On the Effect of Using SysML Requirement Diagrams to Comprehend Requirements: Results from Two Controlled Experiments

Giuseppe Scanniello
Dipartimento di Matematica, Informatica ed Economia
University of Basilicata, Italy
giuseppe.scanniello@unibas.it

Miroslaw Staron, Håkan Burden, Rogardt Heldal
Computer Science and Engineering
Chalmers University of Technology & University of Gothenburg, Sweden
miroslaw.staron@gu.se;
{burden|heldal}@chalmers.se

ABSTRACT

We carried out a controlled experiment and an external replication to investigate whether the use of requirement diagrams of the System Modeling Language (SysML) helps in the comprehensibility of requirements. The original experiment was conducted at the University of Basilicata in Italy with Bachelor students, while its replication was executed at the University of Gothenburg in Sweden with Bachelor and Master students. A total of 87 participants took part in the experiment and its replication. The achieved results indicated that the comprehension of requirements is statistically significant when requirements specification documents include requirement diagrams without any impact on the time to accomplish comprehension tasks. On the basis of our results, we also present and discuss possible implications from the practitioner and researcher perspectives.

Categories and Subject Descriptors

D.2.0 [**Software Engineering**]: General

General Terms

Documentation, Experimentation.

Keywords

Controlled Experiment, Requirements Comprehension, Replication, Software Models, SysML, UML

1. INTRODUCTION

A requirement specifies a capability or a condition that must (or should) be satisfied, a function that a system must implement, or a performance condition that a system must achieve [19]. Sometimes requirements are provided directly by a customer (i.e., person or organization) paying for the system or are generated by the organization that is developing the system [8]. Defects may be caused if requirements are ambiguous, incomplete, inconsistent, silent (unexpressed), unusable, over-specific, and verbose requirements (both functional and non-functional) and will impact on overall quality of a software system [41].

In this context, modeling is very important and becomes even more relevant when computer based systems become larger, complex, and critical to human society. The System Modeling Language (SysML) is a general-purpose modeling language that provides a broad range of tools for engineering computer based systems. For example, the SysML provides multiple ways for capturing requirements and their relationships in both graphical and tabular notations [19]. As far as functional requirements, they can be modeled with use case diagrams and use case narratives. These notations are both in the UML (Unified Modeling Language) [29] and in the SysML. Requirements in the SysML can be depicted also on a requirement diagram (not in the UML). This kind of diagram is considered particularly useful in graphically representing hierarchies of specifications or requirements [17].

Although there are a number of empirical investigations on the UML [9], only a few studies on the SysML have been conducted so far (e.g.,[28]). This lack is even more evident in the context of empirical investigations aimed to study the possible benefits deriving from the use of the additional SysML models in the requirements engineering process.

In this paper, we investigate whether requirements diagrams in SysML provides additional benefits compared to the standard use case diagrams in UML. We present the results of a controlled experiment conducted at the University of Basilicata in Italy with third year Bachelor Students in Computer Science. The goal of this experiment was to study the effect of including requirements analysis diagrams in requirements specification documents. The results indicated that the use of this kind of diagrams improves the comprehension of specification documents without affecting the time to accomplish comprehension tasks. To show that these results were robust, an external replication was carried out at the department of Computer Science and Engineering (CSE[1]) in Gothenburg, Sweden, with Bachelor and Master

[1]This department is shared between Chalmers University of Technology and the University of Gothenburg, in Sweden.

EASE 2014, May 13-14, 2014, London, UK. Copyright 2014 ...$10.00.

Students. Varying the context or the environmental factors contribute some confidence that the effect is not limited to one particular setting and that the original results were not the result of the experimenter's bias. The results of the original experiment were confirmed in the replication. The original experiment and its replication are presented in this paper for the first time.

Paper Structure. In Section 2, we present the design of the experiment and its replication. In Section 3, we present the achieved results. The results are discussed together with possible threats to validity in Section 4. Related work is discussed in Section 5. Final remarks and future work conclude the paper.

2. THE EXPERIMENTS

We carried out two ABBA-type experiments [40] - the original experiment and an external replication. The original experiment (named E-UBAS, from here on) was carried out at the University of Basilicata in June 2012 with 24 third year students from the Bachelor's program in Computer Science. This experiment was replicated at CSE in December 2012 (named R1-UGOT, from here on). The participants in this experiment were 63 students. They were third year students from three Bachelor programs in Information Technology, Computer Science, and Software Engineering and first year students from the Master's program in Software Engineering.

The original experiment and its replication were carried out by following the recommendations provided by Juristo and Moreno [22], Kitchenham et al. [26], and Wohlin et al. [40]. The experiments are reported according to the guidelines suggested by Jedlitschka et al. [21]. For replication purposes, the experiment material is available online[2].

2.1 Goal

Applying the Goal Question Metric (GQM) template [6], the goal of the original experiment and its replication can be defined as:

Analyze SysML requirement diagrams ***for the purpose*** of evaluating requirements comprehensibility ***with respect to*** correctness of comprehension and time to accomplish a comprehension task ***from the point of view of*** the requirements analyst and the developer ***in the context of*** students in Computer Science/Software Engineering.

The use of GQM ensured that important aspects were defined before the planning and the execution of the experiment took place [40].

2.2 Context Selection

The following two systems were used as the **objects** in the original experiment and its replication:

Automobile. It is a mock-up of software for controlling car behavior with use cases about entering the car, anti-lock breaking or operating the climate control of a car.

ESS (Enhanced Security System). The system is designed to detect potential intruders. When an intruder is detected, the operators of the central monitoring station contact the local police or security companies, warning them of the intrusion. The use cases include providing medical/intruder/fire emergency response or investigative data.

The requirements specification documents of these two systems were built on the samples provided in [17]. This book is used to prepare for SysML certification: the OMG Certified Systems Modeling Professional (OCSMP) [18]. The Automobile and ESS systems are used to get the first two levels of the SysML certification. The choice of domains to model can be considered a good compromise of generality and industrial application. A more detailed, industrially relevant problem would be difficult to use at two geographically distinct universities with different profiles, the choice was also suitable for both notations thus minimizing the risk of biased objects of the experiment [2].

One of the authors reviewed all the documentation available in that book and then selected the diagrams and the chunks of the documentation that was of interest for our study. For example, use case narratives was added according to the template suggested by Bruegge and Dutoit [8]. For each experiment, the design choices above allowed reducing internal and external validity threats.

The documentation (including the diagrams) of both the systems was then translated into Italian (for the original experiment) to avoid that different levels of familiarity with English could bias the results. The replication was performed using the documentation in English. This difference was introduced because the official language of instruction at the Gothenburg University is English.

The materials available for the participants were: *(i)* a problem statement; *(ii)* the list of the non-functional requirements together with their unstructured textual descriptions; *(iii)* two requirement diagrams; *(iv)* a use case diagram together with the narratives of its use cases; and *(v)* descriptions of the actors. The Automobile system was specified using 16 non-functional requirements, while ESS was specified using 14. The number of use cases of Automobile and ESS were 8 and 5, respectively. This slight difference in the size is because the requirements specification documents used in the experiments were based on the samples provided in [17]. The used specification documents are available online on the web page of our study.

2.3 Participants

We conducted the original experiment and its replication under controlled conditions using *convenience sampling* from the population of junior software developers with *students as participants*. The participants had the following characteristics (significant differences between these groups are in italics):

E-UBAS. The participants were students of *a software engineering course*. They had passed all the exams related to the following courses: Object Oriented Programming I and II and Databases. In these courses the participants studied and applied the UML [29] on university problems.

R1-UGOT. The participants were students of a *model-driven software development course*. The main goal of this course was in-depth teaching executable modeling. These students attended one of four different programs - a Master program in software engineering or one of three Bachelor programs in IT, computer science, or software engineering. All students had *successfully completed at least 120 ECTS*

[2]The URL of the web page of our study is: www2.unibas.it/gscanniello/SysML/. The reader can find: the experimental package, the raw data, and a reference to a technical report (i.e., gupea.ub.gu.se/handle/2077/32632) with analyses not reported here for space reason.

7. The maximum acceleration of a car is strongly connected to (one or more answers may be correct) ☐ Engine power ☐ Car noise ☐ The number of the cylinders of the engine ☐ The space for the occupants inside the car ☐ The maximum speed					
How much do you trust your answer?					
☐ Unsure	☐ Not sure enough		☐ Sure Enough	☐ Sure	☐ Very Sure
How do you assess the question?					
☐ Very difficult	☐ Difficult		☐ On average	☐ Simple	☐ Very simple
What is the "main" source of information used to answer the question?					
☐ Previous Knowledge	☐ Requirements List	☐ Internet	☐ Use Cases	☐ Use Case Diagram	☐ Requirement Diagrams

Figure 1: A question example from the comprehension questionnaire of Automobile

credits[1]. The modeling experience of these participants can be considered higher than those of E-UBAS.

Although the experience in modeling was different for both groups of participants, all participants in the experiments studied the SysML and the requirement diagram, in particular, for the first time as preparation for the experiments. Before each experiment, the participants attended a seminar of about two hours.

The students participated in the original experiment and its replication on a voluntary basis: we did not force and we did not pay them for their participation. However, we awarded the students for their participation to the experiments with a bonus towards their final mark. They were clearly informed about these conditions. At R1-UGOT 70% of the students of the course attended the experiment and 80% of the students participated in E-UBAS. This shows that only motivated students participated in the original experiment and its replication.

2.4 Variable Selection

We considered the specification documents without requirement diagrams as the *Control Group* and the group with requirement diagrams as the *Treatment Group*. The independent variable in the experiments was *Method*. It is a nominal variable that can assume the following two values: RD (specification document with requirement diagrams) and NORD (specification document without requirement diagrams).

The direct dependent variables are:

Comprehension - the level of correct comprehension of requirements that the participant achieved.

Completion time - the time that the participant spent to accomplish the experimental task.

The variables were measured through questionnaires as **experiment instruments** - one questionnaire for each experiment round. The questionnaire was composed of nine multiple-choice questions. Each question admitted one or more correct answers among a set of five. The comprehension questionnaire of each system was the same independently from the method experimented (RD and NORD). To quantify the quality of answers and the comprehension achieved, we used the approach proposed by Kamsties et al. [24]. In particular, we computed the number of correct responses divided by 9 (i.e., the number of questions in the comprehension questionnaire). We consider a response to a question to be correct if the participant selected all the correct alternatives and no incorrect alternatives were selected. The used measure assumes values in the interval $\in [0, 1]$. A value close to 1 means that a participant got a very good comprehension since he/she answered correctly to all the 9 questions of the questionnaire. Conversely, a value close to 0 means that a participant obtained a low comprehension.

Figure 1 reports a sample question for Automobile. The correct expected answers are: *Engine power* and *The number of the cylinders of the engine*. These answers could be easily derived from both the list of the non-functional requirements and the requirement diagrams. Each response that does not report only these two answers is considered incorrect. For example, if a participant gives either *Engine power* or *The number of the cylinders of the engine* as the answer, the response is incorrect. Although different approached have been proposed in the literature to estimate the comprehension achieved by the participants (e.g., [1], [31]), we opted here for the approach above because it is more suitable for multiple-choice questions and because we were interested in the correct and complete comprehension of requirements [24].

To calculate the second dependent variable - completion time - we used the time (expressed in minutes) to answer the questions of the comprehension questionnaire, which was directly recorded by each participant on the paper copy of that questionnaire. Low values for the time mean that the participants were quicker in completing the experiment. Both comprehension and completion time complement each other - one describes the correctness of requirements comprehension and the other one the efficiency of the participant, while performing requirements comprehension.

We also analyzed the effect of the other independent variables (also called co-factors, from here on):

System. It denotes the system (i.e., Automobile or ESS) used as the experimental object. The effect of the System factor should not be confounded with the main factor. However, for the sake of consistency we analyzed whether this assumption holds.

Trial. It denotes in which experiment trial a particular participant was exposed to the requirement diagram. As the participants worked on two different experimental objects (Automobile and ESS) in two laboratory trials. We analyzed whether the order might affect the results.

2.5 Hypotheses Formulation

The following two null hypotheses have been formulated and tested:

Hn0. The mean value of the comprehension for the RD factor is the same as the mean value of the comprehension variable for the NORD factor.

Hn1. The mean value of the time to complete the task for the RD factor is the same as for the NORD factor.

The alternative hypotheses can be easily derived (e.g., Ha0

[1]120 ECTS is equivalent to 2 years of full studies. 1 year = 60 ECTS, European Credit Transfer System

Table 1: Experiment design

Trial	Group A	Group B	Group C	Group D
First	Automobile, RD	ESS, NORD	Automobile, NORD	ESS, RD
Second	ESS, NORD	Automobile, RD	ESS, RD	Automobile, NORD

- The mean value of the comprehension for the RD factor is **not** the same as the mean value of the comprehension variable for the NORD factor).

Hn0 and Hn1 are both two-tailed because we are interested in the effect of using requirement diagrams and do not expect a positive nor a negative effect. Even though it can be postulated that the participants in the treatment group were provided with additional information it could also be the case that the provided extra information required more time to accomplish a comprehension task. We can hypothesize that this additional information is more suitable to reduce ambiguities and to improve the comprehensibility of requirements, but impose additional burden on remembering extra information thus increasing risk for misunderstandings. Our postulation is supported by the used framework that is suggested by Aranda et al. [2]. This framework is based both on the underlying theory of modeling languages and on cognitive science principles.

2.6 Design of the experiments

We used the within-participants counterbalanced experimental design (see Table 1). This design ensures that each participant works on different experimental objects (Automobile or ESS) in two trials (or runs), using RD or NORD each time. We opted for that design because it is particularly suitable for mitigating possible carry-over effects[4]. As for E-UBAS, we used the participants ability as blocking factor: the groups are similar to each other with respect to the number of high and low ability participants[5]. This experiment is balanced with respect to the number of participants assigned to RD and NORD (each groups contained 6 students). The participants were randomly assigned to the four groups in R1-UGOT. The number of participants in the groups A, B, C, and D were 10, 17, 28, and 8, respectively. The inequality of groups was caused by the fact that no blocking was used and the design was random. In both the original experiment and its replication, we gave a 15-minute break when passing from the first laboratory trial to the second one.

2.7 Experimental Tasks

We asked the participants to perform the following tasks: **Comprehension tasks**. The participants were asked to fill in a paper copy of the comprehension questionnaires (one for each system as summarized in Table 1). We defined the questions to assess several aspects related to the comprehension of requirements. All the questions were formulated using a similar form/schema (see Figure 1). As suggested by Aranda et al. [2], for each question in the comprehension questionnaires we also asked the participants to specify: *(i)* how much they trusted the answers given, *(ii)* the perceived level of difficulty, and *(iii)* the "main" source of information exploited to answer a question. The questions *(i)* and *(ii)* gave insights about the participant's judgment regarding the given answer and the ease in obtaining the information required to answer the question, respectively. Differently, the main source of information allowed us to get qualitative indications on how the participants used the models provided to deal with comprehension tasks. Figure 1 shows a sample question for Automobile, when the participants used RD. Among the possible source of information there is the possibility of choosing requirement diagrams. This source of information was not present among the alternatives when the comprehension task was accomplished without requirement diagrams (i.e., NORD).

Post-experiment task. We asked the participants in E-UBAS to fill in a paper copy of the post-experiment survey questionnaire. This questionnaire contained questions about: the availability of sufficient time to complete the tasks and the clarity of the experimental material and objects. The goal was to obtain feedback about the participants' perceptions of the experiment execution. The questions of the post-experiment survey questionnaire are reported in Table 2.

Table 2: Post-experiment survey questionnaire

Id	Question	Answers
Q1	I had enough time to answer the questionnaire	(1-5)
Q2	The objectives of each task were clear	(1-5)
Q3	The questions were clear to me	(1-5)
Q4	The possible answers were clear to me	(1-5)
Q5	I found useful the exercise from the the practical point of view	(1-5)
Q6	I found useful the requirement diagrams	(1-5)
Q7	I found useful the combination of requirement diagrams and the list of requirements	(1-5)
Q8	The time I spent on requirement diagrams with respect to the total time to accomplish the single task was	(A-E)
1 = Strongly agree, 2 = Agree, 3 = Neutral, 4 = Disagree, 5 = Strongly disagree		
A. $< 20\%$; B. $\geq 20\%$ and $< 40\%$; C. $\geq 40\%$ and $< 60\%$; D. $\geq 60\%$ and $< 80\%$; E. $\geq 80\%$		

2.8 Experiment operation

The participants in the original experiment and its replication first attended an introductory lesson in which the supervisors presented detailed instructions on the experiment to be carried out. The supervisors highlighted the goal of the experiment without providing details on the experimental hypotheses. The participants were informed that the data collected in each experiment were used for research purposes and treated confidentially.

After the introductory lecture, the participants were assigned to the groups A, B, C, and D (see Table 1). No interaction was permitted among the participants, both within each laboratory trial and while passing from the first trial to the second one. No time limit was imposed to accomplish each of the two trials.

To carry out the experiment, the participants first received the material for the first laboratory trial, and when they had finished, the material for the second trial was provided. After the completion of both the trials, the post-experiment questionnaire was given to the participants in E-UBAS.

We asked the participants to use the following experimental procedure: *(i)* specifying name and start-time; *(ii)* an-

[4]If a participant is tested first under the condition A and then under the condition B, he/she could potentially exhibit better or worse performances under the condition B.

[5]The students with average grades below 24/30 were classified as low ability participants, otherwise high, as proposed by Abrahão et al. [1].

swering the questionnaire; and *(iii)* marking the end-time. We did not suggest any approach to browse the requirement specification documents.

2.9 Analysis Procedure

To perform the analysis of the gathered data, we carried out the following steps:

1. We calculated the descriptive statistics of the dependent variables.

2. We tested the null hypotheses using unpaired analyses because the comprehension tasks were accomplished on two different experimental objects (see Table 1). We have planned to use unpaired t-test when the data follow a normal distribution. The normality has been verified using the Shapiro-Wilk W test [35]. A p-value smaller than the α threshold allows us to reject the null hypothesis and to conclude that the data are not normally distributed. If the data will not be assumed to be normally distributed, our non-parametric alternative to the unpaired t-test was the Wilcoxon rank-sum test (also known as the Mann Whitney test) [12]. The chosen statistical tests analyze the presence of a significant difference between independent groups, but they do not provide any information about that difference [23]. Therefore, in the context of the parametric analyses, we used Cohen's d [11] effect size to obtain the standardized difference between two groups. That difference can be considered: negligible ($|d| < 0.2$), small ($0.2 \leq |d| < 0.5$), medium ($0.5 \leq |d| < 0.8$), and large ($|d| \geq 0.8$) [23]. Conversely, we used the point-biserial correlation r in case of non-parametric analyses. The magnitude of the effect size measured using point-biserial is: small ($0 < r \leq 0.193$), medium ($0.193 < r \leq 0.456$), and large ($0.456 < r \leq 0.868$) [23]. We also analyzed the statistical power for each test performed. It is computed on the basis of the test executed. The statistical power is the probability that a test will reject a null hypothesis when it is actually false. The value 0.80 is considered as a standard for the adequacy [15]. The statistical power is computed as 1 minus the Type II error (i.e., β-value). A $\beta - value$ allows understanding how strong the effect size of the tested null hypothesis is. Values above 0.80 are considered the standard for the adequacy of this kind of error. We analyzed statistical power when a null hypothesis can be rejected, the β-value otherwise.

3. To analyze the influence of the considered co-factors, we planned to use a two-way Analysis of Variance (ANOVA) [13] if the data was normally distributed and if their variance is constant. The normality and the variance of the data were tested using the tests of Shapiro and Levene [27], respectively. In case these assumptions are not verified, we would use a two-way permutation test [4], a non-parametric alternative to the two-way ANOVA test.

4. To graphically show the answers of the post-experiment survey questionnaire, we adopted boxplots. These are widely employed since they provide a quick visual representation to summarize data. The responses to the post-experiment questionnaire were analyzed by using the median of the answers to each question.

5. The participants' opinions of each question of the comprehension questionnaire (i.e., how much they trusted the answer given and the perceived level of difficulty) were illustrated by means of descriptive statistics. On the other hand, the answers concerned to the main source of information are graphically summarized by means of mosaic plots. This difference was introduced because the set of possible answers for RD was different from that for NORD regarding the question on the main source of information.

In all the statistical tests, we decided (as custom) to accept a probability of 5% of committing Type-I-error [40] (i.e., the α threshold is 0.05). The R environment[6] for statistical computing has been used in the data analyses.

2.10 Documentation and Communication

The success or the failure of replications may be influenced by the documentation exchanged among experimenters [36] and their communication [38]. To deal with these issues, we used a laboratory package and knowledge sharing mechanisms. A properly management of these issues also reduced the risks related to the consistency across the conditions in the replicated experiment. Consistency is critical especially in case of external replications. All the material used (e.g., specification documents and comprehension questionnaires) in the original experiment was translated from the Italian into English and shared with the replicators, who asked clarifications to the original experimenter when needed.

We began with an initial face-to-face meeting. The results of this meeting were reported in minutes. We exchanged the minutes of this meeting by e-mail in order to agree to a shared common research plan. This phase was relevant to share knowledge among the experimenters and to discuss possible issues related to the study. We used instant messaging tools and e-mails to establish a communication channel in all the phases of our research collaboration.

2.11 Differences between the Original Experiment and the Replication

We introduced some variations in R1-UGOT with respect to E-UBAS:

− The participants in R1-UGOT were more experienced in software modeling than E-UBAS. This alteration was made to better analyze the effect of more highly experienced participants.

− A different group of researchers conducted R1-UGOT (i.e., the external replication). This variation was introduced to deal with external validity threats. However, consistency issues across the different experimenters could be possible. As discussed before, we carefully managed communication among experimenters to administer these issues.

− To familiarize with the used experimental procedure, the participants in the E-UBAS experiment accomplished a training session in which an exercise similar to that would appear in the experimental tasks was accomplished. The participants dealt with the specification document of an AudioPlayer system (details can be found in our experimental package). This exercise was skipped in R1-UGOT because the participants were more experienced in modeling and because of time constraints.

− In E-UBAS the participants filled in a pre- and a post-questionnaire. The results of the pre-questionnaire were used to get information about the participants' ability (the blocking factor). The pre- and the post-questionnaires were not used in R1-UGOT for time constraints and for the large number of participants.

− The participants in E-UBAS could use the Web. This was not possible in R1-UGOT for logistic issues.

[6] www.r-project.org

Table 3: Descriptive statistics

Experiment	Completion time						Comprehension					
	RD			NORD			RD			NORD		
	Med.	Mean	Std. Dev.	Med.	Mean	Std. Dev.	Med.	Mean	Std. Dev.	Med.	Mean	Std. Dev.
E-UBAS	26	26.33	10.483	26	28.04	9.466	0.667	0.657	0.17	0.444	0.449	0.198
R1-UGOT	15	14.95	4.911	15	15.23	4.987	0.56	0.508	0.216	0.44	0.385	0.196

Table 4: Results for Hn0 and Hn1

Experiment	Dependent Variable	#obs for RD	#obs for NORD	p-value	Statistical Power	β-value
E-UBAS	Comprehension	24	24	YES (< 0.001)	0.949	0.051
	Completion time	24	24	NO (0.556)	0.068	0.932
R1-UGOT	Comprehension	63	63	YES (< 0.001)	0.881	0.119
	Completion time	59	56	NO (0.805)	0.064	0.936

3. RESULTS

We present the results of the data analysis following the procedure presented above.

3.1 Descriptive statistics and exploratory data analysis

Table 3 shows the descriptive statistics (i.e., median, mean, and standard deviation) of completion time and comprehension, respectively, grouped by Method.

Comprehension. The comprehension values of the participants in E-UBAS was on average higher with RD. Similar results were achieved on R1-UGOT. In addition, we can observe that the participants in E-UBAS achieved better comprehension values than the participants in R1-UGOT on RD. For NORD, there was a slight tendency in favor of R1-UGOT: the median values are mostly the same, while the mean value is lower for R1-UGOT. A plausible justification for that results is that the participants in E-UBAS were from a more homogenous group than the participants from R1-UGOT (i.e., one program compared to four programs at two different levels).

Completion time. The participants on average spent less time for RD with respect to NORD: 26.33 and 28.04 for E-UBAS and 14.95 and 15.23 for R1-UGOT. Within each experiment, the median values are the same independently from the method used (26 and 15 for E-UBAS and R1-UGOT, respectively). We can also observe that the participants in R1-UGOT spent on average less time than those to E-UBAS to accomplish the tasks with both RD and NORD. This difference could be due to the fact that the participants in R1-UGOT had more experience with software modeling and therefore more familiar with the UML based specification documents.

3.2 Influence of Method

3.2.1 Testing Hn0.

For both the E-UBAS and R1-UGOT, parametric statistical analyses could not be applied. As for E-UBAS, the Shapiro test returned 0.003 and 0.223 as the p-values for RD and NORD, respectively. The p-values on R1-UGOT were 0.086 for RD and 0.016 for NORD.

Table 4 shows the results for the analyses for Influence of Method. We can reject the null hypothesis Hn0 in both the original experiment and its replication. The p-values returned by the Mann-Whitney test were less than 0.01 in both E-UBAS and R1-UGOT, while the statistical power values were both above the 0.80 threshold (i.e., 0.949 for E-UBAS and 0.881 for R1-UGOT).

Table 5: Analysis of co-factors for comprehension

Exp_ID	System	Method vs. System	Trial	Method vs. Trial
E-UBAS	NO (0.373)	NO (0.941)	NO (1)	NO (0.623)
R1-UGOT	YES (< 0.001)	NO (0.596)	NO (0.366)	NO (1)

3.2.2 Testing Hn1.

We used the unpaired t-test in E-UBAS (the Shapiro test returned as the p-values 0.216 and 0.437 for RD and NORD, respectively). This test was not be applied for R1-UGOT. In fact, the Shapiro test returned 0.028 and 0.154 as the p-values for RD and NORD, respectively.

The results shown in Table 4 indicate that Hn1 could not be rejected in both E-UBAS and R1-UGOT. The β-values are always high: 0.932 for E-UBAS and 0.936 for R1-UGOT. It is worth mentioning that the number of observations for R1-UGOT is less than 63 for both RD and NORD. In particular, we did not consider 11 observations (4 for RD and 7 for NORD) in this analysis because the participants did not write their finish time (the time was not provided in the questionnaires). The experimenters were not able to check the start/stop time because many participants simultaneously returned back the questionnaires.

3.3 Effect of co-factors

The results of the analysis of the co-factors is summarized in Table 5. For each experiment, this table reports whether or not a co-factor has any effect on each dependent variable. The results for completion time are not reported because the main factor did not have any effect on that variable. The obtained p-values are shown within brackets. We could apply a two-way ANOVA only for R1-UGOT on System. In all the other cases, we applied a two-way permutation test. In particular, the data were not normally distributed in E-UBAS for RD on Automobile (p-value = 0.01) and for RD in the second laboratory trial (p-value = 0.039). As far as R1-UGOT is concerned, the data were non-normal for NORD in the first laboratory trial (p-value = 0.014). The results about the interaction between Method and the co-factors System and Trial are shown as well.

3.3.1 System

The results show that the effect of System on comprehension was not statistically significant for E-UBAS (p-value =

Table 6: Results for trust and complexity

Experiment	Experiment	Automobile						ESS					
		RD			NORD			RD			NORD		
		Med.	Mean	Std. Dev.	Med.	Mean	Std. Dev.	Med.	Mean	Std. Dev.	Med.	Mean	Std. Dev.
E-UBAS	Trust	3	3.383	0.748	3	3.152	0.782	3	3.385	0.792	3	3.028	0.783
	Complexity	4	3.636	0.65	4	3.514	0.667	4	3.5	0.711	3	3.346	0.715
	Source	6	4.103	2.087	2	2.067	0.943	6	4.769	1.876	2	2.355	1.021
R1-UGOT	Trust	4	3.69	1.141	3	3.203	1.083	3	2.909	1.096	3	2.633	1.026
	Complexity	4	3.612	0.98	3	3.258	0.917	3	3.02	0.983	3	2.667	0.909
	Source	6	3.901	2.188	2	1.761	0.848	6	4.134	2.132	2	2.058	0.861
Trust values: (1) "Unsure"; (2) "Not sure enough"; (3)"Sure enough"; (4) "Sure"; (5) "Very sure" Complexity value: (1) "Very difficult", (2) "Difficult", (3) "On average", (4) "Simple", (5) "Very simple"													

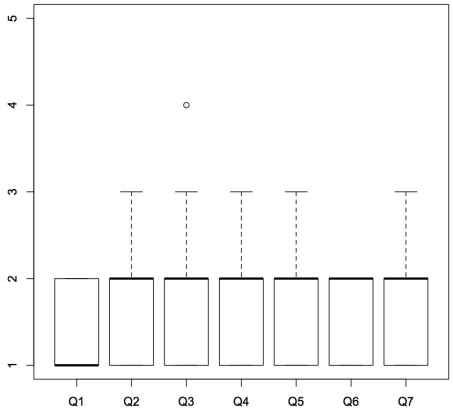

Figure 2: Box-plot of the answers of the post-experiment survey questionnaire

0.373), while it was statistically significant for R1-UGOT (p-value < 0.001). Descriptive statistics suggested that the participants in R1-UGOT obtained better comprehension values when performing the task on Automobile. For that system, the median was 0.56 and the mean 0.542, while 0.33 and 0.352 were the median and the mean value for ESS, respectively. The effect of System could be due to the different levels of familiarity of the participants with the problem domain of the two systems. In both E-UBAS and R1-UGOT the interaction between Method and System was not statistically significant as the obtained p-values were 0.596 and 0.941, respectively.

3.3.2 Trial

The results suggest that the trial effect on comprehension was statistically significant neither in the original experiment nor in the replication. The p-values were 1 and 0.596, respectively. In addition, the interaction between Method and Trial was not statistically significant: the p-values were 0.623 for E-UBAS and 1 for R1-UGOT, respectively. That is, either learning nor fatigue effect was observed.

3.4 Post-experiment survey results

Figure 2 graphically shows the answers to the questions of the post-experiment survey questionnaire. Indeed, the boxplots summarize the answers to the questions from Q1 to Q7 of the participants in E-UBAS. The participants in that experiment judged adequate the time to perform the comprehension task (Q1 - enough time). The median is equal to 1 (strongly agree). Regarding Q2 (objectives perfectly clear), the participants agreed on the fact that the objectives of the experiment were perfectly clear: the median is 2 (agree). For Q3 (questions clear) and Q4 (answers clear) the median are 2 (agree), namely the participants found clear both the questions and the answers of the comprehension questionnaires. The median for Q5 (education perspective) was 2 (agree). The participants found useful the experiment and judged useful the requirement diagrams. The medians for Q6 (usefulness of requirement diagrams) and Q7 (requirement diagrams combined with a requirements list are useful) are 2 (agree).

With respect to Q8 (time spent to analyze requirement diagrams), the median is D. The participants declared to have spent from 60% to 80% of their time to read requirement diagrams, while performing a comprehension task.

3.5 Further Analysis

In this section, we summarize the results of the analyses about the participants' opinion on: how much they trusted the answer given and the perceived level of difficulty. Descriptive statistics of the given answers are reported in Table 6. Figure 3 depicts the mosaic plots about the source of information used for answering the questions of the comprehension questionnaires in E-UBAS and R1-UGOT.

3.5.1 Trusting the given answers

The results suggest that the trusting level increases when participants are provided with requirement diagram (see Table 6). When requirement diagrams were not provided, the participants were less confident on the answers given as the descriptive statistics shown in Table 6 suggest.

3.5.2 Complexity of the questions

The participants in E-UBAS and R1-UGOT overall found the comprehension tasks not so difficult whatever the treatment was. Indeed, the tasks are perceived slightly less complex when the requirement diagrams are included in the requirement specification document (see Table 6).

3.5.3 Source of information

The mosaic plots in Figure 3 suggest that the requirement diagrams are the main source of information for RD, while the list of requirements is the main source of information for NORD. Regarding RD, the light grey rectangle (label 6) is always the largest considering the trials and the experimental objects alone. The second source of information used is the requirement list (label 2), that becomes the first one when the participants accomplished the comprehension task with NORD. It is worth mentioning that the mosaic plot in Figure 3 presents some asymmetries because of the inequality of the groups in Table 1.

4. DISCUSSION

Representations (and also visual notations) can improve the reasoning and the comprehension in several ways [2].

Figure 3: Mosaic plots about the source of information for E-UBAS (a) and R1-UGOT (b)

Based on the paper by Scaife and Rogers [32] and the results presented above, requirement diagrams does not affect offloading. That is, they do not reduce cognitive effort. This kind of diagram makes reasoning and problem solving easier (i.e., re-representation) and due to its graphical notation allows spending cognitive power more effectively (i.e., graphical constraining). This could be possible because relations among requirements are made explicit when using that notation. Also, making explicit requirements derivations, and satisfy and verify relationships could improve comprehension performances. Without requirement diagrams, all this information, that is present in the unstructured textual description of the requirements, has to be inferred, making reasoning more difficult and complex.

The achieved results also suggest that the benefit deriving from the use of requirement diagrams are independent from the UML modeling experience of the participants in the experiments. For both the original experiment and its replication, the effect of Method was statistically significant on the comprehension of requirements. It seemed that modeling experience only affected the task completion time: more experienced participants spent less time (see Table 3).

Although we chose systems on which the participants were familiar with, we observed that for the replication performed in Sweden seemed to be more difficult than Automobile in terms of comprehensibility. These results did not allow us to provide a definitive conclusion about the influence of the co-factor System (i.e., whether ESS was more difficult than Automobile) and could be justified by the participants' varying levels of familiarity with the problem domains of these systems. The effect of System and the fact that no statistically significant interaction was observed between Method and System suggest that the familiarity with the problem domain affected comprehensibility independently from the presence or the absence of requirement diagrams in the specification documents.

4.1 Implications

To judge the implications of our investigation, we adopted a perspective-based approach [5]. In particular, we based our discussion on the *practitioner/consultant* (simply *practitioner* in the following) and *researcher* perspectives using the guidelines by Kitchenham et al. [25]:

• Independently from the participants' experience and profile, the use of requirement diagrams improves the comprehension of requirements. This result is relevant from both the practitioner and the researcher perspectives. From the practitioner perspective, this result is relevant because requirement diagrams can be used as a communication mechanism among analysts, or as a validation tool between analysts and stakeholders. From the researcher perspective, it is interesting to investigate whether variations in the context (e.g., larger and more complex systems and more or less experienced stakeholders) lead to different results.

• The presence of requirement diagrams induces no additional time burden. The practitioner could be interested in that result because requirement diagrams allow stakeholders to get an improved comprehension of requirements without affecting the time to comprehend them. This result is relevant for the researcher because it could be interesting to investigate in which cases the processing and the integration of the information in requirement diagrams and in the specification document could increase/reduce comprehension time.

• The requirements specification documents were realistic enough. Then, we believe that our findings could scale to real projects. To corroborate this assertion, we need further replications with different experimental objects as well as case studies in real software development projects.

• The participants found requirement diagram to be the most relevant source of information to comprehend requirements. This finding is relevant for the researcher, who could be interested in assessing if and how this concern affects benefits stemming from the use of requirement diagram.

• The requirement diagrams are less common in the software industry than the UML diagrams used in our specification documents (e.g., [14, 34]). The results of our study could then promote the adoption of requirement diagrams in industry for both software and system modeling. Transferring a new technology, method, or tool to practitioners is easier when an empirical evaluation is performed and its results show that such a technology solves actual issues [30].

4.2 Threats to Validity

We here present an overview of the main possible threats that could affect the validity of our results.

Conclusion Validity. We minimized threats to conclusion validity coming from statistical methods. We used infer-

ential statistics and in particular, we used parametric test when the assumption was verified, non-parametric tests otherwise. The size of the sample was sufficient to draw statistical conclusions. The composition of the groups in R1-UGOT could also affect the validity of the results.

Internal Validity. This kind of threat has been mitigated thanks to the design of the experiment. Our ABBA design is generally prone to carry-over effects. To check for this threat, we statistically analyzed learning and fatigue effects. The results of the two-way ANOVA and permutation tests showed that the effect of Trial was not statistically significant. Another possible threat concerns the exchange of information among the participants. We prevented that monitoring the participants and asking back the material at the end of each trial.

Construct Validity. The used metrics are widely applied in experiments with purposes similarly to ours (e.g., [24]). Regarding the second concern, we evaluated the participants on either the comprehension they achieved on the requirements nor the time they spent to accomplish the tasks.

External Validity. Possible threats could be related to the complexity/simplicity of the comprehension tasks and the choice of the participants. In our experiment, we used a predefined examples which were significantly simpler than requirement specifications from real automotive domain. The real-world requirement specifications can be over 300 pages long and require significant domain and product knowledge to understand. Having this type of complexity would create internal validity threats - low understanding.

Performing experiments with students might also threaten external validity, thus leading to doubts concerning the representativeness of the participants with regard to software professionals. The tasks to be performed in the experiments did not require a high level of industrial experience, so we believed that the use of students as participants could be considered appropriate, as suggested in literature [10, 20]. Working with students also implies various advantages: the students' prior knowledge is rather homogeneous, a large number of participants might be available [39], there is the chance to preliminarily test experimental design and hypotheses [37], and the cognitive complexity of the experimental objects is not hidden by participants' experience.

5. RELATED WORK

SysML introduces new types of diagrams claiming that these new diagrams increase understanding of complex software systems. However, only a few empirical investigations have been conducted to assess the benefits deriving from the use of these diagrams. For example, Nejati et al. [28] presented a framework to facilitate software design inspections conducted as part of the safety certification process. That framework is based on the SysML and includes a traceability information model, a methodology to establish traceability, and mechanisms to use traceability for extracting slices of models relevant to safety requirements. A supporting tool has also been developed [16]. The authors validated their proposal on one benchmark and one industrial case study. Differently, Briand et al. [7] presented the results of a controlled experiment, which has been conducted to assess an approach devised to establish traceability between requirements and SysML models. That approach was conceived to filter out irrelevant details, easing inspection and understanding. The results indicated a significant decrease in completion time and an increase in correctness.

6. FINAL REMARKS AND FUTURE WORK

Our study expands the studies discussed in Section 5 by evaluating comprehensibility of requirements abstracted with SysML requirement diagrams. We opted for controlled experiments because a number of confounding and uncontrollable factors could be present in real project settings, where it may be impossible to control factors such as learning and/or fatigue effects and to select specific tasks. This kind of study reduces failure risks and it is customary conducted in empirical investigation that take place over years (e.g., [3, 33]). Our results suggest that the use of requirement diagrams significantly improves the correctness of understanding without any effect on the task completion time.

Possible future directions for our research will be focussed on the estimation of both the costs and savings the adoption of requirement diagrams might introduce when modeling a computer based system. Then, it would be worth analyzing whether the effort to model requirements is adequately paid back by a more valuable improved comprehension.

Acknowledgments. We thank the participants in the experiments and Giuseppina Casalaro for her help

7. REFERENCES

[1] S. M. Abrahão, C. Gravino, E. I. Pelozo, G. Scanniello, and G. Tortora. Assessing the effectiveness of sequence diagrams in the comprehension of functional requirements: Results from a family of five experiments. *IEEE Trans. on Soft. Eng.*, 39(3), 2013.

[2] J. Aranda, N. Ernst, J. Horkoff, and S. Easterbrook. A framework for empirical evaluation of model comprehensibility. In *Proceedings of the International Workshop on Modeling in Software Engineering*, Washington, DC, USA, 2007. IEEE Computer Society.

[3] E. Arisholm, L. C. Briand, S. E. Hove, and Y. Labiche. The impact of UML documentation on software maintenance: An experimental evaluation. *IEEE Trans. Softw. Eng.*, 32(6):365–381, 2006.

[4] R. Baker. Modern permutation test software. *In E. Edgington, editor, Randomization Tests, Marcel Decker*, 1995.

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

[6] V. R. Basili and H. D. Rombach. The TAME project: Towards improvement-oriented software environments. *IEEE Trans. Software Eng.*, 14(6):758–773, 1988.

[7] L. C. Briand, D. Falessi, S. Nejati, M. Sabetzadeh, and T. Yue. Traceability and sysml design slices to support safety inspections: A controlled experiment. *ACM Trans. Softw. Eng. Methodol.*, 23(1):9, 2014.

[8] B. Bruegge and A. H. Dutoit. *Object-Oriented Software Engineering: Using UML, Patterns and Java, 2nd edition.* Prentice-Hall, 2003.

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

[10] J. Carver, L. Jaccheri, S. Morasca, and F. Shull. Issues in using students in empirical studies in software engineering education. In *Proc. of the International Symposium on Software Metrics*, pages 239–. IEEE Computer Society, 2003.

[11] J. Cohen. *Statistical power analysis for the behavioral sciences (2nd ed.)*. Lawrence Earlbaum Associates, Hillsdale, NJ, 1988.

[12] W. J. Conover. *Practical Nonparametric Statistics*. Wiley, 3rd Edition, 1998.

[13] J. L. Devore and N. Farnum. *Applied Statistics for Engineers and Scientists*. Duxbury, 1999.

[14] B. Dobing and J. Parsons. How UML is used. *Communications of the ACM*, 49(5):109–113, 2006.

[15] P. Ellis. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, 2010.

[16] D. Falessi, S. Nejati, M. Sabetzadeh, L. Briand, and A. Messina. SafeSlice: a model slicing and design safety inspection tool for SysML. In *Proceedings of European conference on Foundations of Software Engineering*, pages 460–463, New York, NY, USA, 2011. ACM.

[17] S. Friedenthal, A. Moore, and R. Steiner. *A Practical Guide to SysML: Systems Modeling Language*. The MK/OMG Press. Elsevier Science, 2008.

[18] O. M. Group. Omg certified systems modeling professional (ocsmp).

[19] O. M. Group. SysML v1.3.

[20] M. Höst, B. Regnell, and C. Wohlin. Using students as subjects: comparative study of students and professionals in lead-time impact assessment. *Empirical Softw. Engg.*, 5(3):201–214, Nov. 2000.

[21] A. Jedlitschka, M. Ciolkowski, and D. Pfahl. Reporting experiments in software engineering. In F. Shull, J. Singer, and D. I. K. Sjøberg, editors, *Guide to Advanced Empirical Software Engineering*, pages 201–228. Springer London, 2008.

[22] N. Juristo and A. Moreno. *Basics of Software Engineering Experimentation*. Kluwer Academic Publishers, Englewood Cliffs, NJ, 2001.

[23] V. Kampenes, T. Dyba, J. Hannay, and I. Sjoberg. A systematic review of effect size in software engineering experiments. *Information & Software Technology*, 49(11-12):1073–1086, 2006.

[24] E. Kamsties, A. von Knethen, and R. Reussner. A controlled experiment to evaluate how styles affect the understandability of requirements specifications. *Information & Software Technology*, 45(14):955–965, 2003.

[25] B. Kitchenham, H. Al-Khilidar, M. Babar, M. Berry, K. Cox, J. Keung, F. Kurniawati, M. Staples, H. Zhang, and L. Zhu. Evaluating guidelines for reporting empirical software engineering studies. *Empirical Software Engineering*, 13:97–121, 2008.

[26] B. Kitchenham, S. Pfleeger, L. Pickard, P. Jones, D. Hoaglin, K. El Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Trans. on Soft. Eng.*, 28(8):721–734, 2002.

[27] H. Levene. Robust tests for equality of variances. In I. Olkin, editor, *Contributions to probability and statistics*. Stanford Univ. Press., Palo Alto, CA, 1960.

[28] S. Nejati, M. Sabetzadeh, D. Falessi, L. C. Briand, and T. Coq. A SysML-based approach to traceability management and design slicing in support of safety certification: Framework, tool support, and case studies. *Information & Software Technology*, 54(6):569–590, 2012.

[29] OMG. Unified Modeling Language (UML) specification, version 2.0. Technical report, Object Management Group, July 2005.

[30] S. L. Pfleeger and W. Menezes. Marketing technology to software practitioners. *IEEE Software*, 17(1):27–33, 2000.

[31] F. Ricca, M. D. Penta, M. Torchiano, P. Tonella, and M. Ceccato. How developers' experience and ability influence web application comprehension tasks supported by uml stereotypes: A series of four experiments. *IEEE Trans. Software Eng.*, 36(1):96–118, 2010.

[32] M. Scaife and Y. Rogers. External cognition: how do graphical representations work? *International Journal of Human-Computer Studies*, 45(2):185–213, 1996.

[33] G. Scanniello, C. Gravino, M. Genero, J. A. Cruz-Lemus, and G. Tortora. On the impact of UML analysis models on source code comprehensibility and modifiability. *ACM Trans. on Soft. Eng. and Meth.*, 23(2), 2014.

[34] G. Scanniello, C. Gravino, and G. Tortora. Investigating the role of UML in the software modeling and maintenance - a preliminary industrial survey. In *Proc. of the International Conference on Enterprise Information Systems*, pages 141–148. SciTePress, 2010.

[35] S. Shapiro and M. Wilk. An analysis of variance test for normality. *Biometrika*, 52(3-4):591–611, 1965.

[36] F. Shull, M. G. Mendoncça, V. Basili, J. Carver, J. C. Maldonado, S. Fabbri, G. H. Travassos, and M. C. Ferreira. Knowledge-sharing issues in experimental software engineering. *Empirical Software Engineering*, 9(1-2):111–137, March 2004.

[37] D. I. K. Sjøberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N. Liborg, and A. C. Rekdal. A survey of controlled experiments in software engineering. *IEEE Trans. on Soft. Eng.*, 31(9):733–753, 2005.

[38] S. Vegas, N. Juristo, A. Moreno, M. Solari, and P. Letelier. Analysis of the influence of communication between researchers on experiment replication. In *Proc. of the International Symposium on Empirical Software Engineering*, pages 28–37, New York, NY, USA, 2006. ACM.

[39] J. Verelst. The influence of the level of abstraction on the evolvability of conceptual models of information systems. In *Proc. of the International Symposium on Empirical Software Engineering*, pages 17–26, Washington, DC, USA, 2004. IEEE Computer Society.

[40] C. Wohlin, P. Runeson, M. Höst, M. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in Software Engineering*. Springer, 2012.

[41] R. Young. *Effective Requirements Practice*. Addison-Wesley, Boston, MA, 2001.