model and processes. According to the results, the difference between the techniques is not statistically significant regarding understandability, applicability, and ease of use. However, the difference of effectiveness and efficiency is significant. Jung et al. [25] studied a model-based safety analysis approach (component integrated fault trees) with domain experts. The use of models did not result in a number of correct or incorrect solutions that was significantly different, but the subjects considered that the modelling capacities were better in terms of clarity, consistency, and maintainability. Gonschorek et al. [26] have also concluded that Component Fault Trees can be more comprehensible than Fault Trees. When comparing state event fault tree analysis vs. dynamic fault tree analysis and fault tree analysis combined with Markov chains analysis [27], the first was rated as more applicable and the second as more efficient. Cyra and Gorski [28] studied the consistency and accuracy of an approach for argument assessment based on aggregation rules. This study shows that result accuracy and consistency are more similar when applying the rules. Finally, other experiments have evaluated the automatic and manual generation of assurance cases [29].

Table 1. Characteristics of the series of experiments conducted on the use of models to improve the understanding of safety compliance needs

	Pilot experiment [10]	**First experiment [9]**	**Second experiment (this paper)**
Purpose	Validate the design	Test hypotheses	Test hypotheses
Experiment type	Between-subject	Within-subject 2x2 factorial	Within-subject 2x2 factorial
Notation	UML object diagram	UML object diagram	SPEM-like notation and tables
Questionnaire about safety compliance needs	6 questions for each task	7 question for each task; 5 maintained and 2 new for both DO-178C and EN 50128	7 questions for each task; 6 maintained and 1 new for both DO-178C and EN 50128
Number of safety compliance needs to identify	10	11	11
Questionnaire about perceived benefits	Only about models (1 questionnaire)	Only about models (1)	About both models and text (2)
Other changes in the material	N.A.	Adjustment of diagram layout, separate sheets to indicate the start time and the end time	Some minor changes in the wording of the introduction
Number of subjects	15	16	20
Results analysed	1 task per subject	2 tasks per subject	2 tasks per subject
Result summary	Effectiveness slightly better with models, efficiency better with text, both advantages and disadvantages found in model usage	Effectiveness and efficiency better with models, both advantages and disadvantages found in model usage	Effectiveness and efficiency better with models, both advantages and disadvantages found in model usage
Result significance	N.A.	No	Yes

Many publications have investigated the **comprehension of model-based artefacts** in controlled empirical settings. When comparing UML class diagrams and ER ones [30], a better comprehension was achieved using UML. A family of experiments on requirements comprehension with UML sequence diagrams [31] provides evidence that the comprehension is significantly improved when sequence diagrams are used. Another experiment [32] provides more powerful evidence that dynamic modelling with sequence diagrams facilitates requirements comprehension. Cruz-Lemus et al. [33] studied the understandability of UML state charts with composite states. Based on the results, it is not clear that composite states improve diagram understanding. Lange at al. [34] evaluated the understanding of UML models in comparison with a more generic view to understand the interaction among the UML models and to navigate through them. They provided evidence of a better comprehension of the UML model when this generic view is used. The inclusion of object diagrams as a complement of class diagrams does not always lead to significant benefits in terms of design comprehensibility [35]. The style in variability modelling impacts model comprehension [36].

Experiments on the comprehension of software code with UML analysis models [37] suggest that the comprehension is not improved when using models. Nonetheless, a higher level of detail in the models seems to ease the understanding of a system [38,39]. The study of the inclusion of stereotypes in UML diagrams [40,41] has shown that it improves model comprehension. Extensions to and adaptations of the i* requirements language improve its comprehension [42,43,44]. When compared to other languages, i* can be more understandable than KAOS [45], but not always than use cases [46,47]. Other recent pieces of work on i* have focused on factors such as following layout guidelines [48] and gender differences [49]. SysML requirements diagrams were evaluated in a controlled experiment [50] and the result was that they improve requirements comprehension and increase the level of confidence in comprehension.

Regarding experiments on the **comparison of textual and graphical representations**, Razali et al. [51] compared formal UML specification with textual representation and showed that UML expedites system comprehension. In a comparison of textual and graphical representation of Tropos regarding the efficiency in requirements comprehension [52], the subjects stated that they mostly preferred the graphical representation, but they were more efficient using the textual representation. Rodrigues et al. [53] performed a comparison of textual representations of business processes and BPMN models. The results show that both representations appear to be similar for process understanding although understanding performance is better when using BPMN. The use of business process models for user stories appears to significantly lead to a better understanding [54,55], as well as the use of models instead of textual descriptions to understand business processes in some cases [56]. In contrast to these results, Sharafi et al. [52] report that the use of requirements models instead of text did not significantly increase comprehension accuracy. A similar result has been reported for software architecture regarding the communication of design decisions [57]. The results of the subjects that used text predominantly were even better.

Some related experiments have been conducted recently on **security engineering**, which can be regarded as close to safety engineering. Labunets et al. compared security risk assessment with graphical methods and with textual ones [58,59]. The overall effectiveness was similar, but models were more effective for some aspects. The subjects' perceived usefulness and intention to use were higher with the graphical method. When comparing diagrams and tabular representations [60-63], the techniques yielded a similar level of understanding and of perceived effectiveness, but tabular representations seem to be better, especially for simple comprehension tasks. Nonetheless, similar results can also be obtained with the different representations [64]. The use of textual labels also improves the understanding over iconic graphical models.

In summary, the results of related work show that comprehension can be improved with the use of models, but not in all cases. We aim to complement these results by providing new insights in relation to the understanding of safety compliance needs.

3. Experiment Process

We followed the guidelines suggested by Wohlin et al. [65] to conduct the experiment.

The goal is to analyse the use of models to specify safety compliance needs for the purpose of evaluation with respect to effectiveness, efficiency, and perceived benefits of understanding safety compliance needs from the point of view of the researcher in the context of undergraduate students in Computer Science and Engineering.

We formulated three research questions (RQs):

- RQ1. Does the use of models increase the effectiveness of understanding safety compliance needs?
- RQ2. Does the use of models increase the efficiency of understanding safety compliance needs?
- RQ3. Do users find benefits in the use of models to understand safety compliance needs?

We present the planning, operation, and main threats to the validity of the experiment presented in this paper in the following subsections. The planning section includes the information about how the design of the experiment realises the study of the above goal and RQs in a specific way, e.g. about the selection and use of SPEM-like diagrams to create models of safety standards.

3.1 Experiment Planning

The experiment context is a 3rd-year undergraduate course on “Software development projects management” in Computer Science and Engineering at Carlos III University of Madrid, Spain. The students of this course are the subjects. In the course, the students must plan the development and the validation of a software system and design the system following a specific software engineering standard: ESA PSS-05-0 and its associated guides [66]. We regard these students as suitable subjects because of their exposure to having to follow and comply with a safety standard. The course language is English. The percentage of international students is usually around 25%. The Spanish students take the entire degree in English and need to fulfil English-level requirements to be allowed to do so. They must prove a B2 level or above in the first year.

The students have to find safety compliance needs from excerpts of the text of safety standards and from models of the excerpts. They also have to express their opinion about the use of the models.

We formulate three null hypotheses that we aim to test:

1. $H_{1,0}$: There is no statistically significant difference in the effectiveness of understanding safety compliance needs with the text of safety standards and with models.
2. $H_{2,0}$: There is no statistically significant difference in the efficiency of understanding safety compliance needs with the text of safety standards and with models.
3. $H_{3,0}$: There is no statistically significant difference in the perceived benefits of understanding safety compliance needs with the text of safety standards and with models.

The reference p-value to test the hypotheses is 0.05.

We use two independent variables for the experiment:

1. The means used to represent safety compliance needs (text of a safety standard or model).
2. The safety standard considered (EN 50128 integration process or DO-178C requirements process; these standards are different to the one used in the course).

We select these processes because the students worked on similar activities in the course, i.e. equivalent processes in ESA PSS-05-0. We use SPEM-like diagrams to create the models of the safety standards. We use SPEM symbols [11] for Reference artefacts, Reference activities, Reference techniques, and Reference roles. This representation format is extended with the usage of GSN [67] for Reference requirements and of UML composition [68] to represent the parts of Reference elements. SPEM is a standard for modelling software and system processes that prior work has used to represent safety compliance needs (e.g. [69]), as well as industry (e.g. [70]). Other process notations such as BPMN or UML activity diagrams are less suitable because they do not support the specification of e.g. the techniques used in an activity or several participant roles in an activity, thus their use would require a larger extension. GSN is arguably the main notation for safety argumentation [2], and UML is the de-facto software modelling standard [71]. Based on the insights from our previous experiments [9,10], applicability information is presented in tables for clarity. To consider all the types of safety compliance needs [3] (see Section 2.2), we use a process model, an artefact model, a requirements model, and an applicability model. The model elements are tagged with their type to more easily identify it. Prior work has also shown that providing this kind of information, e.g. through stereotypes, can aid in model comprehension (see Section 2.3).

The effectiveness and the efficiency are two dependent variables. The effectiveness allows us to determine the degree to which the subjects are successful in identifying safety compliance needs, whereas the efficiency allows us to determine the degree to which they are successful in relation to the time spent for safety compliance need identification.

We use the F-measure to calculate the effectiveness, as commonly done in software engineering experiments, e.g. [31]. The F-measure is based on the precision and the recall in the identification of safety compliance needs. We use formulas for experiments in which a subject might not answer a question:

$$precision_s = \frac{\sum_i |answer_{s,i} \cap correct_i|}{\sum_i |answer_{s,i}|}$$

$$recall_s = \frac{\sum_i |answer_{s,i} \cap correct_i|}{\sum_i |correct_i|}$$

$$F_s = 2 \times \frac{precision_s \times recall_s}{precision_s + recall_s}$$

$answer_{s,i}$ corresponds to the number of answers that a subject provides and $correct_i$ to the number of correct answers that the subject should provide. Precision concerns the correctness of the responses provided and recall concerns their completeness. The F-measure concerns the balance between correctness and completeness.

As in related studies, e.g. [51], we combine the effectiveness (F-measure) and the time (in minutes) to calculate the efficiency:

$$Effy_s = 100 \times \frac{F_s}{minutes}$$

We consider that both effectiveness and efficiency are relevant from a practical point of view. In addition to correctly and completely understanding safety compliance needs, it is necessary to do it in an acceptable time. Otherwise, a representation technique might not be suitable.

The perceived benefits in understanding safety compliance needs is the third dependent variable. This variable is assessed with a questionnaire that presents statements about the use of the text of safety standards and the use of models to specify and to understand the needs. The questionnaire is based on existing ones [3,9,10]. Unlike in our previous experiments, we asked both about the text and about the models to be able to compare opinions.

The subjects have to complete a questionnaire (object) about safety compliance needs in a model and in a text excerpt (two different tasks). The subjects are randomly divided into four groups in a within-subject 2x2 factorial design [72]:

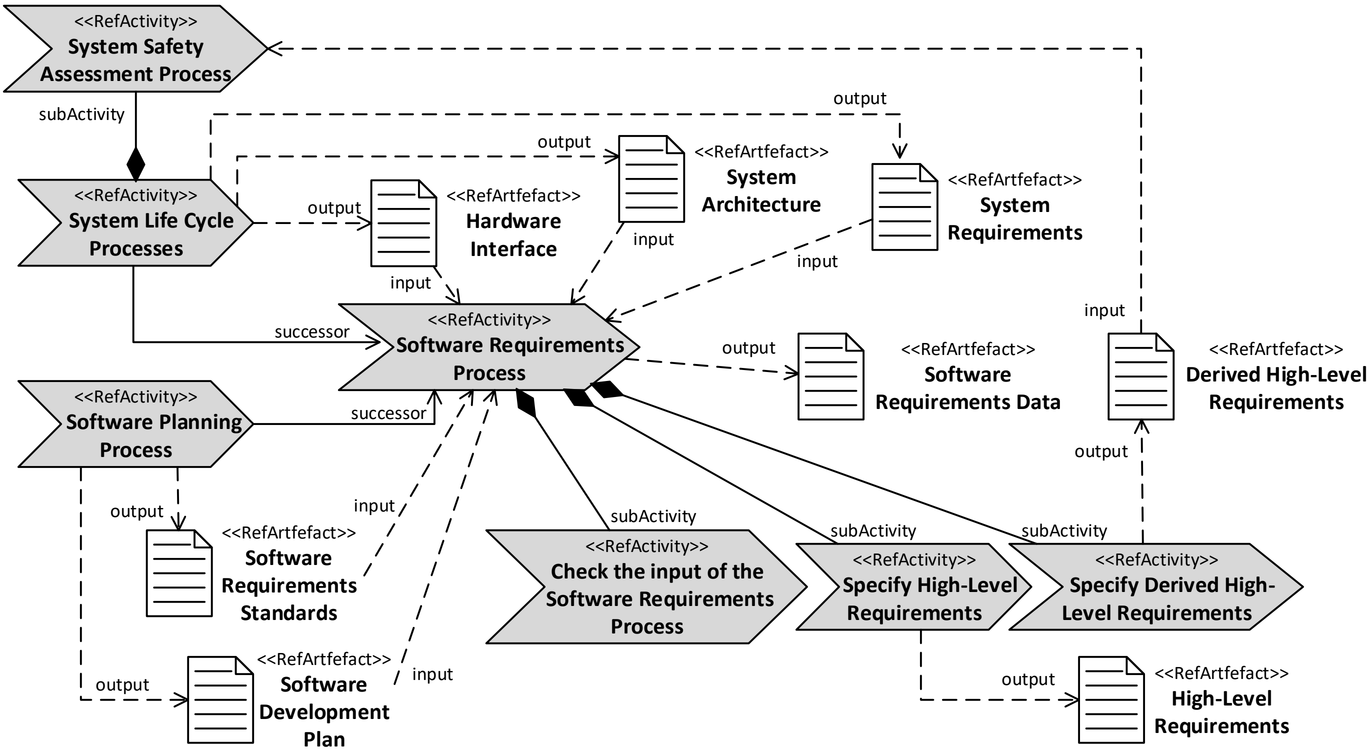
1. DO-178C model (for the first task) and EN 50128 text (for the second task).
2. EN 50128 model and DO-178C text.
3. DO-178C text and EN 50128 model.
4. EN 50128 text and DO-178C model.

Each subject participates in only one group.

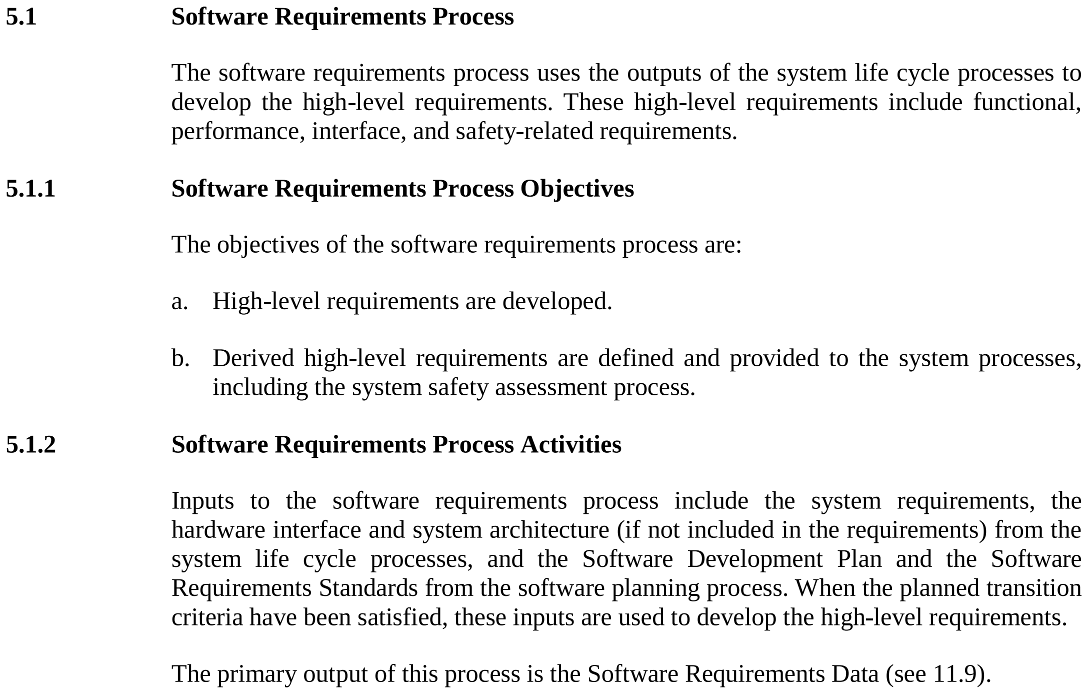
The material for the tasks includes an introductory page, a two-page excerpt of a standard or models that represent the excerpt, and seven free-text questions, e.g. “What information should the High-Level Requirements conform to?” for the DO-178C standard. The material is provided in paper. The subjects should find 11 safety compliance needs to complete the questionnaire correctly; the same in the model and in the text excerpt, e.g. Software Requirements Standards for the question above. When providing their answers, the subjects can refer to more or less than, or exactly to, the 11 needs. This is considered for calculating precision.

The material of the experiment is available online[1]. Figure 2 includes excerpts of the experimental material. It must be noted that the model in Figure 2 (a) is not a representation of only the text in Figure 2 (b) but a complete process model for the requirements process of

[1] https://sites.google.com/site/jldelavara/material/msac2017-uc3m



(a)



5.1 **Software Requirements Process**

The software requirements process uses the outputs of the system life cycle processes to develop the high-level requirements. These high-level requirements include functional, performance, interface, and safety-related requirements.

5.1.1 **Software Requirements Process Objectives**

The objectives of the software requirements process are:

a. High-level requirements are developed.

b. Derived high-level requirements are defined and provided to the system processes, including the system safety assessment process.

5.1.2 **Software Requirements Process Activities**

Inputs to the software requirements process include the system requirements, the hardware interface and system architecture (if not included in the requirements) from the system life cycle processes, and the Software Development Plan and the Software Requirements Standards from the software planning process. When the planned transition criteria have been satisfied, these inputs are used to develop the high-level requirements.

The primary output of this process is the Software Requirements Data (see 11.9).

Figure 2. Excerpts of the experimental material for DO-178C: (a) model and (b) text

DO-178C. In addition, Figure 2 (a) includes information that is not represented in Figure 2 (b) but in other models of the material.

The subjects receive the material for the second task when they finish the first one. For each task, they complete a questionnaire about the ease of understanding safety compliance needs with the text or with the models of safety standards, depending on the task performed.

The statements are the same for both techniques. The second questionnaire has an additional statement about whether models of safety standards are easier to understand than the text. The subjects must also record the time when they start and when they finish each task. They can provide comments at the end of the questionnaires.

Experiment execution is expected to require a maximum of two hours; an hour for training and an hour to perform the tasks. Nonetheless, the subjects can spend the time that they need to complete the tasks. The first author was the main researcher responsible for material preparation because, although the rest of authors had also worked on the topic, he is the author with the widest experience in safety certification. The rest of authors validated the material and, as a result, some minor adjustments were made, e.g. in question wording and the appearance of the diagrams. A presentation on safety assurance for critical systems and on the holistic generic metamodel is used for training, including the semantics of the metamodel elements. The presentation includes models of ESA PSS-05-0 to ensure homogeneous knowledge.

3.2 Experiment Operation

Twenty students participated in the experiment. We ran it in May 2017, the last week of the second semester. These students are different to the subjects of our previous experiments.

The training duration and the average task completion time were close to our plan. At the end of the training, and before performing the tasks, we told the students that the tasks targeted research purposes and that their performance would not impact their course grade. Nonetheless, we explicitly asked the students to do their best and to perform the tasks in an exam-like manner, e.g. without asking other students.

For validation, we checked the data after experiment execution. We did not discard the results from any subject because we did not find any clear indicator of careless response. As further explained in Section 4, we used the Shapiro-Wilk test [73] for normality, analysed statistical significance with the paired t-test [65] for normally distributed samples and with the Wilcoxon test [65] for non-normally distributed ones, and calculated the effect size with Cohen’s d [74] for normally distributed samples and with Cliff’s d [75] for non-normally distributed ones. These tests are suggested in the literature for experiment designs similar to ours [65,72,76,77], and as such have been used, e.g. Ricca et al. [78].

3.3 Validity

Although we planned and executed the experiment carefully, some threats could impact it. We discuss the main ones in this section according to the classification by Wohlin et al. [65]. We also discuss some further aspects in Section 4.

An important aspect of **internal validity** is maturation, as subjects react differently when time passes. Two hours might be a long time for students to participate in an experiment, so a fatigue effect could appear. Nonetheless, the students’ classes last this long. The threat of learning between the first and the second task is reduced by using different standards, different parts of them, and different representation formats for safety compliance needs for each task. Having a break between the training and the tasks and executing the experiment in the morning also mitigated fatigue threats. Reproducibility positively impacts internal validity. In our case, we have run a series of experiments on the understanding of safety compliance needs with models, reproducing the initial experiments after adjusting their design. Making the experimental material publicly available also contributes to reproducibility.

Using different experimental objects in the two tasks mitigated learning effects. However, when validating the data, we realised that most of the subjects (16; 80%) finished the second task faster than the first. We considered that the impact of this threat was not high because of using a within-subject 2x2 factorial design. For confirmation, we run an ANOVA test on the time differences of each group and the results are not statistically significant. The threat seems to affect all the tasks similarly. The size of the experimental material and its complexity can affect experiment results. We use excerpts of only two safety standards to mitigate this threat. We further validated and adjusted the material from the insights gained from [9,10] to reduce its complexity, e.g. by representing applicability information in tables. Nonetheless, the information represented in the material and thus its realism remained identical and suitable. The models still corresponded to semantically equivalent representations whose content had been validated by practitioners.

Although there is a risk of evaluator bias, we reduced it by avoiding telling the subjects the specific goals, research questions, and hypotheses of the study, and that we were among the authors of the metamodel. We also used objective quantitative metrics to measure the dependent variables to reduce this threat. Finally, and as in all questionnaires, there is an inherent threat in the way that the statements about the perceived benefits are formulated.

External validity is related to result generalization. Using students as subjects affects it, as it could be better to use practitioners. Nonetheless, the use of students in experiments on systems and software engineering is a common practice and has been regarded as suitable by the experts on empirical research for over two decades, e.g. [79-82]. According to recent studies, there can be minor differences when practitioners or subjects from academia are used [83]. Using students is valid to advance theories and technologies [84], and students can be considered to be equivalent to novice practitioners [31]. The available evidence further suggests that experience does not greatly help practitioners to improve the understanding of safety compliance needs [4]. Although we cannot claim that the results would be the same with practitioners, it cannot be claimed either that the results would be different.

Since the sample size and the number of students of the course were two important restrictions, we adopted a within-subject 2x2 factorial design to obtain more data. We are also working on a family of follow-up experiments in Spain and Chile to address sample size threats. On the other hand, getting a larger sample that is at least as valid is not easy. It is not usual to find a course and a set of students that have had to follow and comply with a safety standard during the whole course. Recruiting a large sample of suitable practitioners is very difficult as well.

Finally, the DO-178C standard and standards similar to EN 50128, i.e. the IEC 61508 standard and derived ones, seem to be the safety standards that are most frequently used in the industry [4,5]. Their use in the experiment contributes to external validity.

Construct validity refers to the link of the concrete experimental elements and the experimental goal. The interpretation of the safety standards might threat the creation of the experimental material, i.e. the creation of the models that represent excerpts of the standards. We used excerpts of standards for which we had access to models that practitioners had validated to mitigate this threat. Threats from using a specific opinion questionnaire (e.g. misinterpretation) could also appear. We mitigated them by using a questionnaire based on existing ones that experts had validated (practitioners and researchers).

We told the subjects that their performance in the experiment would not affect their course grade to reduce the threat of evaluation apprehension. Reusing material from our previous experiments also contributes to construct validity. We performed a training session to all the participants to mitigate the threat of inadequate pre-operational explication of the material. The interaction of subjects with different treatments is a threat because a confounding effect could appear.

Conclusion validity concerns the ability to draw conclusions that are correct. The use of dependent variables that are widely used in similar experiments, e.g. [22,31], contributes to reliability of measures. Although the random heterogeneity of participants is a threat that might affect our conclusions, this threat is usually reduced when using students with the same or a similar background [65], as in our experiment. We consider that threats from unbalanced groups are mitigated by the 2x2 factorial design and the subjects' similar background. Reproducing our study through a series of experiments contributes to conclusion validity.

The use of tests to analyse the results and to determine their statistical significance and practical importance contributes to conclusion validity. We use parametric tests and non-parametric ones depending on the normality of data, and for the p-value a 0.05 level. Regarding the selection of the population, we randomly created the groups according to the order of the students in the classroom. As in all experiments on the use of models, the use of a given notation impacts conclusion validity. The sample size can also threaten conclusion validity.

4. Results and Interpretation

We present the results of the experiment and interpret them in this section. There is a subsection for each RQ. We further discuss the results in relation to insights presented in prior work.

Regarding the background of the subjects, they did not have knowledge about the standards or the parts of them used in the experiment. They had not been involved in the development of any real safety-critical system and their experience with having to follow standards for the development of safety-critical systems was limited to the course. Their experience with systems or software modelling (e.g. with UML), either in a previous course (Figure 3) or in real projects (Figure 4), was quite homogeneous in our opinion, in addition to very similar to what we expected for 3rd-year undergraduate students studying Computer Science and Engineering. For example, only two students (10%) had not attended any course in which they had to deal with systems or software modelling and no student had a large experience in modelling in practice. Many students (9; 45%) had attended more than one course and several (7; 35%) already had experience with modelling in real projects. Twenty-five percent of the subjects corresponded to international students.

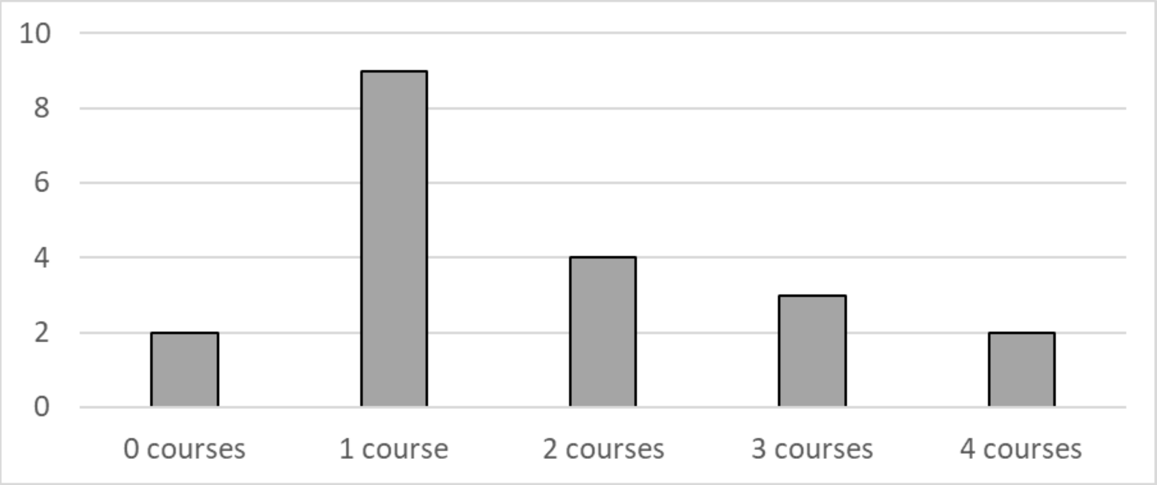


Figure 3. Number of courses in which the subjects had dealt with systems or software modelling

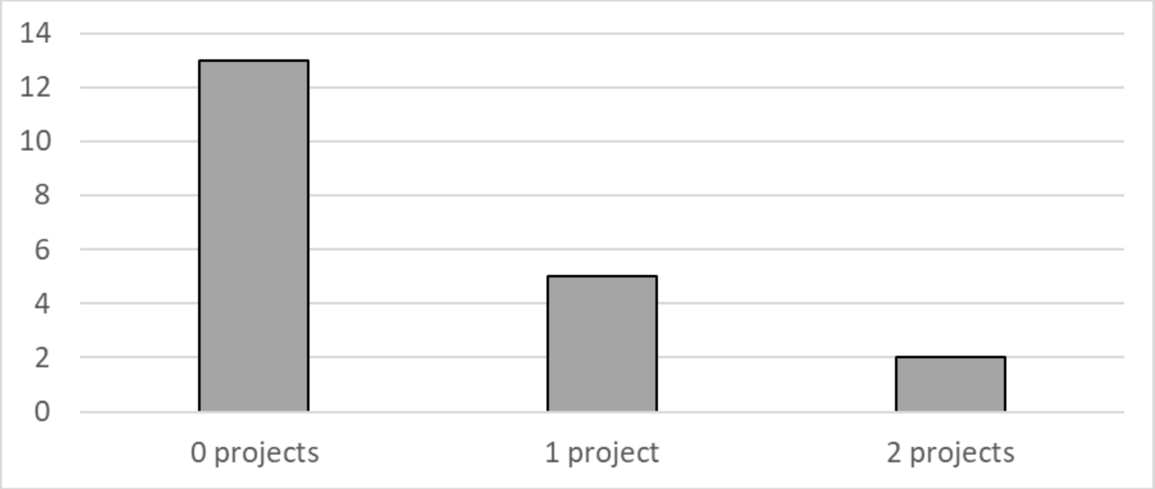


Figure 4. Number of real projects in which the subjects had used systems or software modelling

We do not conduct specific analyses about the possible impact of the experience in courses or in real projects because (1) we consider that the general experience was not heterogeneous, and (2) when checking the results, we did not find any clear indicator of the impact. For example, the students with the largest experience in courses on modelling performed close to the mean or worse than other students with less experience.

4.1 Effectiveness of Understanding (RQ1)

The results about the effectiveness of understanding safety compliance needs are shown in Table 2. The table includes the data about the precision (P), the recall (R), and the F-measure (F) of each subject of each group, considering all the questions as a whole; e.g. the F-measure of a given subject for all the questions. It also shows the mean, the median, and the standard deviation of the set of subjects. The mean values of the metrics are similar to or higher than the mean values in other experiments on safety certification activities, e.g. [22,24]. Therefore, we consider that the subjects' overall effectiveness is acceptable and valid.

In total, the number of non-answered questions is only four (1.4%), covering four different questions. A subject did not answer two questions (one about the DO-178C model and another about the EN 50128 text) and other two subjects did not answer one question (one about the DO-178C text and one about the EN 50128 text, respectively). We consider that this does not have any relevant impact on the results.

The mean effectiveness when using models of safety standards is 22% higher than when using the text of the standards; the median is 26% higher. These initial overall results suggest that the use of models increases the effectiveness of understanding safety compliance needs. In addition, the effectiveness was higher with models for 16 subjects (80%). The highest effectiveness was obtained with models (0.96). The effectiveness was higher than or equal to 0.8 for 12 subjects with models (60%) and only for one with the text (5%). The lowest effectiveness was obtained with the text (0.26), and the effectiveness was lower than or equal to 0.6 for two subjects with models (10%) and for seven with the text (35%).

Table 2. Effectiveness of understanding safety compliance needs with models and with the text of safety standards

		Models			**Text**		
Group	**Subject**	P	R	**F**	P	R	**F**
1	**1**	0.71	0.91	**0.8**	0.64	0.82	**0.72**
	2	0.39	0.64	**0.48**	0.57	0.73	**0.64**
	3	0.67	0.73	**0.7**	0.75	0.55	**0.63**
	4	0.5	0.73	**0.6**	0.67	0.73	**0.7**
	5	0.79	1	**0.88**	0.75	0.82	**0.78**
2	**6**	0.75	0.82	**0.78**	0.38	0.55	**0.44**
	7	0.85	1	**0.92**	0.5	0.73	**0.59**
	8	0.85	0.82	**0.82**	0.75	0.82	**0.78**
	9	0.83	0.91	**0.87**	0.35	0.55	**0.42**
	10	0.75	0.73	**0.73**	0.75	0.82	**0.78**
3	**11**	0.92	1	**0.96**	0.71	0.91	**0.8**
	12	0.91	0.91	**0.91**	0.67	0.91	**0.77**
	13	0.89	0.73	**0.8**	0.67	0.91	**0.77**
	14	0.9	0.82	**0.86**	0.58	0.64	**0.61**
	15	0.89	0.73	**0.8**	0.4	0.55	**0.46**
4	**16**	0.58	0.64	**0.61**	0.25	0.27	**0.26**
	17	0.71	0.91	**0.8**	0.55	0.55	**0.55**
	18	0.5	0.55	**0.52**	0.54	0.64	**0.58**
	19	0.77	0.91	**0.83**	0.58	0.64	**0.61**
	20	0.6	0.82	**0.69**	0.7	0.64	**0.67**
	Mean	0.74	0.81	0.77	0.59	0.69	0.63
	Median	0.76	0.82	0.8	0.61	0.68	0.64
	Std. dev.	0.15	0.13	0.13	0.15	0.16	0.15

The samples for effectiveness are normal according to the Shapiro-Wilk test (p-value > 0.05); thus, we selected the paired t-test for $H_{1,0}$. The test result indicates that the difference in the effectiveness is statistically significant (p-value = 0.0009 < 0.05). As consequence, $H_{1,0}$ can be rejected and the results allow us to claim that the use of models can increase the effectiveness of understanding safety compliance needs. The effect size (Cohen's d) for the effectiveness can be regarded as large (d = 1.00 > 0.8), which contributes to the practical importance of the results.

When having a closer look at the results considering the total 14 questions asked, the subjects made more errors in 11 questions (79%) with the text than with the models. For most of these questions about related pieces of information, e.g. about the input Reference artefacts of a given Reference activity, we have identified large differences in the number of errors. We also checked if receiving first the models or the text of the safety standards influenced the effectiveness results, e.g. if the subjects that received the models first performed better than those who received the models in the second task, when having to understand safety compliance needs with the models. The difference is small on average (0.02 with the models and 0.04 with the text; 3% and 7%, respectively) and is not statistically significant (t-test).

The effectiveness was also higher with models in our previous experiments [9,10], but not so high and the difference lacked statistical significance. A possible explanation can be the change in the notation to represent safety compliance needs (SPEM-like diagrams instead of UML object diagrams), which aimed to make the different types of Reference elements easier to distinguish. The representation of applicability information in tables can also have contributed to increasing effectiveness. Indeed, the amount of errors that the subjects made with models when asked about applicability information has considerably decreased. Another possible reason for the difference in the results is that the sample is different and larger. We will analyse these possibilities in the future in experiment replications.

Almost all the available evidence in prior work shows that using models increases effectiveness of understanding or that the difference with the use of text is not significant (see Section 2.3). Although this outcome depends on the specific notations and their usage purpose, e.g. security risk assessment [59] or requirements engineering [52], the results from the experiment are in general in line with prior work and further provide statistically significant evidence of the benefits of using models to understand safety compliance needs as a specific purpose.

As a main overall conclusion, we argue that the use of models can increase the effectiveness of understanding safety compliance needs. This conclusion is supported by almost all the evidence collected in the experiment, and in our previous experiments despite the lack of statistically significant results. We also argue that the change in the notation from our previous experiments have influenced the results.

4.2 Efficiency of Understanding (RQ2)

The results about the efficiency of understanding safety compliance needs are presented in Table 3. The table shows the time spent on each task (T; in minutes) and the efficiency outcome (Effy) of each subject of each group. It also shows the mean, the median, and the standard deviation of the set of subjects.

Table 3. Efficiency of understanding safety compliance needs with models and with the text of safety standards

		Models		**Text**	
Group	**Subject**	T	**Effy**	T	**Effy**
1	**1**	23.25	**3.44**	21.67	**3.32**
	2	16	**3.02**	22	**2.91**
	3	18.4	**3.78**	18.03	**3.50**
	4	18.6	**3.19**	12.7	**5.46**
	5	23.5	**3.74**	16.75	**4.67**
2	**6**	11.72	**6.68**	17.18	**2.59**
	7	24.3	**3.77**	12.87	**4.61**
	8	27.3	**3**	12.3	**6.36**
	9	16.13	**5.39**	18.77	**2.28**
	10	23.95	**3.04**	18.08	**4.33**
3	**11**	17.58	**5.44**	21.07	**3.8**
	12	12.6	**7.22**	21.57	**3.57**
	13	15.32	**5.22**	19.83	**3.88**
	14	15.37	**5.58**	10.65	**5.72**
	15	8	**10**	14.8	**3.12**
4	**16**	18.62	**3.27**	27.2	**0.96**
	17	13.23	**6.05**	29.87	**1.83**
	18	12.43	**4.12**	27.2	**2.14**
	19	13.8	**6.04**	22.35	**2.72**
	20	10.55	**6.56**	17.23	**3.87**
	Mean	17.03	4.93	19.11	3.58
	Median	16.07	4.71	18.43	3.53
	Std. dev.	5.23	1.83	5.2	1.36

The mean efficiency with models is 38% higher than with the text of safety standards; the median is 33% higher. The mean and median time were also higher with the text (12% and 15%, respectively). These initial overall results suggest that the use of models increases the efficiency of understanding safety compliance needs. In addition, the efficiency was higher with models for 15 subjects (75%), and the highest efficiency was obtained with models (10). The efficiency was above 4 for 11 subjects with models (55%) and for six with the text (25%). The lowest efficiency was obtained with the text (0.96), and the efficiency was lower than or equal to 3 for one subject with models (5%) and for seven with the text (35%).

The sample for efficiency with models is non-normal according to the Shapiro-Wilk test (p-value = 0.02 < 0.05). Therefore, we selected the paired Wilcoxon test for $H_{2,0}$. The test result indicates that the difference in the efficiency is statistically significant (p-value = 0.03 < 0.05). $H_{2,0}$ can thus be rejected and the results allow us to claim that the use of models can increase the efficiency of understanding safety compliance needs. The effect size (Cliff's d) for the efficiency can be regarded as medium-large or large (0.276 < d = 0.4 ≈ 0.428), which contributes to the practical importance of the results.

Efficiency was lower with models than with the text in the pilot experiment [10] and higher in the subsequent experiment [9], although not as high as in this paper. The adjustments and changes in experiment design have probably contributed to the increase in efficiency. The difference was not statistically significant in the previous experiment, and we found both evidence and counterevidence of the positive impact that using models can have on the efficiency of understanding safety compliance needs. In contrast, little evidence of the possible benefits in using text has been found in the experiment reported in this paper.

Regarding related work, both experiments in which the subjects spent less time for some understanding task when using models (e.g. [51]) and experiments in which understanding with text was faster (e.g. [52]) can be found.

We can conclude that the use of models can increase the efficiency of understanding safety compliance needs. However, it appears that an improvement in efficiency of understanding when using models depends on the specific understanding task and modelling notation. Special attention must be paid to experiment design for result reliability, as can be observed when comparing the experiment in this paper with our previous experiments.

4.3 Perceived Benefits in the Use of Models (RQ3)

Figure 5 shows the results about the subjects' perceived benefits in the use of models and in the use of the text of safety standards to understand safety compliance needs. The numbers in the bars indicate the data points of each possible answer for the corresponding statement.

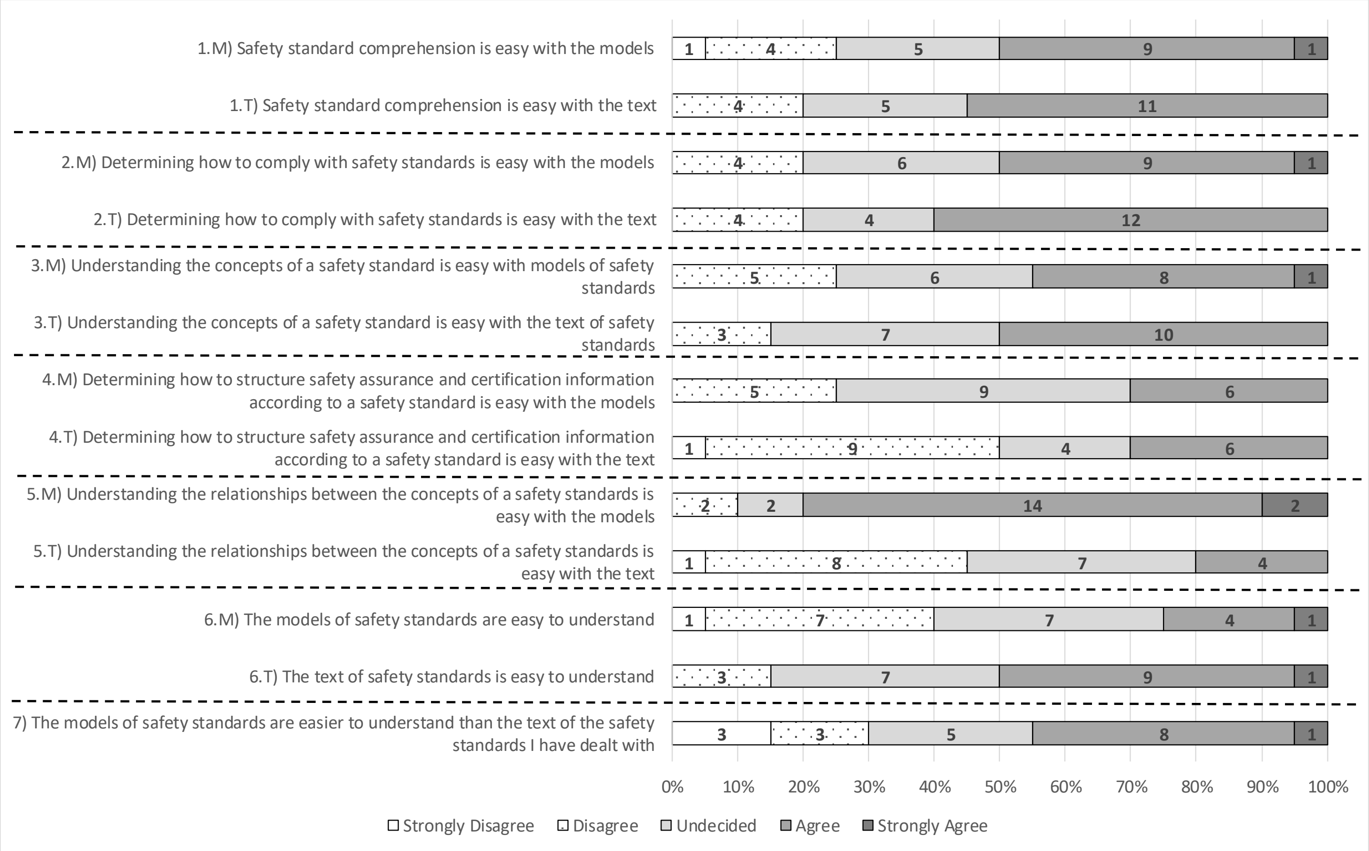


Figure 5. Perceived benefits in the use of models and of the text of safety standards to understand safety compliance needs

For models, the median is *Agree* only for one statement (*"Understanding the relationships between the concepts of a safety standard is easy with the models"*), and for two statements for the text (*"Safety standard comprehension is easy with the text"* and *"Determining how to comply with safety standards is easy with the text"*). Some subjects agreed or disagreed with each statement. The statements with the highest disagreement are *"Determining how to structure safety assurance and certification information according to a safety standard is easy with the text"*, *"Understanding the relationships between the concepts of a safety standard is easy with the text"*, and *"The models of safety standards are easy to understand"*. In total, the number of answers expressing agreement is largely higher than the number of answers expressing disagreement for both models (93% higher; 56 vs. 29) and text (61% higher; 53 vs. 33).

Some results deserve a deeper analysis. If statements 4 and 5, for which a wider disagreement is reported on the benefits of the text, are not considered, the number of answers expressing disagreement is 69% higher for models (22 vs. 13). If statement 5, for which a wider agreement is reported on the benefits of models, is not considered, the number of answers expressing agreement is 23% higher for the text (49 vs. 40). In addition, the statement with the highest number of answers expressing strong disagreement is *"The models of safety standards are easier to understand than the text of the safety standards I have dealt with"*. Therefore, and all in all, someone could argue that the subjects have a preference towards the use of the text of safety standards to understand safety compliance needs, except when they need to understand the relationships between the concepts of a safety standard and how to structure safety assurance and certification information. As counterevidence, *Agree* is the mode for *"The models of safety standards are easier to understand than the text of the safety standards I have dealt with"*. We conjecture that subjects with more experience in modelling, and especially with the notation used in the experiment, would agree more on the benefits of models.

We used the paired Wilcoxon test for $H_{3,0}$. We compared the results for each statement about the models and about the text, e.g. for 1M and 1T in Figure 5. There is a statistically significant difference for *"Determining how to structure safety assurance and certification information according to a safety standard is easy with the models/text"* (p-value = 0.02 < 0.05) and *"The models/text of safety standards are/is easy to understand"* (p-value = 0.01 < 0.05). Therefore, can be rejected $H_{3,0}$ and the results allow us to claim that there can be a difference in the perceived benefits of understanding safety compliance needs with the text of safety standards and with models. The effect size (Cliff's d) is large for *"Determining how to structure safety assurance and certification information according to a safety standard is easy with the models/text"* (d = 0.635 > 0.428) and medium for *"The models/text of safety standards are/is easy to understand"* (0.276 < d = 0.333 < 0.428). We also used the Mann-Whitney test to check whether treatment order (receiving first the model or the text, or receiving first the material about DO-178C or EN 50128) influenced the subjects' opinion about *"The models of safety standards are easier to understand than the text of the safety standards I have dealt with"*. The differences are not statistically significant.

In their comments, the subjects indicated several points. First, they claimed that the number of nodes and edges in the models impacted their understanding. Also, they noted that the relationships between the different models can be difficult to identify, but the structure of safety compliance needs is better represented in models. In addition, they noted that the ease of text understanding varies among text fragments and suggested that a combination of text and models might be the most suitable way to represent safety compliance needs.

When comparing the perceived benefits with the results for RQ1 and RQ2, there seems to be a gap between the subjects' perceptions and their actual performance. Effectiveness and efficiency of understanding safety compliance needs was higher with models, and the difference is statistically significant. However, the subjects seem to prefer the text of the standards in general. This outcome makes us think about two conclusions that might be drawn and might trigger new research. First, and focusing on the experiment, a hypothesis that can be derived from the results is that understanding the relationships between the concepts of a safety standard and how to structure safety assurance and certification information greatly contribute to understanding safety compliance needs. Unlike the rest of statements, there was an ample disagreement on these statements when asking about the text. Second, and from a more general perspective, someone might argue that the results provide evidence that actual performance when applying some technique can easily not match a user's perceived benefits. In other words, actual benefit in use might not be a reliable source to predict user acceptance. Although some users did not like the technique, its use could still be beneficial.

In relation to the subjects' perceived benefits in our precedent experiments [9,10], the median for *"The models of safety standards are easier to understand than the text of the safety standards I have dealt with"* has not been *Agree* in any study. *Undecided* has again been the median regarding the ease of understanding of the models, and the median has even decreased when asking the subjects about certain topics, e.g. about whether models facilitate safety standard comprehension. These results suggest that changing the notation has not contributed to an increase in the perceived benefits in the use of models. In all the experiments, the subjects have found benefits in the use of models but also some limitations. Non-high user acceptance might simply be a consequence of the use of models regardless of their format.

Most of the practitioners that analysed a model of the IEC 61508 standard [8] considered that it was easy to understand, and there was a wide agreement on the ease of understanding the relationships between the concepts of a standard. The model was specified as an UML class diagram, and the practitioners might have experience with this notation. In experiments on security assessment [58,59], the amount of positive characteristics regarding perceived usefulness and perceived ease of use was larger for models than for text.

In summary, although the subjects regard the use of models of safety compliance needs as useful for some purposes, they also find limitations and even seem to prefer the text of safety standards for some tasks. The change in the notation from our previous experiments aimed to improve the understanding of safety compliance needs, and although the achievement of this objective has been shown in relation to the effectiveness and the efficiency of understanding, the change does not seem to have positively impacted the perceived benefits. There appears to be a gap between the subjects' perception and their actual performance that could be further investigated.

5. Conclusion

Safety compliance needs can be difficult to understand from the textual descriptions in safety standards documents. The use of models has been proposed to facilitate the comprehension of these needs, but the empirical evidence in prior studies is insufficient to confirm the benefits of model usage.

In this paper, we have presented an experiment with 20 subjects that worked with textual specifications and with model-based ones to identify and thus understand safety compliance needs. According to the results, the use of models can improve the mean effectiveness of understanding by 22%. The use of models can also provide a higher efficiency rate on average, around 38%. Both results are statistically significant, thus, they confirm, for the first time, that the use of models can increase the effectiveness and the efficiency of understanding safety compliance needs. The effect size for both the effectiveness and the efficiency can also be regarded as large. This contributes to the practical importance of the results.

The subjects also found benefits in using the models, especially to understand the relationships between the concepts of safety standards. However, in general the subjects did not regard the models as easy to understand and regarded the text of safety standards as a better means to determine how to comply with the standards. On average, users were undecided about whether models of safety standards were easier to understand than the text.

Based on the experiment results and on prior studies, we conclude that the use of models can improve the understanding of safety compliance needs, but it does not seem to increase user acceptance. In this sense, the use of a SPEM-like notation instead of UML object diagrams appears to have contributed to improving the effectiveness and efficiency of understanding, but not the perceived benefits. The effect on efficiency also seems to depend on the understanding tasks. It can be argued that there is a gap between the subjects' performance when using models and the benefits that they find in the use. We consider that the conclusions are valid for students, but different ones could be drawn from experiments with practitioners. In addition to the notation change, other aspects that could have contributed to result significance in the experiment are the adjustments in the questions and the larger sample.

Insights that could trigger further research include that the benefits of using models or text seem to vary among understanding tasks. For example, the understanding of relationships between concepts appears to be better with models, but text has outperformed models in prior studies for some tasks. A classification of understanding tasks would be useful, indicating which representation approach would be more suitable to perform each task type. In addition, although the degree of understanding of different notations has received attention in the literature, we consider that more research is necessary.

For practitioners, the results provide evidence that using models could be a better approach to gain awareness of safety compliance needs. We think that three situations in which this use could be especially beneficial are: (1) training on a specific safety standard, as practitioners could more easily identify and understand the needs; (2) comparison of standards, as practitioners could analyse different standards according to the representation of their safety compliance needs in a same format, and; (3) agreement processes with e.g. customers, assessors, or certification authorities, as these stakeholders and system suppliers could base the processes on a more understandable format for the safety compliance needs that the supplier aims to fulfil. Standardisation bodies should consider the inclusion of further models in the safety standards documents so that the documents are more comprehensible.

It must also be taken into account that understanding safety compliance needs is not the only area that can benefit from using models of safety standards in particular and of safety certification information in general. For example, model-based representations of certification information enable the automation of compliance management against a standard [8] and the derivation and composition of new information, e.g. safety argumentation fragments [69]. Model-driven safety compliance can support the specification of correct-by-construction compliance information, the visualisation and analysis of this information in diagrams, compliance information exchange, and the assessment of compliance gaps, among other features [14].

Regarding future work, we plan to replicate the experiment to provide further evidence, or counterevidence, of the improvements from using models to understand safety compliance needs. Experiments with a larger sample and different types of subjects (e.g. practitioners) can provide new insights as well as experiments in which the subjects only use the text or the models of safety standards. We also want to study how the perceived benefits and thus user acceptance could be increased. According to the subjects' feedback, a combination of text and models might be preferred. This combination could further have an impact on the effectiveness and efficiency of understanding. Finally, it would be valuable to analyse the use of different notations, other understanding tasks, and representations of different types of safety certification information. A different task could be used to analyse compliance for a specific system, and different types of safety certification information could be safety cases.

Acknowledgement. The research leading to this paper has received funding from the AMASS (H2020-ECSEL grant agreement no 692474; Spain's MINECO ref. PCIN-2015-262), ENSE (Chile's CONICYT ref. REDI170020), iRel4.0 (H2020-ECSEL grant agreement no 876659), VALU3S (H2020-ECSEL grant agreement no 876852), and Treasure (JCCM ref. SBPLY/19/180501/000270; EC's European Regional Development Fund) projects, and from the Ramon y Cajal Program (Spain's MICINN ref. RYC-2017-22836; EC's European Social Funds). We are also grateful to all the subjects of the experiment for their participation and to the paper reviewers for their valuable suggestions. Finally, the first author thanks Carlos III University of Madrid, Spain, where he was working when the experiment reported in this paper was run.

References

1. L.E.G. Martins, T. Gorscheck, "Requirements engineering for safety-critical systems: A systematic literature review". Information and Software Technology, vol. 75, pp. 71-89 (2016).
2. S. Nair, J.L de la Vara, M. Sabetzadeh, L. Briand, "An extended systematic literature review on provision of evidence for safety certification". Information and Software Technology, vol. 56, pp. 689-717 (2014).

3. J.L. de la Vara, A. Ruiz, K. Attwood, H. Espinoza, R.K., Panesar-Walawege, A. Lopez, I. del Rio, T. Kelly, “Model-based specification of safety compliance needs for critical systems: A holistic generic metamodel”. Information and Software Technology, vol. 72, pp. 16-30 (2016).
4. J.L. de la Vara, M. Borg. K. Wnuk, L. Moonen, “An Industrial Survey on Safety Evidence Change Impact Analysis Practice”. IEEE Transaction on Software Engineering, vol.42, no. 12, pp. 1095-1117 (2016).
5. S. Nair, J.L de la Vara, M. Sabetzadeh, D. Falessi, “Evidence management for compliance of critical systems with safety standards: A survey on the state of practice”. Information and Software Technology. vol. 60, pp. 1-15 (2015).
6. M. García-Valls, J. Escribano-Barreno, J. García-Muñoz, “An extensible collaborative framework for monitoring software quality in critical systems”. Information and Software Technology. vol. 107, pp. 3-17 (2019).
7. L.T. Heeager, P.A. Nielsen, “A conceptual model of agile software development in a safety-critical context: A systematic literature review”. Information and Software Technology. vol. 103, pp. 22-39 (2018).
8. R.K. Panesar-Walawege, M. Sabetzadeh, L. Briand, “Supporting the verification of compliance to safety standards via model-driven engineering”. Information and Software Technology, vol. 55, no.5, pp. 836-864 (2013).
9. J.L. de la Vara, B. Marín, C. Ayora, G. Giachetti, “An Experimental Evaluation of the Understanding of Safety Compliance Needs with Models”. ER 2017, pp. 239-247.
10. J.L. de la Vara, B. Marín, G. Giachetti, C. Ayora, “Do Models Improve the Understanding of Safety Compliance Needs? Insights from a Pilot Experiment”. ESEM 2016, pp- 32:1-32:6.
11. OMG. 2008. Software & Systems Process Engineering Metamodel Specification (SPEM). http://www.omg.org/spec/SPEM/
12. RTCA. DO-178C: Software Considerations in Airborne Systems and Equipment Certification (2011)
13. CENELEC. EN 50128: Railway applications - Communications, signalling and processing systems - Software for railway control and protection systems. 2011
14. J.L. de la Vara, A. Ruiz, H. Espinoza, “Recent Advances towards the Industrial Application of Model-Driven Engineering for Assurance of Safety-Critical Systems”. MODELSWARD 2018, pp. 632-641.
15. H. Stallbaum, M. Rzepka, “Toward DO-178B-compliant Test Models”. MoDeVVa 2010, pp. 25-30.
16. C. Ayora, V. Torres, J.L. de la Vara, V. Pelechano, “Variability management in process families through change patterns”. Information and Software Technology. vol. 74, pp. 86-104 (2016).
17. S. Nair, J.L. de la Vara, A. Melzi, G. Tagliaferri, L. de-la-Beaujardiere, F. Belmonte, “Safety Evidence Traceability: Problem Analysis and Model”. REFSQ 2014, pp. 309-324.
18. J.L. de la Vara, G. Génova, J.M. Álvarez-Rodríguez, J. Llorens, “An analysis of safety evidence management with the Structured Assurance Case Metamodel”. Computer Standards & Interfaces, vol. 50, pp. 179-198 (2017).
19. OPENCOSS project, “D1.4 - Implementation of use cases on top of OPENCOSS platform” (2015)
20. AMASS project, “D1.6 - AMASS demonstrators (c)” (2019)

21. J. Vilela, J. Castro, L.E.G. Martins, T. Gorscheck, “Integration between requirements engineering and safety analysis: A systematic literature review”. Journal of Systems and Software, vol. 125, pp. 68-92 (2017).
22. L. Briand, D. Falessi, S. Nejati, M. Sabetzadeh, T. Yue, “Traceability and SysML design slices to support safety inspections: A controlled experiment”. ACM T. Softw. Eng. Meth., vol. 23, no. 1, pp. 9:1-9:43 (2014).
23. T. Stålhane, G. Sindre, “An experimental comparison of system diagrams and textual use cases for the identification of safety hazards”. Int. J. of Inform System Modeling and Design (IJISMD), vol. 5.1, pp. 1-24 (2014)
24. A. Abdulkhaleq, S. Wagner, “A controlled experiment for the empirical evaluation of safety analysis techniques for safety-critical software”. EASE 2015, pp. 16.
25. J. Jung, K. Hoefig, D. Domis, A. Jedlitschka, M. Hiller, “Experimental Comparison of Two Safety Analysis Methods and Its Replication”. ESEM 2013, p. 223-232.
26. T. Gonschorek, M. Zeller, K. Höfig, F. Ortmeier, “Fault Trees vs. Component Fault Trees: An Empirical Study”. SAFECOMP 2018 Workshops, pp 239-251.
27. A. Mouaffo, D. Taibi, K. Jamboti, “Controlled experiments comparing fault-tree-based safety analysis techniques”. EASE 2014, article no. 46.
28. L. Cyra, J. Górski, “Support for argument structures review and assessment”. Reliability Engineering & System Safety, vol. 96.1, pp. 26-37 (2011).
29. A. L. Oliveira, “A model-based approach to support the systematic reuse and generation of safety artefacts in safety-critical software product line engineering”. PhD Thesis. Universidade de São Paulo, Brazil, (2016).
30. A. de Lucia, C. Gravino, R. Oliveto, G. Tortora, “An experimental comparison of ER and UML class diagrams for data modelling”. Empirical Software Engineering, vol. 15, no. 5, pp. 455-492 (2010).
31. S. Abrahão, C. Gravino, E. Insfran, G. Scanniello, G. Tortora, “Assessing the Effectiveness of Sequence Diagrams in the Comprehension of Functional Requirements: Results from a Family of Five Experiments”. IEEE Transactions on Software Engineering, vol. 39, no. 3, pp. 327-342 (2013).
32. S. Abrahão, E. Insfran, C. Gravino, G. Scanniello, “On the effectiveness of dynamic modeling in UML: Results from an external replication”. ESEM 2009, pp. 468-472.
33. J.A. Cruz-Lemus, M. Genero, M.E. Manso, S. Morasca, M. Piattini, “Assessing the understandability of UML statechart diagrams with composite states”. Empirical Software Engineering, vol. 14, no. 6, pp. 685-719 (2009).
34. C.F.J. Lange, M.R.V. Chaudron, “Interactive Views to Improve the Comprehension of UML Models - An Experimental Validation”. ICPC 2007, pp. 221-230.
35. M. Torchiano, G. Scanniello, F. Ricca. G. Reggio, M. Leotta, “Do UML object diagrams affect design comprehensibility? Results from a family of four controlled experiments”. Journal of Visual Languages and Computing, vol. 41, pp. 10-21 (2017).
36. I. Reinhartz-Berger, K. Figl, O. Haugen, “Investigating styles in variability modeling: Hierarchical vs. constrained styles”. Information Software Technology, vol. 87, pp. 81-102 (2017).
37. G. Scanniello, C. Gravino, M. Genero, J.A. Cruz-Lemos, G. Tortora, “On the Impact of UML Analysis Models on Source-Code Comprehensibility and Modifiability”. ACM T. Software Engineering Methods, vol. 23, no. 2, pp. 13:1-13:26 (2014).
38. A.M. Fernández-Sáez, M. Genero, D. Caivano, M.R.V. Chaudron, “Does the level of detail of UML diagrams affect the maintainability of source code? A family of experiments”. Empirical Software Engineering. vol. 21, no. 1, pp. 212-259 (2016).

39. A. Nugroho, “Level of detail in UML models and its impact on model comprehension: A controlled experiment”. Information Software Technology, vol. 51, no. 12, pp. 1670-1685 (2009).
40. M. Staron, L. Kuzniarz, C. Wohlin, “Empirical assessment of using stereotypes to improve comprehension of UML models: A set of experiments”. Journal of Systems and Software, vol. 79, no.5, pp. 727-742 (2006).
41. J.A. Cruz-Lemus, M. Genero, D. Caivano, S. Abrahão, E. Insfrán, J.A. Carsí, “Assessing the influence of stereotypes on the comprehension of UML sequence diagrams: A family of experiments”. Information Software Technology, vol. 53, no. 12, pp. 1391-1403 (2011).
42. M. Morandini, A. Marchetto, A. Perini, “Requirements comprehension: A controlled experiment on conceptual modeling methods”. EmpiRE 2011, pp. 53-60.
43. J.M. Morales, E. Navarro, P. Sánchez, D. Alonso, “A family of experiments to evaluate the understandability of TriStar and i* for modelling teleo-reactive systems”. Journal of Systems and Software, vol. 114, pp. 82-100 (2016).
44. M.A. Teruel, E. Navarro, V. López-Jaquero, F. Montero, J. Jaén, P. González, “Analyzing the understandability of Requirements Engineering languages for CSCW systems: A family of experiments”. Information and Software Technology, vol. 54, no. 11, pp.1215-1228 (2012).
45. J.M. Morales, E. Navarro, P. Sánchez, D. Alonso, “A controlled experiment to evaluate the understandability of KAOS and i* for modeling Teleo-Reactive systems”. Journal of Systems and Software, vol. 100, pp. 1-14 (2015).
46. I. Hadar, I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, A. Susi, “Comparing the comprehensibility of requirements models expressed in Use Case and Tropos”. Information Software Technology, vol. 55, no. 10, pp. 1823-1843 (2013).
47. F.L. Siqueira, “Comparing the comprehensibility of requirements models: An experiment replication”. Information Software Technology, vol. 96, pp. 1-13 (2018).
48. M. Santos, C. Gralha, M. Goulão, J. Araújo, A. Moreira, J. Cambeiro, “What is the Impact of Bad Layout in the Understandability of Social Goal Models?”. RE 2016, pp. 206-215.
49. C. Gralha, M. Goulão, J. Araújo, “Analysing Gender Differences in Building Social Goal Models: A Quasi-Experiment”. RE 2019, pp. 165-176.
50. G. Scanniello, M. Staron, H. Burden, R. Heldal, “On the effect of using SysML requirement diagrams to comprehend requirements: results from two controlled experiments”. EASE 2014, pp. 49.
51. R. Razali, C.F. Snook, M.R. Poppleton, P.W. Garratt, R.J. Walters, “Experimental Comparison of the Comprehensibility of a UML-based Formal Specification versus a Textual One”. EASE 2007, pp. 1-11.
52. Z. Sharafi, A. Marchetto, A. Susi, G. Antoniol, Y.G. Guéhéneuc, “An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension”. ICPC 2013, pp. 33-42.
53. R.A. Rodrigues, M.O. Barros, K. Revoredo, L.G. Azevedo, H. Leopold, “An Experiment on Process Model Understandability Using Textual Work Instructions and BPMN Models”. Brazilian Symposium on Software Engineering, pp. 41-50 (2015).
54. M. Trkman, J. Mendling, M. Krisper, “Using business process models to better understand the dependencies among user stories”. Information and Software Technology. vol. 71, pp. 58-76 (2016).

55. M. Trkman, J. Mendling, P. Trkman, M. Krisper, "Impact of the conceptual model's representation format on identifying and understanding user stories". Information and Software Technology. vol. 116, article no. 106169 (2019).
56. A. Ottensooser, A. Fekete, H.A. Reijers, J. Mendling, C. Menictas, "Making sense of business process descriptions". Journal of Systems and Software, vol. 85, no. 3, pp. 596-606 (2012).
57. W. Heijstek, T. Kühne, M.R.V. Chaudron, "Experimental Analysis of Textual and Graphical Representations for Software Architecture Design". ESEM 2011, pp. 167-176.
58. K. Labunets, F. Massacci, F. Paci, L.M.S. Tran, "An Experimental Comparison of Two Risk-Based Security Methods". ESEM 2013, pp 163-172.
59. K. Labunets, F. Paci, F. Massacci, R. Ruprai, "An Experiment on Comparing Textual vs. Visual Industrial Methods for Security Risk Assessment". EmpiRE 2014, pp. 28-35.
60. K. Labunets, Y. Lib, F. Massacci, F. Paci, M. Ragosta, B. Solhaug, K. Stølen, A. Tedeschi, "Preliminary Experiments on the Relative Comprehensibility of Tabular and Graphical Risk Models". SESAR Innovation Days (2015).
61. K. Labunets, F. Massacci, F. Paci, "On the Equivalence Between Graphical and Tabular Representations for Security Risk Assessment". REFSQ 2017, pp. 191-208.
62. K. Labunets, F. Massacci, F. Paci, S. Marczak, F.M. de Oliveira, "Model comprehension for security risk assessment: an empirical comparison of tabular vs. graphical representations". Empirical Software Engineering, vol. 22, no. 6, pp. 3017-3056 (2017).
63. K. Labunets, F. Masacci, A. Tedeschi, "Graphical vs. Tabular Notations for Risk Models: On the Role of Textual Labels and Complexity". ESEM 2017, pp. 267-276
64. K. Labunets, "No Search Allowed: What Risk Modeling Notation to Choose?". ESEM 2018, article no. 20.
65. C. Wohlin, P. Runeso, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslen, "Experimentation in Software Engineering" (2nd ed.). Springer (2012).
66. ESA. Software engineering and standardisation (2006) http://www.esa.int/TEC/Software_engineering_and_standardisation/TECBUCUXBQE_0.html
67. T. Kelly, R. Weaver, "The goal structuring notation - a safety argument notation". Proceedings of the dependable systems and networks 2004 workshop on assurance cases. 2004.
68. OMG. 2017. Unified Modeling Language (UML). http://www.omg.org/spec/UML
69. G. Gallina, E. Gómez-Martínez, C. Benac-Earle, "Promoting MBA in the rail sector by deriving process-related evidence via MDSafeCer". Computer Standards & Interfaces, vol. 54, pp. 119-128 (2017).
70. IBM. 2013. Successful compliance with IEC 61508 safety standards. https://www.ibm.com/developerworks/rational/library/compliance-IEC-61508-safety-standards/index.html
71. M. Fowler, "UML distilled: a brief guide to the standard object modeling language". Addison-Wesley Professional (2004).
72. S. Vegas, C. Apa, N. Juristo, "Crossover Designs in Software Engineering Experiments: Benefits and Perils". IEEE Transactions on Software Engineering, vol. 42, no. 2, pp. 120-135 (2016).

73. S.S. Shapiro, M.B., Wilk, "An analysis of variance test for normality (complete samples)". Biometrika, vol. 52, no. 3–4, pp. 591–611 (1965).
74. J. Cohen, “Statistical Power Analysis for the Behavioral Sciences”. Routledge (1988).
75. N. Cliff, “Dominance statistics: ordinal analyses to answer ordinal questions”. Psychological Bulletin, vol. 114, no. 3, pp. 494–509 (1993)
76. N. Juristo, A.M. Moreno, “Basics of Software Engineering Experimentation”. Springer (2001).
77. B. Kitchenham, L. Madeyski, D. Budgen, J. Keung, P. Brereton, S. Charters, S. Gibbs, A. Pohthong, “Robust Statistical Methods for Empirical Software Engineering”. Empirical Software Engineering, vol. 22, no. 2, pp. 579-630 (2017)
78. F. Ricca, M. Di Penta, M. Torchiano, P. Tonella, M. Ceccato, “How Developers’ Experience and Ability Influence Web Application Comprehension Tasks Supported by UML Stereotypes: A Series of Four Experiments”. IEEE Transactions on Software Engineering, vol. 36, no. 1, pp. 96-118 (2010).
79. C. Wohlin, A. Gustavsson, M. Höst, C. Mattsson, “A framework for technology introduction in software organizations”. Conference on Software Process Improvement, pp. 167-176 (1996).
80. M. Höst, B. Regnell, C. Wohlin, “Using students as subjects - a comparative study of students and professionals in lead-time impact assessment”. Empirical Software Engineering, vol. 5, no. 3, pp. 201-214 (2000)
81. P. Runeson, “Using students as experiment subjects - an analysis on graduate and freshmen student data”. EASE 2003, pp. 95-102.
82. M. Svahnberg, A. Aurum, C. Wohlin, “Using students as subjects - an empirical evaluation”. ESEM 2008, pp. 288-290.
83. I. Salman, A.T. Misirli, N. Juristo, “Are Students Representatives of Professionals in Software Engineering Experiments?”. ICSE 2015, pp. 666-676.
84. D. Falessi, N. Juristo, C. Wohlin, B. Turhan, J. Münch, A. Jedlitschka,M. Oivo, “Empirical software engineering experts on the use of students and professionals in experiments”. Empirical Software Engineering, vol. 23, no. 1, pp. 452-489 (2018).