Assessing the Effectiveness of Sequence Diagrams in the Comprehension of Functional Requirements: Results from a Family of Five Experiments

Silvia Abrahão, Carmine Gravino, Emilio Insfran,
Giuseppe Scanniello, *Member*, *IEEE Computer Society*, and Genoveffa Tortora, *Senior Member*, *IEEE*

Abstract—Modeling is a fundamental activity within the requirements engineering process and concerns the construction of abstract descriptions of requirements that are amenable to interpretation and validation. The choice of a modeling technique is critical whenever it is necessary to discuss the interpretation and validation of requirements. This is particularly true in the case of functional requirements and stakeholders with divergent goals and different backgrounds and experience. This paper presents the results of a family of experiments conducted with students and professionals to investigate whether the comprehension of functional requirements is influenced by the use of dynamic models that are represented by means of the UML sequence diagrams. The family contains five experiments performed in different locations and with 112 participants of different abilities and levels of experience with UML. The results show that sequence diagrams improve the comprehension of the modeled functional requirements in the case of high ability and more experienced participants.

Index Terms—Documentation, software engineering, requirements specifications

1 Introduction

Functional requirements are validated to establish whether they provide an accurate account of stakeholders' needs [1]. Adapted from the problem of validating scientific knowledge, requirements validation is the task of making sufficient empirical observations to verify whether a real-world problem is properly captured [2]. To assess this fact, we should verify that the associated models are properly interpreted and understood by stakeholders [3]. An incorrect interpretation and comprehension of models would increase the cost needed to fix them later in the development process [4], [5].

Several approaches for representing functional requirements have been proposed in the past, and of these, behavioral modeling is a common part of those most widely employed [2], [6], [7]. Behavioral modeling involves the modeling of the dynamic and/or functional behavior of users and software systems. A functional model is built in the requirements elicitation phase and is then used during the analysis phase to build the analysis object models and the dynamic models (DM) [3], [8].

In this paper, we present the results of a family of five controlled experiments carried out to investigate whether the use of dynamic models, represented by the Unified Modeling Language (UML) sequence diagrams [9], improves the comprehension of the modeled functional requirements. The original experiment was conducted with a group of computer science undergraduate students from the University of Basilicata (UniBas) in Italy [10].

To further investigate the results of the original experiment, we carried out four replications whose participants had different backgrounds and experience in modeling with UML. These replications were conducted with computer science master's degree students from the University of Salerno (UniSa) in Italy, software engineering master's degree students from the Universitat Politècnica de València (UPV) in Spain [11], PhD students at UPV, and Spanish software professionals. To increase our confidence in the results attained, the latter two replications were based on functional requirements of software systems from different domains.

The paper is structured as follows: In Section 2, we present the family of experiments, while in Section 3 we provide details on the design of the individual experiments, including their definition, the context selection, the hypotheses, and the instrumentation. We report on the results of each experiment and their interpretations in Section 4. In Section 5, we discuss the results of the family of experiments. Possible threats to validity are presented in Section 6. In Section 7, we report on related works, while final remarks conclude the paper.

- *S. Abrahão and E. Insfran are with the Department of Computer Science, Universitat Politècnica de València c/Camino de Vera, s/n, Valencia 46022, Spain. E-mail: {sabrahao, einsfran}@dsic.upv.es.*
- *C. Gravino and G. Tortora are with the Department of Management and Information Technology, University of Salerno, via Ponte Don Melillo, 84084 Fisciano (SA), Italy. E-mail: {gravino, tortora}@unisa.it.*
- *G. Scanniello is with the Dipartimento di Matematica, Informatica, e Economia, University of Basilicata Viale Dell'Ateneo, Macchia Romana, Potenza 85100, Italy. E-mail: giuseppe.scanniello@unibas.it.*

Manuscript received 15 Dec. 2010; revised 14 Nov. 2011; accepted 1 Apr. 2012; published online 26 Apr. 2012.
Recommended for acceptance by B. Cheng.
For information on obtaining reprints of this article, please send e-mail to: tse@computer.org, and reference IEEECS Log Number TSE-2010-12-0372.
Digital Object Identifier no. 10.1109/TSE.2012.27.

2 The Family of Experiments

An increasing understanding exists within the software engineering community that empirical studies are needed to create, improve, or assess processes, methods, and tools for software development [12], [13], [14] and maintenance [15], [16]. An empirical study is generally an act or operation by which to discover something that is unknown or to test hypotheses [17]. Research strategies include controlled experiments, qualitative studies, surveys, and archival analyses [18], [19]. In order to achieve greater validity of the results, replications are necessary [20], [21]. Unfortunately, in software engineering the greater part of empirical studies tend to be isolated and not replicated. For example, Sjøberg et al. [22] observe that 20 out of 113 controlled experiments are replications and, of these, 15 are differentiated replications.[1]

The concept of replication is extended to that of the "family of experiments" reported by Basili et al. [23]. A family is composed of multiple similar experiments that pursue the same goal to build the knowledge needed to extract significant conclusions.

In this section, we present the family of experiments that we performed. The method adopted is an extension of the five steps proposed by Ciolkowski et al. [24], in which the fifth step, "Family data analysis," has been replaced with "Family data analysis and meta-analysis."

2.1 Step 1: Experiment Preparation

The goal of our study is to analyze the use of dynamic models (represented by UML sequence diagrams) with the purpose of assessing whether they improve the comprehension of the modeled functional requirements for different categories of users in terms of experience and ability. Two are the perspectives of the family: from the researcher's point of view, to investigate the support provided by sequence diagrams in the comprehension of functional requirements, and from the project manager's point of view, to evaluate the possibility of adopting these models to enhance requirements representations.

2.2 Step 2: Context Definition

The groups of participants were:

- undergraduate students in computer science with some knowledge of the UML,
- master's students in computer science, many of them being professionals,
- PhD students in computer science, who have attended various software engineering courses,
- software industry professionals, who design and develop software systems.

Regarding the participants' ability, the students were classified according to the average grades attained in their academic degrees. As suggested in [25], for Italian students, a Grade Point Average (GPA)[2] less than or equal to 24 is considered to be low; otherwise, it is high. For Spanish students, an average below 9/10 is considered to be low; otherwise, it is high. Since in the literature there are no similar controlled experiments conducted in Spain, we selected this threshold according to the Spanish experimenters' experience. Due to several differences in the organization of Italian and Spanish universities (e.g., *in Spain the selection procedure of the master's students limits access to only a few selected students*), the use of the same converted threshold of the original experiment was practically impossible. For the replication conducted with professionals, we used a similar threshold based on the grades that they had achieved in the specialization course "Modeling with the UML" at UPV.

All the students were given one point in their final grades, regardless of their performances. The students' and professionals' participation in the experiments was voluntary.

2.3 Step 3: Material and Experimental Tasks

The material was composed of two experimental objects (containing a set of models with an attached comprehension questionnaire), the training material, and a postexperiment survey questionnaire. The experimental objects were the models of functional requirements of an e-commerce system and other typical management information systems (MIS) (e.g., course management). The comprehension questionnaire included multiple-choice questions. The questionnaires used in the replications conducted in Spain were translated from Italian into English and then into Spanish. The training materials included: 1) a set of slides containing an introduction to both the UML diagrams considered in the experiments (i.e., class diagrams and sequence diagrams) and the use case template employed, along with examples of UML models; 2) a set of slides describing the procedure to be followed in the experiments; 3) comprehension tasks related to UML models. The experimental material and the raw data are available for download at www.scienzemfn.unisa.it/scanniello/RE_Exp1/.

At the end of each experiment, the participants were asked to fill in a postexperiment survey questionnaire. The questionnaire aimed at gaining sufficient insight to strengthen and explain the results.

2.4 Step 4: Individual Experiments

Fig. 1 summarizes the family of experiments. The rectangles represent the experiments and are grouped by the experimental objects used. The figure also shows the execution order of the experiment (e.g., first experiment), the kind of replication (e.g., external replication), the participants involved and their number, and the name associated with each experiment (e.g., Italy 1).

The second and third experiments (Italy 2 and Spain 1) were differentiated replications of the original experiment (i.e., Italy 1) in different settings with different participants. In particular, Italy 2 was an *internal* replication (the same experimenters as the original one), while Spain 1 was an *external* replication (conducted by different experimenters to verify whether the results were independent of both the experimenters and the experimental setting). In Italy 2 and Spain 1, the same experimental objects as in Italy 1 were employed. The fourth and fifth experiments (Spain 2 and 3) were differentiated replications whose goal was to repeat the first experiment with different participants and experimental objects. An independent group of experimenters conducted these experiments (*external* replications).

1. Differentiated replications introduce variations in essential aspects of the experimental conditions. One prominent variation concerns the executions of replications with different kinds of participants.

2. In Italy, the exam grades are expressed as integers and assume values between 18 and 30. The lowest grade is 18, while the highest is 30.

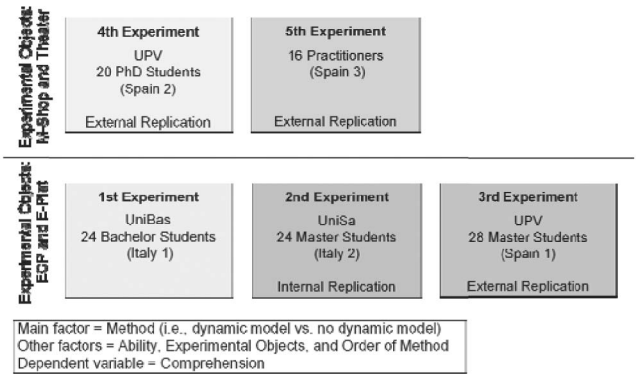


Fig. 1. Experiments in the family.

2.5 Step 5: Family Data Analysis and Meta-Analysis

The results of each individual experiment and the family of experiments were collected and analyzed. We also performed a meta-analysis to aggregate the results since the experimental conditions were very similar for each experiment. We performed a meta-analysis to obtain stronger results with regard to the contribution of sequence diagrams to comprehend functional requirements.

3 Design of Individual Experiments

In this section, we present the design of the experiments according to the guidelines for experimental software engineering proposed by Juristo and Moreno [18] and Wohlin et al. [19]. With regard to the replications, we discuss only their differences with respect to the original experiment, thus avoiding useless redundancies. We conclude the section discussing issues related to the documentation used in the external replications and the communication among the experimenters.

3.1 The Original Experiment

The functional requirements were modeled by employing the method suggested in [8], in which a requirements analysis document includes a functional model, an analysis object (or conceptual) model, and a dynamic model. A functional model focuses on the software functionality, while an analysis object model focuses on the individual concepts of the problem domain that will be manipulated by the software system. Finally, a dynamic model concerns the software system behavior of the meaningful use cases presented in the functional model. In this study, use case diagrams were employed to describe the functional model, while class diagrams and sequence diagrams were used to describe analysis object model and dynamic model, respectively.

3.1.1 Planning

Context. The two experimental objects were selected from the requirements analysis specifications of the following two systems:

- **ECP**—An e-commerce platform from which CDs and books can be ordered via the Internet from an online catalogue.
- **E-Plat**—A software system for the management of courses, lecturers, and students of a university.

We randomly selected the functional requirement "Processing an order" for the first system (ECP in the following) and then used the corresponding models. The functional requirement "Adding a course" and its models were similarly selected for the second system (E-Plat in the following). Fig. 2 shows samples of the ECP models used in the experimentation.

To assess the complexity of the models used and to identify possible mistakes, we carried out a pilot experiment with a master's student at UniBas. From the pilot, we deduced that the complexity of the models was comparable because the student spent almost the same time on each of the two tasks using sequence diagrams (30 minutes on average).

A group of 24 computer science undergraduate students at UniBas was involved as participants. These students were enrolled in the software engineering course at UniBas held in the first semester, from September 2007 to January 2008. One of the main topics of that course is the modeling of object-oriented systems using the UML. The participants also had experience in object-oriented programming and web technology. During the course, the students were grouped in teams, and each team was involved in the modeling and development of a software system. Two young researchers were responsible for coordinating the projects, which were based on requirements analysis, on the high-level design of the software system, and on an incremental development of the subsystems identified. For the experiment, the students' participation was voluntary.

Hypotheses formulation. We formulated the following null hypothesis, which is one-sided since we expected a positive effect of sequence diagrams on the comprehension of the models:

- H_{n0}: The comprehension of functional requirements *does not significantly improve* when participants are provided with models that include UML sequence diagrams.

In the case of rejection of the null hypothesis with relatively high confidence, it is possible to formulate the alternative hypothesis that admits a positive effect:

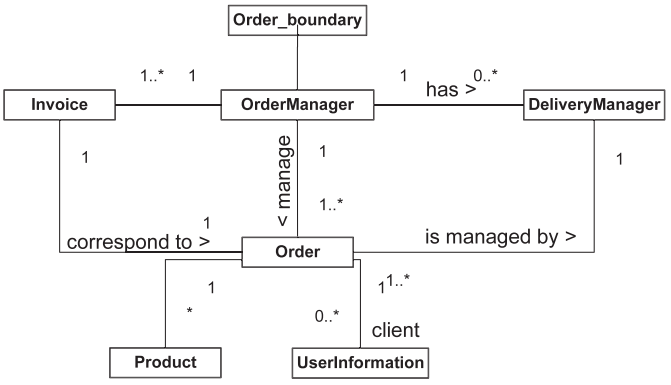
- H_{a0}: The comprehension of functional requirements *significantly improves* when participants are provided with models that include UML sequence diagrams.

The participants who performed the comprehension tasks with NO_DM received as documentation a functional model and an analysis object model. The participants who performed the tasks with DM also received a dynamic model.

The *Control Group* is "requirements specification *without* dynamic models abstracted using sequence diagrams," while the *Treatment Group* is "requirements specification *with* dynamic models abstracted using sequence diagrams." Therefore, the only independent variable (or main factor) is Method, which is a nominal variable that can assume two possible values: Dynamic Model (DM) and NO Dynamic Model (NO_DM).

To test the null hypothesis, we considered a measure based on the comprehension of the models representing functional requirements. The comprehension questionnaire consisted of multiple-choice questions. Each question

Use case: OrderProcessing
ID: 3
Description: User creates the order for one or more products. The products are delivered after the user pays for them
Actor: User
Flow of events: 1. User selects the functionality "Processing an Order" 2. System shows a form to the User 3. User inserts the information about the product/s related to the order created 4. When the information about the product/s is inserted, User submits the order information. 5. System creates the order and notifies User. 6. User pays for the order. 7. System starts the delivery process of the products. 8. System notifies that the payment has been correctly accepted.
Post-conditions: 1. The order has been paid for 2. The products are ready to be delivered.



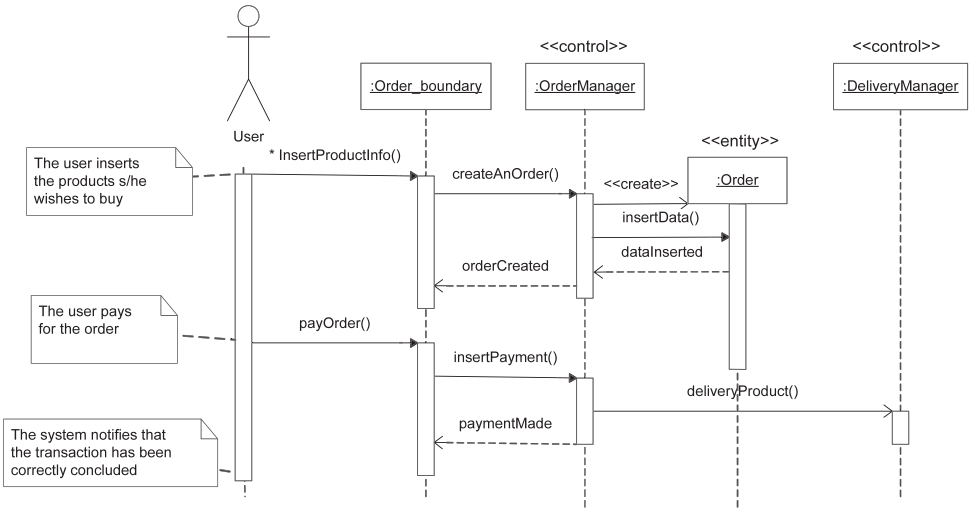


Fig. 2. Samples of ECP models.

admitted one or more correct answers among a set of four. The number of possible answers was the same for each question. The purpose of this questionnaire was to rapidly evaluate various comprehension aspects of the models and then the abstracted functional requirements. A sample question of the comprehension questionnaire is shown in Fig. 3. The correct answer is "Create an order" and can be easily derived from both the use case and the sequence diagrams shown in Fig. 2.

We expected that the stakeholders' comprehension of the models would increase when they were provided with additional information (sequence diagrams in our case). However, stakeholders' ability and experience, and the way in which the sequence diagrams are built (e.g., the level of detail and the use of complex functionality of the notation) may affect their comprehension of the requirements. Therefore, here we are interested in investigating whether the comprehension is affected by the stakeholders' ability and experience, along with whether there is some interaction between the participants' ability and the main factor under study. In order to evaluate the participants' comprehension of the models, we assessed the answers they provided to the comprehension questionnaire in terms of correctness and completeness [25]. The usage of these evaluation criteria was possible since each question in the comprehension questionnaire admitted one or more correct answers. We measured correctness and completeness by using an information retrieval-based approach [26]. In particular, we used an aggregate formulation of recall and precision [27], [28]:

$$recall_s = \frac{\sum_i |answers_{s,i} \cap correct_i|}{\sum_i |correct_i|}, \quad (1)$$

$$precision_s = \frac{\sum_i |answers_{s,i} \cap correct_i|}{\sum_i |answer_{s,i}|}, \quad (2)$$

Q3. The end-user of the system can*:
- ☐ Add a delivery.
- ☐ Modify an order.
- ☐ Modify the data of an invoice.
- ☐ Create an order.

*Mark the right answer/s

Fig. 3. A sample question for ECP.

TABLE 1
Design of the Controlled Experiments

	Groups			
	A	B	C	D
Task1	ECP, DM	ECP, NO_DM	E-Plat, DM	E-Plat, NO_DM
Task2	E-Plat, NO_DM	E-Plat, DM	ECP, NO_DM	ECP, DM

where $answer_{s,i}$ is the set of responses that the participants s provided for the question i, while $correct_i$ is the set of correct responses expected for this question.

In order to obtain a single measure that shows the comprehension achieved by the participants, we used the harmonic mean of precision and recall [26], called F-measure, which allowed us to obtain a balance between correctness and completeness:

$$F\text{-}measure_s = 2 * \frac{precision_s * recall_s}{precision_s + recall_s}. \quad (3)$$

To test the null hypothesis, the following dependent variable was considered:

- **Comprehension**: The F-measure of the precision and recall values of all the questions of the comprehension questionnaire. A value close to 1 means that the participant answers were very good, while a value close to 0 means very bad.

Since we were also interested in assessing the effect of ability on the comprehension of the models, the Ability ordinal variable was additionally considered and its effect was also analyzed. The values that the Ability cofactor can assume are Low and High.

Experiment design. We adopted a counterbalanced design by dividing the participants into four groups, namely, A, B, C, D (see Table 1). The participants in each group were asked to perform the tasks without interacting with each other. The two tasks were performed sequentially, with an interval of 15 minutes. The chosen design mitigates possible learning effects.

We used a prequestionnaire to gather some information about the participants (e.g., the industrial working experience and the GPA) to assess their ability level and experience. Then, using the GPA, we split the participants into the four groups, where the High- and Low-ability participants were equally distributed.

Other factors to be controlled. Other factors (also denominated as cofactors) may influence the results. In particular, different extraneous factors may have an undesirable effect on the comprehension of the modeled functional requirements, and this effect may be confused with the effect of Method:

- **Object**. The complexity of the models used and the participants' familiarity with the application domain of the systems may affect the comprehension of the functional requirements.
- **Order of method**. The order in which the participants perform the tasks may produce learning effects, thus biasing the results.

3.1.2 Operation

Preparation and execution. The experiment took place in a single room. All the participants attended an introductory lesson in which detailed instructions on the tasks to be performed were presented. Only the goal of the experiment was highlighted, and details of the experimental hypothesis were not provided. We informed the participants about the procedure to follow in the execution of the experiment. The participants were told that their answers would be treated anonymously, and were also informed (with the exception of the participants of Spain 3) that their grade for the course would not be affected by their performance in the experiment. We additionally informed the participants (including those of Spain 3) of the pedagogical purpose of the exercises. Before starting the introductory lesson, we provided the participants with a set of instructional slides and a set of slides describing the experimental procedure.

TABLE 2
The Postexperiment Survey Questionnaire

Id	Question	Answers
Q1	I had enough time to perform the tasks.	(1-5)
Q2	The task objectives were perfectly clear to me.	(1-5)
Q3	The tasks I performed were perfectly clear to me.	(1-5)
Q4	Judge the difficulty of the task in the e-commerce system.	(A-E)
Q5	Judge the difficulty of the task in the system for the management of courses, lecturers, and students.	(A-E)
Q6	Assess your experience level in analysis object modeling and the UML class diagrams.	(A-E)
Q7	Assess your experience level in dynamic modeling and the UML sequence diagrams.	(A-E)
Q8	What is the most useful method with which to comprehend and interpret functional requirements?	DM, NO_DM, Neutral
Q9	Which models did you perceive as the most useful to comprehend functional requirements?	Functional, Analysis Object, Dynamic
1 = Strongly agree, **2** = Agree, **3** = Neutral, **4** = Disagree, **5** = Strongly disagree		
A = Very high, **B** = High, **C** = Medium, **D** = Low, **E** = Very low		

After the introductory lesson, the participants started with the experiment. The experiment's execution was controlled, no interaction between participants was allowed, and no time limit was imposed. We asked the participants to accomplish the two tasks with a break of 15 minutes between them. Each participant received only the material needed to accomplish the task in hand. No further information or clarifications on the models were provided during the execution of each task.

After the comprehension tasks, the participants were asked to fill in the postexperiment survey questionnaire shown in Table 2. It includes questions to assess the overall quality of the material provided, the perceived usefulness of the models, the clearness of the tasks, and the goals of the experiment. The postexperiment survey questionnaire was mostly the same for all the experiments in the family. The only difference was for Q4 and Q5, which for Spain 2 and 3 concerned different tasks.

Analysis procedure. We used statistical tests, interaction plots [29], and clustered bar charts [19] to analyze the data collected. Table 3 summarizes the analyses performed. In particular, we tested the null hypothesis by using the nonparametric Wilcoxon test [30]. We used this test since a statistical difference between two-dependent groups (paired analysis) was under investigation. In contrast, to analyze the effect of Ability, we performed an unpaired analysis by applying the nonparametric Mann-Whitney test [30].

TABLE 3
Analyses Performed

Factor/Cofactors and Their Interaction	Investigation
Method	Wilcoxon test
Ability	Mann–Whitney test
Method vs. Ability	Interaction Plot
Object	Wilcoxon test
Order of Method	Mann–Whitney test
Responses to the post experiment survey questionnaire	Clustered bar charts

The counterbalanced experimental design also permitted us to analyze the interaction of Ability with the main factor (i.e., Method). We used interaction plots[3] to study the presence of such a possible interaction.

The Wilcoxon test was employed to analyze the effect of Object on the comprehension of the models. To test the effect of Order of Method, we used a method similar to that proposed by Briand et al. [31]. In particular, let: *Diff(-NO_DM)* be the differences in comprehension of those participants who performed the tasks using NO_DM first and then DM, and *Diff(DM)* be the differences in comprehension of those participants who performed the tasks using DM first and then NO_DM. Unlike Briand et al., we used the nonparametric Mann-Whitney test to verify whether Diff(NO_DM) is greater than Diff(DM). The rationale behind the Order of Method analysis lies in the fact that the capability of using sequence diagrams may improve during the tasks. As a result, we verified the null hypothesis $H_{0d} : \mathrm{Diff(NO\_DM)} = \mathrm{Diff(DM)}$. The alternative hypothesis was H_{ad}: $\mathrm{Diff(NO\_DM)} > \mathrm{Diff(DM)}$.

We adopted nonparametric tests because of the sample size and the mostly nonnormality of the data. Both these tests have been chosen because they are very robust and sensitive, and have been used in experiments similar to ours in the past, e.g., [25], [31]. As usual, in all the tests, we decided to accept a probability of 5 percent of committing a Type-I-Error [19], i.e., of rejecting the null hypothesis when it is actually true. Statistical tests allow the presence of a significant difference between dependent or independent groups to be verified, but they do not provide any information about this difference. Therefore, we used the Cohen's "d" [32] to obtain the standardized difference between two groups that can be considered *negligible* for $|d| < 0.2$, *small* for $0.2 \leq |d| < 0.5$, *medium* for $0.5 \leq |d| < 0.8$, and *large* for $|d| \geq 0.8$. In the context of paired analyses, we used the difference between the means of the distributions divided by the standard deviation of the (paired) differences between the samples, while in the context of unpaired analyses we used the difference between the means of the distributions divided by the pooled standard deviation [25].

To graphically show the answers of the postexperiment survey questionnaire, we adopted clustered bar charts. These are widely employed since they provide a quick visual representation to summarize data.

3. Interaction plots are simple line graphs in which the means of the dependent variables for each level of one factor are plotted over all the levels of the second factor. The resulting lines are parallel when there is no interaction and nonparallel when interaction is present.

3.2 Italy 2

The participants were 24 students from a computer science master's degree program at UniSa, who were enrolled in the advanced software engineering course. The experiment was conducted as part of a series of laboratory exercises carried out within this course. The course was held in the second semester (from March 2008 to June 2008). All the students of Italy 2 had received a bachelor's degree in computer science from UniSa.

The participants had a reasonable level of technical maturity and knowledge of UML-based, object-oriented software development, and software project management. The students were familiar with concepts of the requirements engineering process and had experience in supervising teams of developers.

3.3 Spain 1

The participants were 33 students enrolled in a master's degree program in software engineering, formal methods, and information systems at UPV. They were asked to perform the experiment as a part of a series of optional laboratory exercises conducted within the software engineering with the models course. This course was held in the first semester, from September 2008 to January 2009, and was selected because it is a specialized teaching unit in which students learn techniques regarding conceptual modeling, the UML, and model-driven engineering.

3.4 Spain 2

This experiment varied the manner in which Italy 1 was run with the intention of increasing confidence in the experimental results by testing the same hypothesis as before, but altering the details of the experiment to increase external validity. Different experimental objects were selected from the following two systems:

- **M-Shop**—A software system for managing the sales in a music shop.
- **Theater**—A software system for managing a theater's ticket reservations.

We randomly selected the functional requirement "Search Album by Singer" for the first system (M-Shop in the following). For the second system (Theater in the following), we selected "Buy Theater Ticket." We defined the questions in the comprehension questionnaires for the M-Shop and Theater models such that their complexity could be considered comparable with the ECP and E-Plat questions.

As for Italy 1, we conducted a pilot experiment. We involved a research fellow from UniSa. The results indicated that 45 minutes was sufficient to accomplish each task. The time needed for the tasks, the size of the models of the experimental objects, and the experience of the participant enabled us to deduce that the objects were more complex than those used in Italy 1, 2, and Spain 1.

The participants were 20 PhD students enrolled in the software engineering, formal methods and information systems PhD program at UPV. The experiment was conducted in the course "Software Engineering with Models" as part of a series of laboratory exercises. The course was held in the first semester (from September 2008 to February 2009).

TABLE 4
Descriptive Statistics for the Comprehension*-Dependent Variable

Exp.	DM			NO_DM		
	Med.	Mean	Std. Dev.	Med.	Mean	Std. Dev.
Italy 1	0.625	0.633	0.198	0.630	0.612	0.159
Italy 2	0.685	0.673	0.115	0.565	0.526	0.089
Spain 1	0.458	0.423	0.172	0.353	0.356	0.111
Spain 2	0.727	0.727	0.088	0.591	0.618	0.166
Spain 3	0.636	0.631	0.122	0.546	0.534	0.119

The values for Comprehension range between 0 and 1, where $0 = null$ comprehension and $1 = perfect$ comprehension.

3.5 Spain 3

Here, we employed the same design and experimental objects as used in Spain 2. The participants involved were 16 professionals from Spanish software development companies. The professionals had various backgrounds and various levels of experience in modeling with UML. They primarily worked as software analysts and programmers.

The professionals previously attended the specialization course "Modeling with the UML," which was aimed at teaching the principles of object-oriented analysis and design. They were classified as Low- or High-ability participants based on the grades achieved in that course. In particular, they were classified as Low ability for grades less than 9/10 and High otherwise. The professionals' participation was voluntary.

3.6 Documentation and Communication

Issues such as documentation [33] and communication between experimenters [34] may influence the success of replications. Shull et al. in [33] discuss how deficiencies in laboratory packages and documentation are one of the weak points within the discipline that makes it difficult to use replication to advance knowledge. As a possible solution, the authors propose better laboratory packages and the use of knowledge sharing mechanisms.

With regard to documentation, the experimenters of the original experiment translated into English all the material, initially written in Italian. This material included the postexperiment survey questionnaire, the comprehension questionnaire, the data collection forms, and the UML models. The experimenters of the external replications then translated this material into Spanish. The definition, planning, and experimental operation were provided in another document, which also described the script of the comprehension tasks. In the document, we also discussed the rationale behind the design choices taken in the original experiment highlighting all the information useful to reproduce the experimental conditions. The experimenters involved in the external replications were also provided with a previous publication concerning the original experiment [10]. The groups of experimenters additionally exchanged the training material to reproduce the same experimental setting used in Italy 1.

Although documentation is a key factor in being able to carry out a replication, communication among experimenters is even more important [34]. The interaction between the groups of experimenters started with an initial face-to-face meeting where the main ideas behind Italy 2 and Spain 1 were discussed. The experimenters produced and shared minutes of the meeting. To assist the experimenters in all the phases of the study, e-mail communication and virtual meetings were used. The experimenters occasionally also used instant messaging tools.

4 Results and Discussion

In this section, we discuss the results of each individual experiment by analyzing: the influence of the method, the effect of Ability, Objects and Order of Method, and the results of the postexperiment survey questionnaire.

4.1 Influence of Method

Table 4 shows the descriptive statistics for the Comprehension variable. The descriptive statistics show that the mean comprehension scores obtained for the participants when using sequence diagrams were superior to those obtained when not using them. Upon analyzing the comprehension mean scores, we can observe that the participants from Spain 1 had lower scores than the other participants. One possible reason for this result might be that the students involved in Spain 1 were less trained in behavioral modeling than the other participants.

As mentioned in Section 3.1.2 and shown in Table 3, we used the paired Wilcoxon test to study the effect of Method and the Cohen's "d" to obtain the standardized difference between two groups. Table 5 summarizes the results for each experiment. The results of the Wilcoxon test revealed that H_{n0} can be rejected for all the experiments with the exception of Italy 1. Table 5 also shows the number of participants who benefitted from DM (# of $DM > NO\_DM$), those who achieved the same results with both DM and NO_DM (# of $DM = NO\_DM$), and those who obtained better comprehension scores with NO_DM (# of $DM < NO\_DM$). For example, 11 out of 24 participants of Italy 1 benefitted from sequence diagrams, while 10 participants did not benefit from them.

The results of the replications diverge from those we found in the original experiment. One possible reason for this difference might be that the participants in the original experiment were less experienced with UML.

TABLE 5
The Wilcoxon Test Results for Each Experiment in the Family

Exp.	#obs	Hypothesis Rejected? (p-value)	Effect Size (Cohen's d)	# of DM > NO_DM	# of DM = NO_DM	# of DM < NO_DM
Italy 1	48	No (0.26)	Negligible (0.11)	11/24	3/24	10/24
Italy 2	48	**Yes (<0.001)**	Large (0.99)	20/24	1/24	3/24
Spain 1	56	**Yes (0.027)**	Small (0.43)	16/28	6/28	6/28
Spain 2	40	**Yes (0.017)**	Medium (0.55)	12/20	3/20	5/20
Spain 3	32	**Yes (0.01)**	Medium (0.64)	10/16	4/16	2/16

TABLE 6
Influence of Ability on the Comprehension-Dependent Variable Grouping Observations by Method

Experiment	DM	NO_DM
Italy 1	No (0.203)	No (0.610)
Italy 2	No (0.264)	No (0.726)
Spain 1	No (0.252)	**Yes** (0.041)
Spain 2	No (0.640)	No (0.440)
Spain 3	No (1.00)	No (0.502)

4.2 Influence of Ability

The results of the Mann-Whitney test when applied to the observations grouped by Method and Ability are shown in Table 6.

Italy 1—The results of the original experiment showed that Ability did not significantly influence Comprehension. In addition, the plot in Fig. 4 shows that the lines are almost parallel, and High-ability participants always achieved higher scores, regardless of whether or not they used sequence diagrams.

Italy 2—The results of the Mann-Whitney test revealed that Ability had no significant effect on Comprehension. However, the interaction plot in Fig. 4 suggests that there was interaction between Method and Ability. Although the Low-ability participants achieved better comprehension scores than the High-ability ones when using sequence diagrams, this difference can be considered as meaningless (the difference is less than 0.048). This result and the fact that the High-ability participants achieved better comprehension when they did not use sequence diagrams suggest that these models appear to fill the gap between the Low- and High-ability participants. It can also be noted that both High- and Low-ability participants benefitted from sequence diagrams.

Spain 1—The Mann-Whitney test showed that Ability affected the comprehension scores when sequence diagrams were not used. In particular, Low-ability participants achieved better comprehension with a large effect size. In addition, the interaction plot in Fig. 4 indicates that some interaction between Method and Ability was present. High-ability participants achieved better results when using sequence diagrams. It would thus appear that High-ability participants benefitted more from sequence diagrams than Low-ability participants. Conversely, Low-ability participants achieved better results when sequence diagrams were not used. The performance of High-ability participants may be motivated by their habit of using sequence diagrams. The lack of these diagrams may cause the search of tentative interpretations of functional requirements.

Spain 2—The results of the data analysis revealed that Ability had no significant effect on Comprehension. Furthermore, the interaction plot in Fig. 4 indicates an interaction between Method and Ability. In particular, Low-ability participants obtained better results than High-ability ones when performing tasks without sequence diagrams, while High-ability participants achieved better results when using sequence diagrams. These results are similar to those obtained in Spain 1, except that the effect of Ability is not

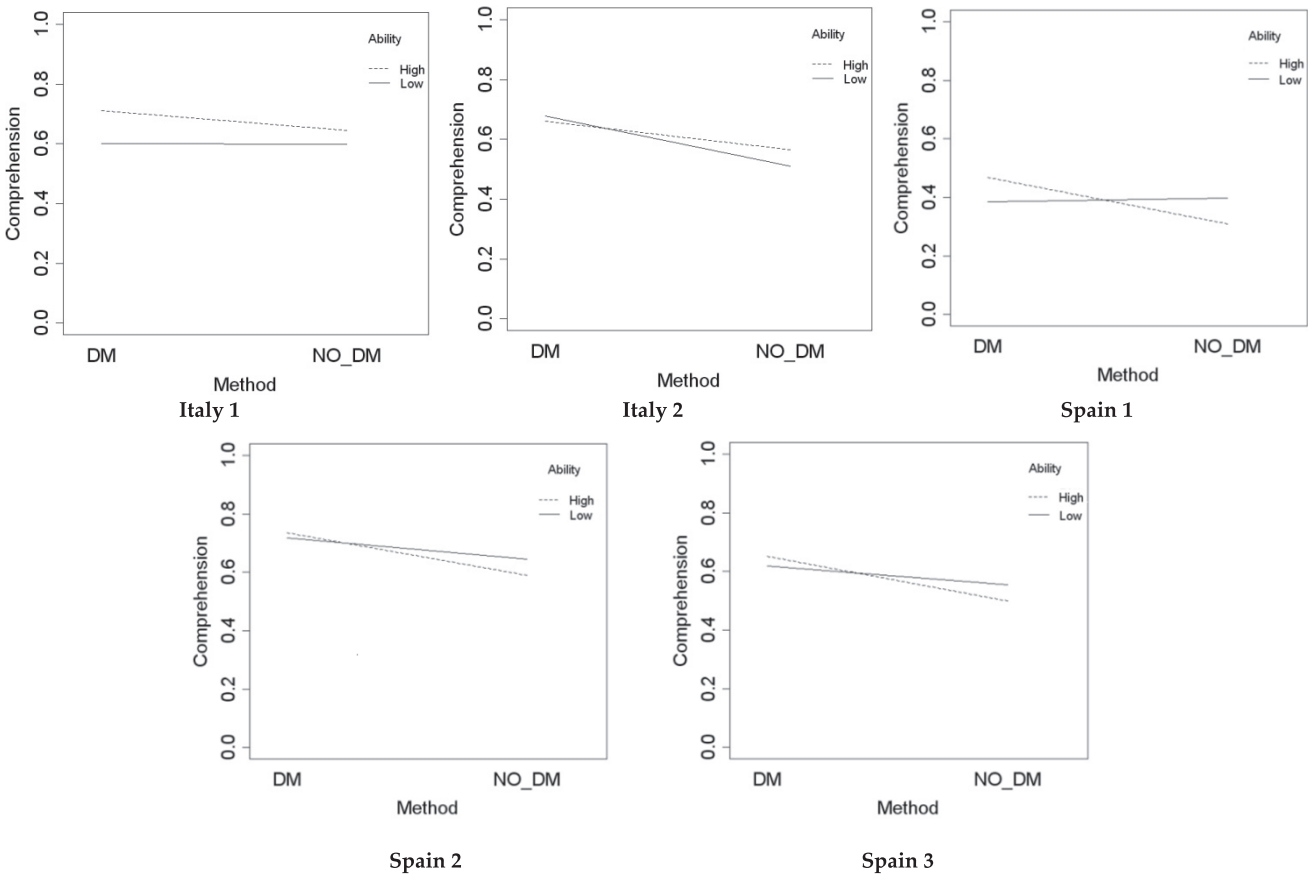


Fig. 4. Interaction between Method and Ability.

TABLE 7
Summary of the Results with Regard to Object and Order of Method

Experiment	Object	Order of Method
Italy 1	NO (0.637)	No (0.823)
Italy 2	NO (0.679)	No (0.941)
Spain 1	**YES (<0.001)**	No (0.306)
Spain 2	NO (0.177)	No (0.859)
Spain 3	NO (0.344)	No (0.479)

statistically significant for NO_DM. As for Spain 1, the High-ability participants might have done erroneous interpretations of the requirements due to the lack of the sequence diagrams. Unlike Spain 1, both the High- and Low-ability participants benefitted from sequence diagrams.

Spain 3—The results of the Mann-Whitney test showed that Ability did not influence the results attained. The interaction plot in Fig. 4 shows that there is an interaction between Method and Ability. Low-ability participants obtained better results than High-ability ones when performing comprehension tasks without sequence diagrams, while High-ability participants achieved better results when using sequence diagrams. These results are coherent with those obtained in Spain 1 and 2 since High-ability participants benefitted more from sequence diagrams than Low-ability participants. Unlike Spain 1, the participants benefitted from sequence diagrams.

4.3 Influence of Object and Order of Method

Table 7 summarizes the results of the statistical analysis conducted to assess the effect of Object and Order of Method. The results of the Wilcoxon test indicated that the effect of Object was not statistically significant in any of the experiments except for Spain 1. In this experiment, the participants who performed the comprehension task on E-Plat achieved a significantly better comprehension of the software models with a large effect size. This result might have been caused by the perceived complexity of E-Plat and ECP and the participants' familiarity with the problem domain of the systems. A more exhaustive analysis of Object is shown in the appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TSE.2012.27.

To verify the influence of Order of Method, we tested the null hypothesis H_{0d} by applying the Mann-Whitney test, which revealed that this hypothesis could not be rejected in the original experiment (p-value = 0.823). The same conclusion was proven to hold in all the replications. Therefore, we can conclude that the participants in all the experiments did not have a significantly greater Comprehension score in the second task.

4.4 Survey Questionnaire Results

The responses provided to the questions in the postexperiment survey questionnaire (see Table 2) are visually summarized in Fig. 5. The figure also shows the responses to each question. We grouped together the questions that admit the same possible answers.

We analyzed the responses to the postexperiment survey questionnaires of Italy 1, 2, and Spain 1 together since they are based on the same experimental objects. Similarly, we analyzed the responses of Spain 2 and 3, together.

The analysis of the responses to question Q1 showed that the majority of the participants strongly agreed that the time needed to carry out the experiments was appropriate, thus suggesting that they did not perceive time pressure. The responses to Q2 and Q3 revealed that the objectives and the tasks were considered clear in all the cases even if seven participants out of 28 in Spain 1 responded *Disagree* to Q3.

With regard to Q4, Italy 1 and 2 participants answered *Medium*. These participants also answered *Medium* to Q5, suggesting that the difficulty of both the tasks was medium and comparable. Conversely, 16 of the 28 participants to Spain 1 answered *High* for Q4, while 14 responded *Medium* to Q5. Therefore, they perceived the task on E-Plat to be more difficult than that on ECP. This result partially justifies the data analysis results presented in Section 4.2.

The majority of the participants of Italy 1, 2, and Spain 1 answered *Medium* to Q6, namely, the statement used to detect the participants' experience level in analysis object modeling and class diagrams. The majority of the participants of Italy 1 indicated that their knowledge of dynamic modeling and the UML sequence diagrams was medium. For Italy 2, the responses to Q7 were equally distributed between *High* and *Medium*. Conversely, for Spain 1, a few participants (9 out of 24) responded *Low* to Q7.

The responses to Q8 for the first three experiments revealed that the participants generally considered DM to be more useful. With regard to Italy 1, this is a case in which a controlled experiment provides insight into the difference between the perceived usefulness of a given method and the effective advantage of using it.

Finally, the participants found dynamic modeling to be more useful in Italy 1 and 2 (14 and 17, respectively). The functional modeling was considered to be more useful in the case of Spain 1 (13 out of 28).

With regard to Spain 2 and 3, a huge number of participants strongly agreed on the fact that the time needed to carry out the experiments was appropriate (see bars of Q1 for Spain 2 and 3 in Fig. 5). The responses to Q2 revealed that the task objectives were considered to be clear in all cases. The majority of the participants answered *Strongly Agree* to this statement (10 out of 20 for Spain 2 and nine out of 16 for Spain 3). Similar results were achieved in Q3 for Spain 2. The responses to Q3 were not concordant for Spain 3: Seven participants out of 16 responded *Agree*, while six responded *Disagree*.

The responses to Q4 in Spain 2 indicated that the participants mostly answered *High* (10 out of 20) or *Medium* (eight out of 20), thus suggesting that they judged the difficulty of the comprehension tasks for M-Shop to be *Medium/High*. The majority of the participant of Spain 2 answered *Medium* to Q5, while the participants of Spain 3 responded either *High* or *Medium*.

The responses to Q6 suggest that the participants' knowledge of analysis object modeling and the UML class diagrams was mostly medium (12 out of 20 for Spain 2 and eight out of 16 for Spain 3). We obtained similar results for Q7. The participants judged their experience with dynamic modeling and sequence diagrams to be *Medium* or *Low*.

The responses to Q8 revealed that the participants generally considered DM to be more useful in the execution of the comprehension tasks. For Spain 2 and 3, 18 and 13 participants answered DM, respectively.

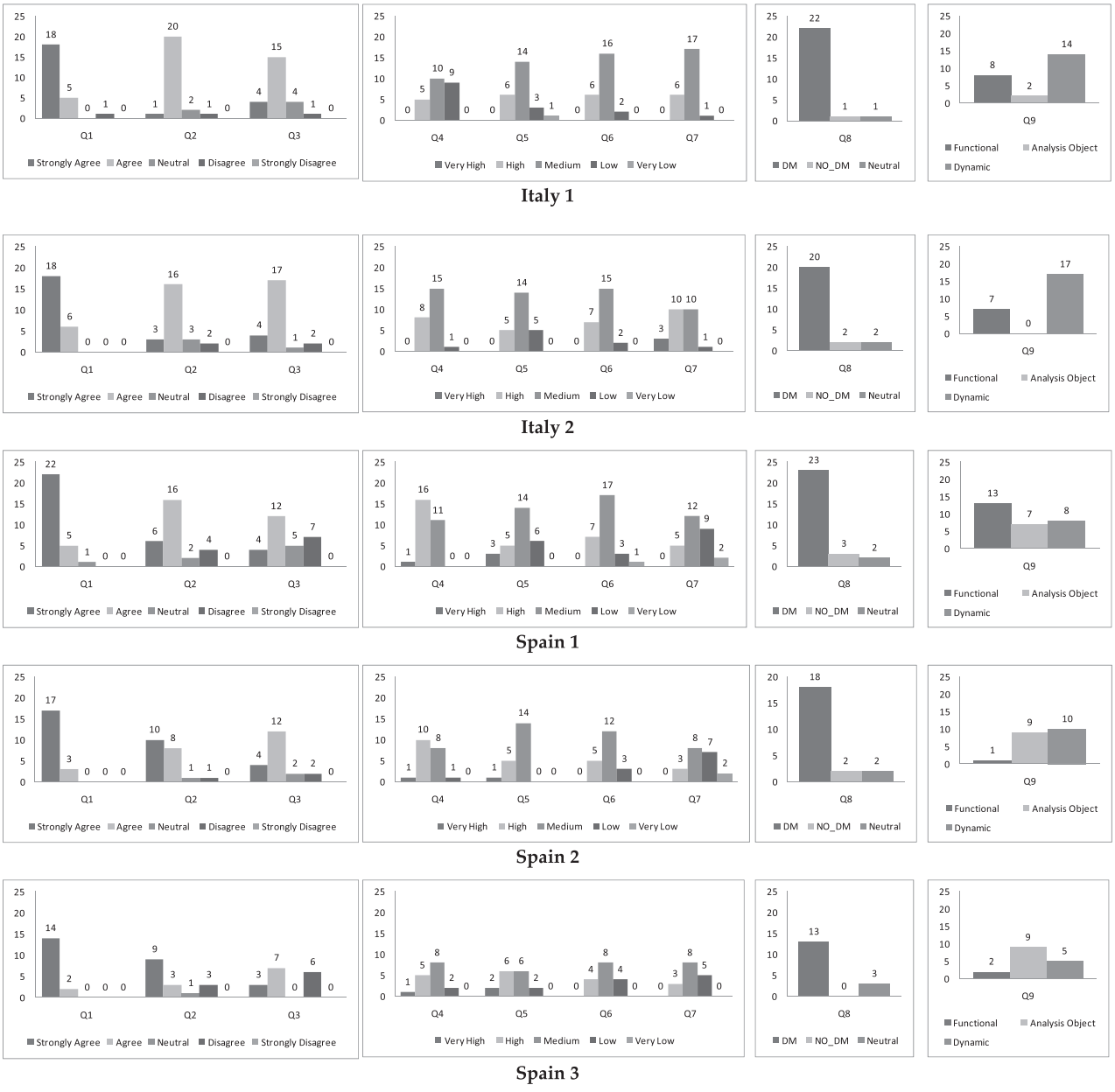


Fig. 5. Participants' responses.

Finally, the responses to Q9 revealed that, for Spain 2, the participants generally considered object and dynamic modeling to be equally useful in the execution of comprehension tasks (nine out of 20 answered *Analysis Objects* and 10 out of 20 *Dynamic*). Object modeling was considered to be more useful for Spain 3.

5 Family Data Analysis

In this section, we provide a summary of the results obtained. We first present an analysis of the results in the context of the family of experiments, followed by the results of a meta-analysis that aggregates the empirical findings obtained in each experiment.

5.1 Summary of Results

A summary of the experiments and their results is provided in Table 8. The main result of the family of experiments indicates that support was found for hypothesis H_{a0} in all the experiments, with the exception of the original one. This indicates that the use of sequence diagrams significantly improved the comprehension of the models representing functional requirements when participants have an adequate level of experience (i.e., at least a Bachelor's degree in Computer Science). The results of the family of experiments also open up a number of issues:

- **Comprehension.** The comprehension mean scores of the Spain 1 participants were lower than those of the other experiments (see Table 4). This concern might be due to the students' low familiarity with the modeling method used in the study (i.e., behavioral modeling). Moreover, the comprehension mean scores obtained by the Spanish PhD students and professionals are comparable to those of the Italian students. This could have been caused by the

TABLE 8
Summary of the Experiments

Exp.	Context	Description	Type of replication	# of participants			Objects	Hypothesis Rejected?	Method vs. Ability	Influence of co-factors
				All	Ability					
					High	Low				
Italy 1	Bachelor Students at UniBas	Original experiment	–	24	7	17	ECP E-Plat	No	Yes	No
Italy 2	Master's Students at UniSa	Different environment and participants	Internal	24	7	17	ECP E-Plat	Yes	Yes	No
Spain 1	Master's Students at UPV	Different environment and participants	External	28	13	15	ECP E-Plat	Yes	Yes	Influence of Ability on NO_DM
Spain 2	PhD Students at UPV	Replication that varies the manner in which Italy 1 is run	External	20	10	10	M-Shop Theater	Yes	Yes	No
Spain 3	Spanish Professionals	Replication that varies the manner in which Italy 1 is run	External	16	10	6	M-Shop Theater	Yes	Yes	No

complexity of the experimental objects used in Spain 2 and 3 that are more complex than those used in the other experiments.

- **Participants' experience.** The Spanish students and professionals and the Italian students from Italy 2, had more experience than those of Italy 1. This finding, together with the results of each experiment, indicates that more experienced participants obtained greater benefit from sequence diagrams.
- **Influence of ability.** The interaction plots provided further information useful in improving our understanding of what type of participants (High or Low ability) will benefit from the use of sequence diagrams. In particular, all the experiments indicated that there was some interaction between Method and Ability. Furthermore, the experimental results indicated that the High-ability participants obtained better results than the Low ability ones, with the exception of Italy 2, in which a slight difference in favor of Low-ability participants was observed. This result may indicate the presence of a possible ability threshold that the participants should have, signifying that High-ability participants benefit from the use of sequence diagrams. As for Italy 2, the difference between Low- and High-ability participants when using sequence diagrams can be considered as meaningless (the difference is less than 0.048), deducing that all the participants benefitted from the dynamic models.

The experiments allowed us to gather knowledge concerning the conditions under which the use of sequence diagrams is more effective. According to the previously discussed results, we can conclude that High ability and more experienced participants benefit more from sequence diagrams than Low ability and less experienced ones. The results also showed that the Object complexity did not influence the participants' comprehension, with the exception of Spain 1 (see Section 4.3).

In summary, the results support the hypothesis that the comprehension of functional requirements significantly improves when a stakeholder is provided with class diagrams and use cases together with sequence diagrams. Running a family of experiments rather than a single experiment provided us with more evidence of the external validity, and thus the generalization of the study results. The same hypothesis was tested and confirmed (with very few exceptions) in five different environments using different experimental objects and four types of participants. Each replication provided further evidence of the confirmation of the hypothesis. Thus, we can conclude that the general goal of the empirical validation has been achieved.

5.2 Meta-Analysis

Although several statistical methods exist for aggregating and interpreting results obtained from interrelated experiments [35], [36], [37], [38], we used meta-analysis because it allowed us to get more general conclusions. Meta-analysis is a set of statistical techniques for combining the different effect sizes of the experiments. The estimation of effect sizes can be used after comparing studies to evaluate the average impact across studies of an independent variable on the dependent one. Since measures may come from different settings and may be nonhomogeneous, a standardized measure must be obtained for each experiment: These measures must be combined to estimate the global effect size of a factor. In our study, we considered that Method was the main factor in the family of the experiments.

The meta-analysis was performed by using the Meta-Analysis v2 tool [39]. As reported in [40], we employed the mean value obtained using sequence diagrams (i.e., m_1) minus the mean value achieved when not using them (i.e., m_2) to calculate the effect sizes for Comprehension for each of the individual experiments, and from these values, we obtained the Hedges' g metric [36], [41], which was used as a standardized measure. This measure expresses the magnitude of the effect of Method.

To obtain the overall conclusion, we calculated the Z-score based on the mean and standard deviation of the Hedges' g statistics of the experiments. We used correlation coefficients which provided the effect sizes that have a normal distribution (z_i) once they had been transformed by the Fisher transformation [42]. The global effect size is obtained by using the Hedges' g metric, with the weights proportional to the experiment size:

$$\bar{Z} = \frac{\sum_i w_i z_i}{\sum_i w_i}, \tag{4}$$

where $w_i = 1/(n_i - 3)$ and n_i is the sample size of the ith experiment. The higher the value of Hedges' g, the higher the corresponding correlation coefficient is.

Table 9 summarizes the results of the meta-analysis: For each experiment, it reports the effect size, the values of the Hedges' g metric, and its significance. The effect size is

TABLE 9
The Hedges' g Metric Values
for the Comprehension-Dependent Variable

Experiment	Effect Size (Hedges' g)	Significance
Italy 1	Small (0.092)	No (p=0.518)
Italy 2	Medium (0.563)	**Yes (p<0.001)**
Spain 1	Small (0.260)	No (p=0.053)
Spain 2	Small (0.341)	**Yes (p=0.033)**
Spain 3	Medium (0.423)	**Yes (p=0.019)**
Global Effect Size	Small (0.319)	**Yes (p<0.001)**

rated *small* (0 to 0.37), *medium* (0.38 to 1), or *large* (above 1) [41], depending on the standardized difference between the two means m_1 and m_2. For example, an effect size of 0.5 indicates that $m_1 = m_2 + (0.5^*d)$, where d is the standard deviation, i.e., a positive value means that sequence diagrams improve the comprehension of the models measured by the dependent variable defined.

Fig. 6 shows the forest plot (or blobbogram) as provided by the tool used. On the left-hand side, the experiments are reported in chronological order from the top downward. On the right-hand side, the effect of Hedges' g metric is plotted for each experiment by a square whose dimensions are proportional to the weight of the experiment in the meta-analysis. The estimations for studies with a large sample size are more accurate, signifying that they make a greater contribution to the overall effect. The square is proportional to the number of participants and its position with regard to the x-axis indicates the Hedges' g value. The confidence intervals of each experiment are represented by horizontal lines. Here, we have considered a confidence interval of 95 percent for each experiment. The confidence interval $[-1, 0]$ indicates a negative correlation, whereas the confidence interval $[0, 1]$ indicates a positive correlation. The diamond in the last row represents the overall conclusion. The summary measure is the center line of the diamond, while the associated confidence interval is the lateral tips of the diamond.

The effect size obtained varies between small and medium in all the cases. Despite the fact that the first and third experiments (i.e., Italy 1 and Spain 1) did not produce significant results (see Table 9), the overall results of the meta-analysis present a significant positive effect; thus, we can reject the null hypothesis, namely, "*the comprehension of functional requirements does not significantly improve when participants are provided with models that include sequence diagrams.*" The meta-analysis therefore strengthens the alternative hypothesis, which can be easily derived.

6 Threats to Validity

6.1 Internal Validity

The threats to internal validity are relevant in those studies that attempt to establish a causal relationship. In our case, the participants were students and professionals with varying knowledge and backgrounds. The Italy 1 participants were less experienced in the UML than those of the replications. However, all the participants found the experimental tasks to be clear, as the results of the postexperiment survey questionnaire revealed.

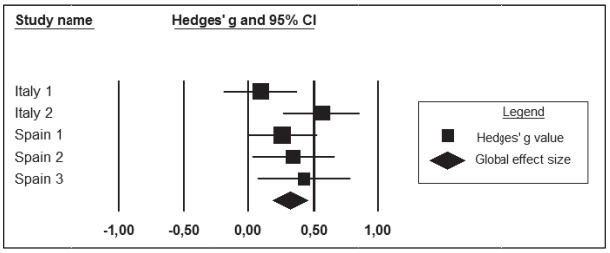


Fig. 6. Comprehension of the models meta-analysis.

A key question in the threat to internal validity is whether the differences observed are due to the learning effect. This fact was mitigated in the experiments because each participant worked on different experimental objects in the two tasks. We also assessed the effect of Order of Method using statistical tests.

Another issue concerns the possible information exchanged among the participants. Supervisors monitored them to avoid communication biases both while performing the tasks and during the time interval between the two tasks. Further, the participants were asked to give back all the material at the end of each task.

With regard to Italy 2 and Spain 1, a different threat to internal validity was present. It was possible that the participants in these experiments might have been able to obtain information on the tasks from the Italy 1 participants. However, they did not know the participants of Italy 2 and Spain 1 because they resided in different regions and countries. With regard to Spain 2 and 3, the participants could not exchange information because we conducted the experiments in two different contexts and the participants did not know each other.

Finally, the selection of different objects in the study may have affected the instrumentation validity and thus biased the results. We mitigated this threat by conducting pilot experiments to assess both the complexity of the objects and to attempt to identify mistakes in the experimental material.

6.2 External Validity

External validity refers to the approximate truth of conclusions involving generalizations within different contexts. When designing the experiments, we took into account possible threats to validity coming from the involvement of students as the participants [43], [44]. Our first concern was to select groups of participants widely representative of software professionals. The participants had knowledge of UML and dynamic modeling as is the case with the majority of young software professionals working in small/medium companies in Italy and Spain. With regard to the participants of Italy 1, there is not a great difference between them and professionals [23], [45]. In fact, these kinds of students will soon be integrated into the software industry market, so they can be considered as widely representative of young professionals [46]. Similar considerations are possible on the participants of Italy 2 and Spain 1. The participants of Spain 2 are not far from software professionals, as the experimental results also suggested.

A threat that might affect the external validity concerns the size and complexity of the tasks used. We decided to use relatively small tasks since a controlled experiment requires that participants complete the assigned tasks in a

limited amount of time. To confirm or contradict the achieved results, we plan to conduct case studies with larger and more complex tasks.

6.3 Construct Validity

The construct validity may have been influenced by the measures used to obtain a quantitative evaluation of the comprehension achieved on the experimental objects, the comprehension questionnaires, and the postexperiment survey questionnaire. A proper design of the experiments allowed us to mitigate these threats. Note that the measures used have been employed in several previous controlled experiments [25], [28], [47].

The comprehension questionnaire might represent a threat to construct validity because its author knew the hypothesis and the diagrams of the experimental objects. We defined the questions of this questionnaire to avoid affecting the results in favor of DM or NO_DM. They were also formulated to be simple but not obvious, and of comparable complexity. The questions expected multiple-choice answers to reduce ambiguities and the time required for completion. Another by-product of this choice was in the evaluation of participants' comprehension that could be computed in a repeatable manner. Supervisors collected the questionnaires at the end of each task. In each experiment, two researchers worked together to analyze the answers and compute the precision, recall, and F-measure values.

The use of a single-dependent variable could have biased the results. Another possible threat could concern the choice of classifying participants in the two classes High and Low. This choice is widely used in similar studies (e.g., [25]) and is also related to the sample size.

The postexperiment survey questionnaire was designed to capture the participants' perception of the tasks. Our objective was principally to support and explain the quantitative results of the experiments by providing qualitative insight from the questionnaire data. We designed this questionnaire using standard methods and scales [48].

To avoid evaluation apprehension, the students were not graded on the results they obtained. Moreover, the participants were not aware of the experimental hypothesis. Ease of comprehension/interpretation of the models was the sole criterion examined as it represents a key issue for establishing agreement among different stakeholders.

6.4 Conclusion Validity

Threats to conclusion validity concern the issues that affect the ability to draw a correct conclusion. In this study, threats to conclusion validity concern the selection of the participants, the data collection, the measurement reliability, and the validity of the statistical tests. With regard to the selection of the populations, we drew fair samples and conducted our experiments with participants belonging to these samples. F-measure allowed us to assess in an objective way the comprehension achieved by the participants. We chose the Wilcoxon and Mann-Whitney nonparametric statistical test for their robustness and sensitivity [30].

7 Related Work

In this section, we discuss the related literature concerning empirical studies aimed at: 1) assessing the use of UML behavior diagrams, 2) evaluating the influence of ability and experience in comprehension tasks with regard to the use of UML. A systematic literature review concerning empirical evaluations on the models and forms used in UML can be found in [49].

7.1 UML Behavioral Diagrams

Otero and Dolado [50] presented a comparison of sequence, collaboration, and statechart diagrams. The empirical study revealed that the comprehension of dynamic models generally depended on the diagram type and on the complexity of the functionality. They also found that software design documents were more comprehensible when sequence diagrams were used to model dynamic behavior, coherently with our results. In a subsequent study [51], the same authors observed that the specification of the dynamic behavior using the Open Modeling Language was faster to comprehend and easier to interpret than UML.

Glezer et al. [52] performed a controlled experiment to investigate the comprehensibility and quality of sequence and collaboration diagrams in two application domains: Management Information Systems (MIS) and Real-Time (RT) systems. The results indicated that collaboration diagrams were easier to comprehend than sequence diagrams in the RT systems, but not in the case of the MIS systems. From this result and from the findings presented in this paper, we derive the suggestion to compare the use of sequence and collaboration diagrams in the various phases of the software development process. Moreover, Glezer et al. observed that the kind of system used in the experimental object affected the comprehension of the models. Several studies could be necessary to investigate the effect of the problem/solution domain on the comprehensibility of UML models. The main difference with respect to our study is that the authors did not focus on models produced in the early phases of the development process.

Nugroho et al. [53] presented an empirical study to investigate the relationship between the level of detail in UML models and the defect density of the corresponding implementation. The study was conducted within an industrial context and revealed that classes modeled in sequence diagrams with a higher level of detail present a lower number of defects. The authors conducted a case study with professional programmers instead of a controlled experiment, and confounding factors were not controlled [19].

In the studies discussed so far, the contribution of the UML sequence diagram in the comprehension of models produced in the requirements engineering process is not considered and the effect of ability and experience of the involved participants is not analyzed.

7.2 Participants' Ability and Experience in the Comprehension of the UML

Briand et al. [31] established that training is required to achieve better results when the UML diagrams are complemented with the use of Object Constraint Language (OCL) [54]. The focus of the experiment is on the contribution that OCL provides in increasing the comprehension of some UML diagrams (i.e., class, sequence, and statechart diagrams). They found that OCL improved an engineer's ability to understand, inspect, and modify a system. Similarly to our study, the authors found some interaction between the main factor and the participants' ability.

Ricca et al. [25] presented the results of four experiments carried out to assess the effectiveness of the UML stereotypes [55] in the design of web applications. The experiments involved participants with different levels of experience and ability. In contrast with our results, the participants with a low ability achieved significant benefits from the use of the considered notation (i.e., stereotypes), while participants with a high ability obtained a comparable comprehension with or without the notation. The authors thus concluded that the use of stereotypes reduces the gap between novice and experienced software engineers.

As for the study presented here, the analysis of participants' ability and experience is crucial to gaining a deeper understanding of the benefit deriving from the use of a notation rather than from another.

8 Conclusion

We have reported the results of a family of five experiments conducted to assess whether dynamic models represented in terms of UML sequence diagrams improve stakeholders' comprehension when dealing with functional requirements. We carried out the experiments focusing on the comprehension, without considering any time/efficiency evaluation, in different locations, and with participants including students and professionals of different abilities and levels of experience.

The context of the original experiment was a group of computer science undergraduate students from the University of Basilicata. The results showed that the participants judged the use of sequence diagrams to be useful, although the effect of sequence diagrams was not statistically significant. Possible reasons for this may be that the participants were familiar with the domain of the specifications employed or that the participants had no adequate previous experience in modeling with UML. To verify these findings, we carried out four replications of this experiment with professionals and more experienced students.

The results of the replications revealed that the sequence diagrams significantly improved the comprehension of the modeled functional requirements. A meta-analysis confirmed this result with stronger evidence. A possible explanation is that more experienced and high-ability participants benefit more when information is structured as in the sequence diagrams. We plan to investigate this aspect in the future.

Further investigations could concern: 1) empirical studies involving several groups of professionals with different levels of experience, 2) experiments to assess the effect of providing the information to the participants in an incremental way, 3) experiments to analyze the effect of different behavior diagrams in the comprehension of software models. Another interesting aspect to be investigated will be regarding the comprehension of nonfunctional requirements, for which several methodologies could be necessary.

Acknowledgments

The authors wish to thank all the participants in the experiments. This research was partially supported by the MULTIPLE project (with ref. TIN2009-13838).

References

[1] A. Finkelstein, "Requirements Engineering: An Overview," *Proc. Asia-Pacific Software Eng. Conf.*, 1993.
[2] M. Jackson, *Software Requirements and Specifications: A Lexicon of Practice, Principles and Prejudices.* Addison Wesley, 1995.
[3] B. Nuseibeh and S. Easterbrook, "Requirements Engineering: A Roadmap," *Proc. Conf. Future of Software Eng.*, pp. 35-46, 2000.
[4] B.W. Boehm, *Software Engineering Economics.* Prentice-Hall, 1981.
[5] T. Nakajo and H. Kume, "A Case History Analysis of Software Error Cause-Effect Relationships," *IEEE Trans. Software Eng.*, vol. 17, no. 8, pp. 830-838, Aug. 1991.
[6] A. Davis, *Software Requirements: Objects, Functions and States.* Prentice Hall, 1993.
[7] R.J. Wieringa, *Requirements Engineering: Frameworks for Understanding.* Wiley, 1996.
[8] B. Bruegge and A. Dutoit, *Object-Oriented Software Engineering Using UML, Patterns, and Java.* Prentice Hall, 2004.
[9] OMG. Unified Modeling Language (UML) Specification, version 2.0, 2005.
[10] C. Gravino, G. Scanniello, and G. Tortora, "An Empirical Investigation on Dynamic Modeling in Requirements Engineering," *Proc. Int'l Conf. Model Driven Eng. Languages and Systems,* pp. 615-629, 2008.
[11] S. Abrahão, E. Insfran, C. Gravino, and G. Scanniello, "On the Effectiveness of Dynamic Modeling in UML: Results from an External Replication," *Proc. Third Int'l Symp. Empirical Eng. and Measurement,* pp. 468-472, 2009.
[12] V.R. Basili, R.W. Selby, and D.H. Hutchens, "Experimentation in Software Engineering," *IEEE Trans. Software Eng.*, vol. 12, no. 7, pp. 733-743, July 1986.
[13] V.R. Basili, "The Role of Experimentation in Software Engineering: Past, Current, and Future," *Proc. Int'l Conf. Software Eng.*, pp. 442-449, 1996.
[14] N. Fenton, "How Effective Are Software Engineering Methods?" *J. Systems and Software,* vol. 22, no. 2, pp. 141-146, 1993.
[15] M. Colosimo, A. De Lucia, G. Scanniello, and G. Tortora, "Evaluating Legacy System Migration Technologies through Empirical Studies," *Int'l J. Information and Software Technology,* vol. 51, no. 12, pp. 433-447, 2009.
[16] W.J. Dzidek, E. Arisholm, and L.C. Briand, "A Realistic Empirical Evaluation of the Costs and Benefits of UML in Software Maintenance," *IEEE Trans. Software Eng.*, vol. 34, no. 3, pp. 407-432, May/June 2008.
[17] V.R. Basili, "The Experimental Paradigm in Software Engineering," *Proc. Int'l Workshop, Experimental Software Eng. Issues: Critical Assessment and Future Directions,* 1993.
[18] N. Juristo and A.M. Moreno, *Basics of Software Engineering Experimentation.* Kluwer Academic Publishers, 2001.
[19] C. Wohlin, P. Runeson, M. Host, M.C. Ohlsson, B. Regnell, and A. Wesslen, *Experimentation in Software Engineering—An Introduction.* Kluwer, 2000.
[20] F. Shull, J.C. Carver, S. Vegas, and N. Juristo, "The Role of Replications," *Empirical Software Eng.*, vol. 13, no. 2, pp. 211-218, 2008.
[21] B. Kitchenham, "The Role of Replications in Empirical Software Engineering—A Word of Warning," *Empirical Software Eng.*, vol. 13, no. 2, pp. 219-221, 2008.
[22] 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. Software Eng.*, vol. 31, no. 9, pp. 733-753, Sept. 2005.
[23] V.R. Basili, F. Shull, and F. Lanubile, "Building Knowledge through Families of Experiments," *IEEE Trans. Software Eng.*, vol. 25, no. 4, pp. 456-473, July/Aug. 1999.
[24] M. Ciolkowski, F. Shull, and S. Biffl, "A Family of Experiments to Investigate the Influence of Context on the Effect of Inspection Techniques," *Proc. Sixth Int'l Conf. Empirical Assessment in Software Eng.*, pp. 48-60, 2002.
[25] F. Ricca, M. Di 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.*, vol. 36, no. 1, pp. 96-118, Jan./Feb. 2010.
[26] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval.* Addison-Wesley, 1999.

[27] G. Antoniol, G. Canfora, G. Casazza, A. De Lucia, and E. Merlo, "Recovering Traceability Links between Code and Documentation," *IEEE Trans Software Eng.*, vol. 28, no. 10, pp. 970-983, Oct. 2002.
[28] T. Zimmermann, P. Weissgerber, S. Diehl, and A. Zeller, "Mining Version Histories to Guide Software Changes," *IEEE Trans. Software Eng.*, vol. 31, no. 6, pp. 429-445, June 2005.
[29] J.L. Devore and N. Farnum, *Applied Statistics for Engineers and Scientists.* Duxbury Press, 1999.
[30] W.J. Conover, *Practical Nonparametric Statistics,* third ed. Wiley, 1998.
[31] L. Briand, Y. Labiche, M. Di Penta, and H. Yan-Bondoc, "An Experimental Investigation of Formality in UML-Based Development," *IEEE Trans. Software Eng.*, vol. 31, no. 10, pp. 833-849, Oct. 2005.
[32] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences,* second ed. Lawrence Earlbaum Assoc., 1988.
[33] F. Shull, M.G. Mendonça, V.R. Basili, J. Carver, J.C. Maldonaldo, S. Fabbri, G.H. Travassos, and M.C. Ferreira, "Knowledge-Sharing Issues in Experimental Software Engineering," *Empirical Software Eng.*, vol. 9, nos. 1/2, pp. 111-137, 2004.
[34] S. Vegas, N. Juristo, A.M. Moreno, M. Solari, and P. Letelier, "Analysis of the Influence of Communication between Researchers on Experiment Replication," *Proc. Int'l Symp. Empirical Software Eng.*, pp. 28-37, 2006.
[35] G.V. Glass, B. McGaw, and M.L. Smith, *Meta-Analysis in Social Research.* Sage Publications, 1981.
[36] L.V. Hedges and I. Olkin, *Statistical Methods for Meta-Analysis.* Academia Press, 1985.
[37] R. Rosenthal, *Meta-Analytic Procedures for Social Research.* Sage Publications, 1986.
[38] J.A. Sutton, R.K. Abrams, R.D. Jones, A.T. Sheldon, and F. Song, *Methods for Meta-Analysis in Medical Research.* John-Wiley & Sons, 2001.
[39] Biostat, Comprehensive Meta-Analysis v2, 2006.
[40] J.A. Cruz-Lemus, M. Genero, M.E. Manso, S. Morasca, and M. Piattini, "Assessing the Understandability of UML Statechart Diagrams with Composite States—A Family of Empirical Studies," *Empirical Software Eng.*, vol. 14, no. 6, pp. 685-719, 2009.
[41] V. Kampenes, T. Dybå, J.E. Hannay, and D.I.K. Sjøberg, "A Systematic Review of Effect Size in Software Engineering Experiments," *Information and Software Technology,* vol. 49, nos. 11/12, pp. 1073-1086, 2007.
[42] R.A. Fisher, "Frequency Distribution of the Values of the Correlation Coefficient in Samples of an Indefinitely Large Population," *Biometrika,* vol. 10, no. 4, pp. 507-521, 1915.
[43] J. Carver, L. Jaccheri, S. Morasca, and F. Shull, "Issues in Using Students in Empirical Studies in Software Engineering Education," *Proc. Int'l Software Metrics Symp.*, pp. 239-249, 2003.
[44] J.E. Hannay and M. Jørgensen, "The Role of Deliberate Artificial Design Elements in Software Engineering Experiments," *IEEE Trans. Software Eng.*, vol. 34, no. 2, pp. 242-259, Mar./Apr. 2008.
[45] M. Höst, B. Regnell, and C. Wholin, "Using Students as Subjects—A Comparative Study of Students and Professionals in Lead-Time Impact Assessment," *Proc. Fourth Conf. Empirical Assessment and Evaluation in Software Eng.*, pp. 201-214, 2000.
[46] B.A. Kitchenham, S. Pfleeger, D.C. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary Guidelines for Empirical Research in Software Engineering," *IEEE Trans. Software Eng.*, vol. 28, no. 8, pp. 721-734, Aug. 2002.
[47] B. Anda, K. Hansen, I. Gullesen, and H.K. Thorsen, "Experiences from Using a UML-Based Development Method in a Large Safety-Critical Project," *Empirical Software Eng.*, vol. 11, no. 4, pp. 555-581, 2006.
[48] A.N. Oppenheim, *Questionnaire Design, Interviewing and Attitude Measurement.* Pinter Publishers, 1992.
[49] D. Budgen, A.J. Burn, O.P. Brereton, B. Kitchenham, and R. Pretorius, "Empirical Evidence about the UML: A Systematic Literature Review," *Software: Practice and Experience,* vol. 41, no. 4, pp. 363-392, 2011.
[50] M.C. Otero and J.J. Dolado, "An Initial Experimental Assessment of the Dynamic Modelling in UML," *Empirical Software Eng.*, vol. 7, no. 1, pp. 27-47, 2002.
[51] M.C. Otero and J.J. Dolado, "An Empirical Comparison of the Dynamic Modeling in OML and UML," *J. Systems and Software,* vol. 77, no. 2, pp. 91-102, 2005.
[52] C. Glezer, M. Last, E. Nachmany, and P. Shoval, "Quality and Comprehension of UML Interaction Diagrams: An Experimental Comparison," *Information and Software Technology,* vol. 47, no. 10, pp. 675-692, 2005.
[53] A. Nugroho, B. Flaton, and M.R.V. Chaudron, "Empirical Analysis of the Relation between Level of Detail in UML Models and Defect Density," *Proc. Int'l Conf. Model Driven Eng. Languages and Systems,* pp. 600-614, 2008.
[54] OMG, Object constraint language (OCL) specification, version 2.0, 2005.
[55] J. Conallen, *Building Web Applications with UML.* Addison-Wesley Object Technology Series, 1999.

Silvia Abrahão is an associate professor at the Universitat Politècnica de València, Spain. She has served as a member of the management committee of two European COST actions: "Towards the Maturity of IT Usability Evaluation" and "Towards the Integration of Transectorial IT Design and Evaluation." She has been awarded several grants to support her research work; in particular, she has been a visiting researcher at the Software Engineering Institute, Carnegie Mellon University, the Belgian Laboratory of Computer-Human Interaction, and Guent University. Currently, she is the Spanish editor for the ERCIM News magazine and serves regularly on the program committees of several international conferences and workshops. Her research interests include quality assurance in software product lines, model-driven development, integration of usability in software development processes, web quality, and empirical software engineering.

Carmine Gravino received the Laurea degree in computer science (cum laude) in 1999 and the PhD degree in computer science from the University of Salerno, Italy, in 2003. Since March 2006, he has been an assistant professor in the Department of Mathematics and Informatics, University of Salerno. His research interests include software metrics and techniques to estimate web application development effort, software-development environments, design pattern recovery from object-oriented code, evaluation and comparison of notations, methods, and tools supporting software development and maintenance. He has published more than 60 papers on these topics in international journals, books, and conference proceedings.

Emilio Insfran received the MS degree in computer science from Cantabria University, Spain, in 1994 and the PhD degree from the Universitat Politècnica de València, Spain, in 2003. He is an associate professor in the Department of Information Systems and Computation, Universitat Politècnica de València, Spain, and a member of the Software Engineering and Information Systems Research Group. He was a visiting research scientist in the Department of Computer Science, University of Twente, The Netherlands, in 1999, and in the Department of Information Systems, Brigham Young University, Utah, in 2001. His research interests are requirements engineering, model-driven software development, software quality, and software engineering environments and tools. He has published more than 40 journal and conference papers and has been working on a number of national and international projects.

Giuseppe Scanniello received the Laurea and PhD degrees, both in computer science, from the University of Salerno, Italy, in 2001 and 2003, respectively. In 2006, he joined the Department of Mathematics and Computer Science, University of Basilicata, Potenza, Italy, where he is currently an assistant professor and leads the Software Engineering Group. His research interests include requirements engineering, empirical software engineering, reverse engineering, reengineering, workflow automation, migration, wrapping, integration, e-learning, global software engineering, cooperative supports for software engineering, and visual languages. He has published more than 80 papers in international journals, books, and proceedings of refereed conferences. He serves on the organizing and program committees of several international conferences in the field of software engineering and is a member of the IEEE Computer Society.

Genoveffa Tortora has been a full professor in computer science at the University of Salerno, since 1990, where she has been department chair and then dean of the Faculty of Sciences. She has been (1993-1999) the vice-president of GRIN (Gruppo di Informatica), the Italian Association of University Professors of Computer Science. She has been general chair, program chair, and a program committee member of several international conferences and serves on the editorial boards of several scientific journals. She has been responsible for several national research projects, and has been an evaluator and scientific committee member of European and national projects. She has coauthored more than 200 papers published in scientific journals or proceedings of refereed conferences, and has coedited three books. At the University of Salerno, she founded and directed the Department of Mathematics and Informatics and the Software Engineering, GIS, VR, and Visual Computing Labs. Her research interests are in the software engineering and information systems areas, and include software development environments, human-computer interaction, visual languages, databases and geographic information systems, image processing ,and biometric systems. She is a senior member of the IEEE and a member of the ACM and the IAPR.