On the Effectiveness of the UML Object Diagrams: A Replicated experiment

Giuseppe Scanniello  
Dipartimento di Matematica e Informatica  
Università degli Studi della Basilicata  
giuseppe.scanniello@unibas.it

Filippo Ricca  
DISI, Università di Genova  
Genova, Italy  
filippo.ricca@disi.unige.it

Marco Torchiano  
Politecnico di Torino  
Torino, Italy  
marco.torchiano@polito.it

Background: In the modeling of object oriented software systems, the UML object diagrams are recognized very useful to complement class diagrams. However, up to now, there exists only one experiment [Torchiano 2004] that investigates this concern.

Aim: To confirm or contradict the findings of the original experiment, we have conducted a replication and the achieved results have been presented in this paper. Both the replication and the original experiment have been conducted to investigate whether the use of object diagrams to complement class diagrams affects the comprehension of software systems.

Method: The replication has been conducted with a group of 24 graduated subjects in Computer Science of the University of Basilicata. The experiment adopts a counterbalanced design, thus ensuring that each subject work on two comprehension tasks, experimenting each time class and object diagrams together or class diagrams alone. The comprehension on each task has been assessed using a questionnaire-based approach. In particular, we have measured the comprehension level of each subject using an information retrieval based approach that allowed us to get a balance between correctness and completeness of the answers.

Results: The results show that the subjects significantly benefit from the use of object diagrams in the comprehension of software systems, thus confirming and strengthening the findings of the original experiment.

Conclusions: It is advisable to complement the usual class diagrams with object diagrams to increase the understandability of software systems. To raise the generalizability of the results, replications of this study are necessary especially with professional software engineers.

1. INTRODUCTION

The software engineering community has developed a number of methods and approaches to model software systems. Most of them use the Unified Modeling Language (UML) [OMG 2010] as a notation to represent structural and behavioral system properties. Regarding the structural properties, UML provides six different diagrams: class, object, deployment, package, component, and composite structure.

The most widely employed UML structural diagrams are probably the class and the object diagrams. In particular, the former can be employed to model the structure of a software system by showing its classes, the attributes of each class, and the relationships between the classes of the system. Differently, object diagrams can be used to provide a complete or partial view of the structure of a software system at run-time, at a specific instant in terms of object instances, attributes values, and links between the instances.

In the modeling of object oriented software systems, the object diagrams are recognized useful to complement class diagrams; in particular, to explain the mining of a class diagram [Fowler and Kendall 2010]. This is especially true in the early phases of the development process, where class diagrams complemented with object diagrams are used to capture the concepts of the problem domain of the system to develop and to highlight relationships between the identified concepts [Bruegge and Dutoit 2003].

Despite the relevance of the object diagrams as a complement of the class diagrams, the contribution they provide has been marginally investigated in the past by means of experimental
studies. Currently, there exists only one controlled experiment [Torchiano 2004], executed in 2003, assessing whether the use of object diagrams supports the comprehension of the problem domain of the software system to be developed. The context of that experiment is constituted of a group of 17 graduated students in Computer Science at the Politecnico di Torino in Italy. The results show that the subjects benefit from the use of object diagrams.

To confirm or contradict the findings of that experiment, we present in this paper a replication with a group of 24 graduated subjects in Computer Science of the University of Basilicata. With respect to the original experiment, we have deliberately introduced variations to improve both the experimental material and the data analysis.

The paper is organized as follows: Section 2 discusses a subset of the related literature concerning experiments aimed at assessing the use of UML in comprehension tasks. Section 3 provides an overview of the original experiment and then describes the design and execution of the replication. In this section differences between these experiments are highlighted as well. Section 4 presents the data analysis, while Section 5 discusses the results of the replication and the threats to validity. Final remarks and future direction for our work conclude the paper.

2. RELATED WORK

In this section, we discuss a subset of the related literature concerning experiments aimed at assessing the use of UML based software models in comprehension tasks. The interested reader can find a systematic literature review on UML in [Budgen et al. 2010].

Andrew and Drew [Andrew and Drew 2009] empirically investigate whether the use case diagrams improve the effectiveness of use cases by providing visual clues. The investigation is conducted providing a group of students only with use cases (control group) or use cases augmented with use case diagrams (treatment group). The results show that students employing both use case diagrams and use cases achieve a significantly higher level of understanding.

Controlled experiments have been also conducted to assess the effectiveness of other UML diagrams. For example, in [Gravino et al. 2008] is presented a controlled experiment with Bachelor students to assess whether the comprehension of software requirements is influenced by the use of the dynamic models abstracted by employing sequence diagrams. The data analysis reveals that there is not a significant difference in the comprehension of system requirements when using or not dynamic models. On the other hand, the results of an external replication conducted with more experienced subjects (i.e., Master students) are presented in [Abrahao et al. 2009]. Differently from the original experiment, the data analysis shows that the use of dynamic models facilitates the interpretation and comprehension of software requirements.

With regard to the comprehensibility of class diagrams, an empirical investigation is proposed in [Yusuf et al. 2007]. The authors used eye-tracking equipment to assess subjects’ comprehension in the context of software design problems. The results indicate that more experienced subjects prefer to use stereotype information, coloring, and layout to promote the exploration and the navigation of class diagrams. Differently, in [Lucia et al. 2010] the results of three controlled experiments are presented. The goal of these experiments consists in investigating whether the class diagrams are more comprehensible than the entity relationship diagrams in the comprehension and the execution of maintenance tasks. The results shows that the subjects are not better supported when provided with class diagram.

In [Gross and Doerr 2009] two controlled experiments to compare the UML activity diagrams and the event-driven process chains are reported. The main objective of that comparison concerns the model understandability from the perspective of both requirements engineers and customers. In the case of requirements engineers, better performances have been obtained when they use activity diagrams. On the other hand, the customers do not get significant differences in terms of business process understandability, when they use both the business process visual notations.

An experiment is reported in [Byckling et al. 2006] to assess whether the comprehensibility of object models is increased when additional information are added to class and sequence diagrams with respect to the behavior of their attributes. The results indicate that the comprehensibility is increased when the role of the variables is expressed in compact manner to the diagrams on which the study is focused. Similarly, the experiment presented in [Nugroho 2009] focus on the level of detail of some UML diagrams. The results show a significant effect of the level of detail on the comprehension of software models.

Empirical studies are also conducted to assess the impact of the comprehension of UML diagrams on software maintenance tasks. For example, in [Otero and Dolado 2005] is presented a study
to assess whether documentation based on UML may significantly improve the functional correctness of changes as well as the design quality when complex tasks have to be accomplished. Similarly, in [Gravino et al. 2010] a controlled experiment to assess the contribution of requirements UML based models on the comprehension of source code is presented. The study reveals that the comprehension is not increased in case subjects are provided with requirements models.

In the literature the effect of employing the UML stereotypes in the execution of comprehension tasks is investigated as well. For example, a series of controlled experiments to assess the effectiveness of stereotypes in class diagrams to comprehend Object-Oriented applications in the telecommunication domain is shown in [Staron et al. 2006]. This experiment is conducted with students and professionals. The results indicate that the use of stereotypes significantly helps both the kind of subjects to improve comprehension. Similarly, in [Ricca et al. 2007; 2010] a set of experiments with bachelor, master, and PhD students is presented to assess the understanding of the Conallen's stereotypes in the context of the design of Web applications. Differently to [Staron et al. 2006], stereotypes seem little useful in understanding. The main finding of that experiment is that stereotypes reduce the gap between experienced and less experienced subjects.

3. EXPERIMENTATION SETUP

In this section we first present an overview of the original experiment an then show the design and execution of the replication. We followed the guidelines proposed in [Juristo and Moreno 2001, Wohlin et al. 2000]. For replication purposes, the experimental package (in Italian) and the raw data are available on the web.

3.1. Overview of the Original Experiment

In the following we provide an overview of the original experiment.

3.1.1. Experiment context, goal, and design

The experiment was conducted within a research laboratory at the Politecnico di Torino (Italy) with a group of 17 students of the first year of the master program in Computer Science. The experiment was a regular class exercise of a basic software engineering course.

The goal consisted in assessing whether the use of the UML object diagrams as a complement to the class diagrams could improve the comprehension of the problem domain of a given software system. The original experimenter used a factorial design with two groups, one treatment, and four objects. Subjects were randomly assigned to groups, while the treatment consisted in providing (+) or not (−) object diagrams. The experimental objects were four (some details are shown in Table 1) and they were based on the following four systems:

- File System (FS). A system to manage folders, files, and links. Folders can contain other elements (e.g., files), while links refer to other elements in the file system (e.g., folders). The class diagram (up) and the object diagram (down) used in the experimental tasks are shown in Figure 1.

- Roads (R). It is used to describe maps made up of cities connected by means of roads. Each road starts and ends in a city. Furthermore, a road is characterized by a length;

- Trains (T). A system to manage timetables, trains, and paths. The interested reader can find the class diagram and the object diagram of this system in [Torchiano 2004];

- Catalogue (C). A catalogue is implemented to collect category of items (e.g., cars) and to describe items (e.g., car models) based on a set of features (e.g., number of doors) that can have a set of possible values.

<table>
<thead>
<tr>
<th>Object</th>
<th>Class diagram</th>
<th>Relations</th>
<th>Object diagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>R</td>
<td>2</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 1: Characteristics of objects.

Figure 1: The class and object diagrams of FS.
The first experimental group performed the comprehension task FS+, R−, T+, and C−, while the second one did T−, C+, FS−, R+. The comprehension level achieved by the subjects was evaluated using a questionnaire (per experimental object) that included four multiple-choice questions with only one correct answer. The score (i.e., number of correct answers) achieved by a subject on a given questionnaire indicated his/her comprehension level on the associated object. An example of comprehension question for FS is: “Which object types can belong to a Folder object?”. To answer this question² is sufficient to look at the aggregation in the class diagram (see Figure 1) or to observe the example provided by the object diagram (if present).

3.1.2. Results
The data analysis was performed considering the experimental objects together and alone. The analysis on each object separately revealed a significant statistical difference in favor of the object diagrams only on FS and T. Moreover, a significant positive effect of the object diagrams was revealed when all the objects were considered together.

The analysis of the magnitude of the observed differences using the Cohen’s “d” standardized difference [Cohen 1988] between two groups revealed that the effect size is medium for FS and T and negligible for R and C. Overall, the effect is small when considering all the objects together (see Section 5.1 for further details).

3.2. The replication
We present here the design of a replicated experiment and discuss the differences it introduces with respect to the original one.

3.2.1. Context
The replication has been conducted with 24 students of the master program in Computer Science at the University of Basilicata (Italy). The experiment represented an optional activity of an advanced software engineering course. Before the experiment, the subjects had passed the exams of the following courses: basic and advanced object oriented programming languages, database system modeling, and basic software engineering.

3.2.2. Hypotheses
Similarly to the original experiment, the perspective of the replication is from the point of view of the researchers, investigating on the effectiveness of objects diagrams in the comprehension of software systems, and of the project managers, evaluating the possibility of adopting objects diagrams in the development and evolution of the software systems of their organizations. Accordingly, we have investigated the following null hypotheses:

\[ H_{00} : \text{The presence of object diagrams does not significantly improve the comprehension level of the subjects to perform comprehension tasks.} \]

\[ H_{00} : \text{There is no significant difference in terms of effort when using or not object diagrams to perform a comprehension task.} \]

\[ H_{ed0} : \text{The presence of object diagrams does not significantly affect the perceived difficulty in performing comprehension tasks.} \]

\[ H_{cda} : \text{The presence of object diagrams does not significantly affect the perceived confidence in comprehension tasks outcomes.} \]

Based on the results of the original experiment, the null hypotheses \( H_{00}, H_{ed0}, \) and \( H_{cda} \) are one-tailed since we expect a positive effect of the object diagrams on the subject performances. On the other hand, \( H_{00} \) is two-tailed since we cannot postulate a difference expectation in terms of effort.

The objective of the statistical analysis will be rejecting the above defined null hypotheses, thus accepting the corresponding alternative ones (i.e., \( H_{0a}, H_{ta} H_{eda}, \) and \( H_{cda} \)) that can be easily derived.

3.2.3. Design
The experiment adopts a counterbalanced design (see Table 2), thus ensuring that each subject work on two comprehension tasks (each consisting of two experimental objects), experimenting alternatively class and object diagrams together or class diagram alone. Similarly to the original experiment, the experimental objects were paired in assignments as follows:

- **A1**: the pair of objects FS and T;
- **A2**: the pair of objects R and C.

<table>
<thead>
<tr>
<th>Task</th>
<th>Group1</th>
<th>Group2</th>
<th>Group3</th>
<th>Group4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A1+</td>
<td>A1−</td>
<td>A2+</td>
<td>A2−</td>
</tr>
<tr>
<td>2</td>
<td>A2−</td>
<td>A2+</td>
<td>A1−</td>
<td>A1+</td>
</tr>
</tbody>
</table>

Table 2: Experiment design. Class diagram completed with object diagram (+), only class diagram (-)

3.2.4. Selected variables
In this replication the control group is class diagrams with no object diagrams, while the treatment group is class diagrams completed with object diagrams. Thus, the only independent variable is Treatment, which is a nominal variable (with values \( \{+, −\} \)).
The dependent variables selected to investigate the defined null hypotheses are the comprehension level and comprehension effort.

The comprehension level dependent variable is used to measure the subjects' comprehension. Similarly to [Kuzniarz et al. 2004], it has been assessed by employing a comprehension questionnaire (in our case 4 open questions per object). As suggested in [Aranda et al. 2007], the subjects were asked to specify the source of information used to answer the questions of this questionnaire. To this end, we added a further question for each question: “Did the answer come from the class diagram, object diagram, description of the system, previous Knowledge?”. For each question, the subjects were also asked to judge both his/her confidence on the provided answer and the complexity of the question: “How much do you trust your answer?” (possible answers: Unsure, Not sure enough, Sure Enough, Sure, and Very Sure) and “How do you judge this question?” (possible answers: Very difficult, Difficult, Medium, Simple, and Very Simple).

The comprehension level of the subjects has been measured using an information retrieval based approach as in [Ricca et al. 2007]. To this end, we defined as: \( A_{s,i} \), the set of answers (a list of string items) provided by the subject \( s \) on the question \( i \), and \( C_{i} \), the correct set of items of the question \( i \). Accordingly, the correctness of the provided answers has been measured using the precision measure, while the completeness of the answers using the recall measure:

\[
\text{precision}_{s,i} = \frac{|A_{s,i} \cap C_i|}{|A_{s,i}|} \quad \text{recall}_{s,i} = \frac{|A_{s,i} \cap C_i|}{|C_i|}
\]

We used the harmonic mean between precision and recall to aggregate them and to get a single value representing a balance between correctness and completeness:

\[
F\text{-Measure}_{s,i} = \frac{2 \cdot \text{precision}_{s,i} \cdot \text{recall}_{s,i}}{\text{precision}_{s,i} + \text{recall}_{s,i}}
\]

To obtain a single measure of the comprehension level achieved by each subject, we computed the overall average of the F-Measure values of all the questions. This mean assumes values ranging from 0 to 1. A value close to 0 indicates a very bad comprehension level of the system, whilst a value close to 1 means a very good understanding.

The second dependent variable, comprehension effort, measures the time, expressed in minutes, used to accomplish a task and it was directly recorded by the subjects during the experiment.

3.2.5. Execution and material

To execute the tasks, the subjects were asked to use the following procedure: (i) specify name and start-time in the comprehension questionnaire; (ii) answer independently the questions by consulting the provided material; (iii) mark the end-time of the task.

To perform the experiment the subjects were provided with the following hard copy material:

- the problem statement\(^3\) of the experimental objects;
- the comprehension questionnaires to be filled in (the questionnaires for Task2 were provided to the subjects once the first task was accomplished);
- a post-experiment questionnaire to be filled in after the two tasks.

The post-experiment questionnaire aimed at gaining insights about the subjects’ behavior during the experiment and at better explaining the obtained quantitative results. The questionnaire is composed of 13 questions concerning the availability of sufficient time to complete the tasks, the clarity of the objects, the ability of subjects to understand them, and the perceived usefulness of the object diagrams and class diagrams. All the questions include closed answers according to a five point Likert scale: (1) strongly agree, (2) agree, (3) not certain, (4) disagree, (5) strongly disagree. For space reasons, the results of the post experiment survey questionnaire are only marginally discussed in the following.

3.2.6. Analysis procedure

In all our statistical tests we decided (as it is customary) to accept a probability of 5% of committing Type-I error [Wohlin et al. 2000], i.e., rejecting the null hypothesis when it is actually true. Because of the sample sizes (24 subjects) and mostly non-normality of the data we adopted non-parametric tests to reject the null hypotheses. In particular we selected Mann-Whitney test for unpaired analysis and Wilcoxon test for paired analysis because they are very robust and sensitive [Juristo and Moreno 2001, Wohlin et al. 2000].

While the statistical tests allow for checking the presence of significant differences, they do not provide any information about the magnitude of such a difference. Accordingly, we used the Cohen’s “d” standardized difference between two groups [Cohen

---

\(^3\)A problem statement describes the current situation, the functionality the system should support, and the environment in which it will be deployed.
1988]. Typically, it is 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 \( d \geq 0.8 \).

Additional variables (also named co-factors in the following) have been controlled, such as Assignment \( (\in \{A1, A2\}) \), Task \( (\in \{Task1, Task2\}) \) and Group \( (\in \{Group1, Group2, Group3, Group4\}) \) by means of a two-way Analysis of Variance (ANOVA), mainly to discover interactions with the treatment, and with the Mann-Whitney test.

3.3. Differences between the experiments

To improve the material and the data analysis, some modifications with respect to the original experiment have been introduced:

- **Comprehension questionnaire.** We removed mistakes and some sources of possible confusion. Furthermore, we modified the questionnaire by turning the closed questions of the original experiment into open ones. The rationale for executing this modification relies on the fact that open questions should reduce as much as possible that subjects guess and hence give correct answers by chance.

- **Dependent variables.** Apart from comprehension, we added effort as dependent variable. This was added to further investigate the effect of object diagrams in performing comprehension tasks.

- **Experiment design.** In the replication, a within-subjects counterbalanced experimental design was used. This modification was introduced to analyze the effect of the main factor and the co-factors better. Other modifications have been introduced as a natural consequence of having used a different design. For example, a break between the two tasks was provided, thus expecting a lower fatigue effect with respect to the original experiment.

- **Group composition.** We used the information gathered in a pre-questionnaire to equally distribute high and low ability subjects among the four groups of Table 2. Differently, in the original experiment the subjects were randomly assigned to the groups.

- **Post-experiment survey questionnaire.** In order to get insights about the subjects’ behavior during the experiment and to better explain the obtained results a post-experiment questionnaire was employed in the replication.

- **Data analysis.** Bearing in mind the new dependent variable introduced (comprehension effort), we considered a new null hypothesis. Two additional null hypotheses have been considered to investigate the perceived difficulty and confidence to perform comprehension tasks. Moreover, in the replication we accepted the standard probability of 5% of committing Type-I-error instead of 17% as done in the original experiment.

- **Subjects’ experience.** The subjects involved in the replication are more experienced than the original ones since they have already passed a basic software engineering course and they are attending an advanced software engineering course.

4. ANALYSIS AND RESULTS

In this section the results of the data analysis are presented with respect to the defined null hypotheses. We conclude the section discussing the effect of the co-factors and briefly presenting the results of the post-experiment questionnaire.

4.1. Comprehension level

Table 3 reports some descriptive statistics of comprehension level, one-tailed tests and effect size (Cohen’s \( d \)) grouped by assignments and experimental objects. The overall comparison (i.e., without partitioning by assignment/object) is visually presented in Figure 2 by means of boxplots. From them, it is apparent that students with object diagrams (OD) outperformed in comprehension students without them (NO-OD). We evaluate the first hypothesis overall. Both the paired \( (p = 1.4 \cdot 10^{-5}) \) and unpaired Mann-Whitney \( (p = 3.4 \cdot 10^{-6}) \) provide evidence that there exists a significant difference in terms of comprehension level as effect of the presence of object diagrams. Therefore, in general, we can reject the null hypothesis \( H_{0,\text{D}} \).
On the Effectiveness of the UML Object Diagrams
Giuseppe Scanniello et al.

<table>
<thead>
<tr>
<th>Assignment</th>
<th>Object</th>
<th>NO OD</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>OD</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>Mann-Whitney p-value</th>
<th>Wilcoxon p-value</th>
<th>Cohen d</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>All</td>
<td>0.58</td>
<td>0.60</td>
<td>0.21</td>
<td></td>
<td>0.79</td>
<td>0.75</td>
<td>0.14</td>
<td></td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>1.18</td>
</tr>
<tr>
<td>A1</td>
<td>FS + T</td>
<td>0.59</td>
<td>0.62</td>
<td>0.23</td>
<td></td>
<td>0.83</td>
<td>0.78</td>
<td>0.12</td>
<td></td>
<td>&lt;0.001</td>
<td>—</td>
<td>1.28</td>
</tr>
<tr>
<td></td>
<td>FS</td>
<td>0.55</td>
<td>0.59</td>
<td>0.19</td>
<td></td>
<td>0.78</td>
<td>0.75</td>
<td>0.10</td>
<td></td>
<td>&lt;0.001</td>
<td>—</td>
<td>1.48</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.63</td>
<td>0.69</td>
<td>0.27</td>
<td></td>
<td>0.88</td>
<td>0.88</td>
<td>0.12</td>
<td></td>
<td>0.007</td>
<td>—</td>
<td>1.19</td>
</tr>
<tr>
<td>A2</td>
<td>R + C</td>
<td>0.57</td>
<td>0.57</td>
<td>0.19</td>
<td></td>
<td>0.75</td>
<td>0.75</td>
<td>0.15</td>
<td></td>
<td>&lt;0.001</td>
<td>—</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>0.61</td>
<td>0.59</td>
<td>0.20</td>
<td></td>
<td>0.76</td>
<td>0.78</td>
<td>0.13</td>
<td></td>
<td>0.035</td>
<td>—</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.53</td>
<td>0.57</td>
<td>0.18</td>
<td></td>
<td>0.74</td>
<td>0.73</td>
<td>0.17</td>
<td></td>
<td>0.008</td>
<td>—</td>
<td>1.22</td>
</tr>
</tbody>
</table>

Table 3: Descriptive statistics of comprehension level and results of the Mann-Whitney test

From a practical point of view, the difference can be considered large (\(d = 1.18\)). The mean comprehension level improvement achieved with object diagrams is of 21 points (see means of the “All” row in Table 3), i.e., 36.2\%\(^4\).

Substantially analogous results can be observed for the primary measures, precision (\(p = 4.15 \cdot 10^{-5}\) and \(d = 1.53\)) and recall (\(p = 1.6 \cdot 10^{-5}\) and \(d = 1.58\)).

Focusing on the separate assignments we can observe a significant difference both for A1 and A2. Detailing further on objects (see boxplots in Figure 3), it appears that a significant difference exists for all four objects (see Table 3). Thus, the first null hypothesis can be rejected for FS, R, C and T.

4.2. Comprehension effort

Figure 4 shows the overall boxplots of comprehension effort versus the treatment. Apparently, students without object diagrams (NO OD) employed more time than students with object diagrams (OD). Means and medians are respectively: 39'52" and 40 minutes for NO OD; 36'5" and 35 minutes for OD. A paired two-tailed Wilcoxon test returned \(p = 0.39\), therefore we cannot reject the overall null hypothesis \(H_{0o}\).

As far as the effect size is concerned, we obtained a standardized Cohen’s \(d = 0.36\), which can be considered small in practical terms. Even analyzing the four objects separately no significant difference was found (the results of an unpaired two-tailed Mann-Whitney test are: \(p = 0.7\) for FS, \(p = 0.06\) for R, \(p = 0.66\) for T and \(p = 0.93\) for C). The comparison is visually presented in Figure 5 by means of boxplots.

4.3. Perceived complexity and confidence in the answers

Ease of understanding has been measured in terms of both perceived complexity of the understanding tasks and confidence in the answers. Perceived complexity is lower and confidence in the answers is higher when object diagrams are present, although by a negligible measure (\(d = 0.2\)) the former and by a small amount (\(d = 0.22\)) the latter. We applied the one-tailed paired Mann-Whitney test that returned non significant differences for both complexity (\(p = 0.24\)) and confidence (\(p = 0.17\)). Therefore, we rejected neither null hypotheses \(H_{eco}\) nor \(H_{ede}\).

4.4. Source of information

The proportion of the sources of information used by the students to answer globally the questions are the following. For OD: class diagram 45\%, object diagram 51\%, description of the system 3\% and previous knowledge 1\%. For NO OD: class diagram 77\%, description of the system 12\% and previous knowledge 11\%. We can observe that the class

\(^4\)The percentage comes from 0.58+0.58*x%=0.79
diagrams represent a relevant source of information, while object diagrams are predominant when present (i.e., in the treatment group).

4.5. Co-factors

We have analyzed the effects of co-factors (Assignment, Task and Group) on both comprehension level and effort. On the overall data set, we found no significant effect of Assignment on comprehension level ($p = 0.19$) and no interaction with Treatment ($p = 0.48$). Similarly, no significant results were obtained analyzing the effects (and interactions) of the co-factors Task and Group on the comprehension level. The results were confirmed by Mann-Whitney test.

Differently, we found a significant effect of Assignments on effort ($p = 0.008$) without interaction with Treatment ($p = 0.66$). A marginal interaction ($p = 0.06$) was also observed between Treatment and Task as shown in Table 4. Finally, we found no significant effect of Group on effort ($p = 0.18$) and no interaction with Treatment ($p = 0.14$). Also in this case, results were confirmed by the Mann-Whitney test.

4.6. Post-questionnaire results

The students judged sufficient the time to complete the task (median=1), they also found perfectly clear: the objectives of experiment, description of the systems (median=2), and comprehension questions (median=2). They found the object diagrams useful (median=2). Moreover, students consider object diagrams more useful for understanding the cardinality aspects of the associations (median=2) and in presence of direct or indirect self-loop associations that generate a graph-like structure (median=2). Finally they found very useful the experiment from the pedagogical point of view (median=1).

Table 4: Two-way ANOVA of Treatment and Task vs. Effort

<table>
<thead>
<tr>
<th>Source</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>3</td>
<td>86.26</td>
<td>28.75</td>
<td>1.66</td>
<td>0.2011</td>
</tr>
<tr>
<td>Task</td>
<td>1</td>
<td>58.59</td>
<td>58.59</td>
<td>1.13</td>
<td>0.2914</td>
</tr>
<tr>
<td>Treatment:Task</td>
<td>3</td>
<td>189.84</td>
<td>63.28</td>
<td>3.65</td>
<td>0.0592</td>
</tr>
<tr>
<td>Residuals</td>
<td>92</td>
<td>4786.29</td>
<td>52.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Comparison between the experiments in terms of comprehension level

5. DISCUSSION

In the following we discuss the obtained results and the threats that may affect their validity.

5.1. Discussion on the results

Table 5 summarizes the results in terms of comprehension level (both significance and effect size) of our replication and provides, as a comparison, those obtained by the original experiment. Globally, (see row “All”) the results confirm and strengthen the findings of the original experiment, thus increasing our confidence in the benefits deriving from the use of object diagrams: the subjects significantly benefit from the use of object diagrams.

A direct and detailed comparison of the results between the original study and the replication is made difficult by the differences concerning material and data analysis (see Section 3.3). In particular, the modifications of the comprehension questionnaire (closed questions turned into open questions) and the ensuing different way to measure the comprehension level (score vs. F-measure) forbids direct numerical correspondence.

At the macroscopic scale, we observe in the replication a stronger significance and an a larger effect when compared to the original study. Indeed, the results of the replication revealed benefits in all the four systems, while this is not true in the original experiment (see Table 5). Comparing the improvement amount of comprehension level attained for the four objects, we can enforce the speculation, provided in the original paper, that very simple diagrams containing a few classes benefit less from an additional object diagram: the simplest object – i.e. R counting just two classes (see Table 1) – see the least benefits. We can not draw further conclusions concerning which class “structure” can
enjoy most comprehension improvement from an accompanying object diagram.

Regarding the comprehension effort, the experiment did not reveal any significant difference. This feature, not investigated in the original study, suggests that the additional information — consisting in the additional object diagram — does not require extra effort to be analyzed.

Finally, the benefit deriving from the use of the object diagrams is also consistent with the opinions of the students. Difficulty of the questions is lower and confidence of the answers is higher when object diagrams are present. Moreover, object diagrams are the predominant source of information. In addition, the data analysis of the post-questionnaires confirm the usefulness of object diagrams.

5.2. Threats to validity

The adopted design enabled us to mitigate Internal validity threats. For example, the effect of both learning and fatigue effects have been mitigated since a break of 10 minutes between the tasks was given to the subjects. This was also confirmed by the data analysis. Another issue concerns the information exchanged among the subjects. This was prevented as much as possible by monitoring them. In addition, students were not evaluated on their performance to avoid apprehension.

External validity may be threatened when experiments are performed with students, throwing into doubt the representativeness of the subjects with respect to software professionals. However, performed tasks did not require high levels of industrial experience, so we believe that this experiment can be considered appropriate, as suggested in the literature [Basili et al. 1999]. Another possible threat concerns the size and complexity of the tasks. Hence, we plan to conduct replications with more complex tasks. Replications with professionals are planned as well.

In this study, the Construct validity threats are related to the metrics used to get a quantitative evaluation of the subjects’ comprehension, the comprehension questionnaires, and the post-experiment survey questionnaire. Similarly to [Kuzniarz et al. 2004, Ricca et al. 2007; 2010] we used questionnaires and the provided answers were evaluated using an information retrieval based approach to avoid as much as possible any subjective evaluation. Furthermore, the comprehension questionnaires were defined to be complex enough without being too obvious. The comprehension effort was measured by means of proper time sheets and validated qualitatively by researchers. Although this may not be very accurate, this is a widely adopted way of measuring the effort to accomplish tasks in controlled experiments. The post-experiment survey questionnaire was designed using standard way and scales.

Conclusion validity concerns the data collection, the reliability of the measurement, and the validity of the statistical tests. Non parametric tests (i.e., Mann-Whitney and Wilcoxon) were used to reject the null hypotheses. Two-way ANOVA was used to detect the possible effects of the co-factors and the interactions between each co-factor and the main factor. Even if all the assumptions/conditions for using ANOVA were not checked this test is quite robust and has been extensively used in the past to conduct analysis similar to ours. In addition, the effect of the co-factors was also verified by the Mann-Whitney test.

6. CONCLUSIONS

In this paper, we have presented a replication of an experiment for assessing the effectiveness of complementing class diagrams with object diagrams. The results indicate a clear improvement in the comprehension level of software systems when object diagrams are present (the mean was increased by 36.2%) with no significant impact on the effort to accomplish comprehension tasks. Differently from the original experiment, we found a statistical difference for the comprehension level in all the four objects and a larger effect size.

Future replications will aim at investigating: (i) the effects of changing the domain and the complexity of the tasks; (ii) the motivation of the observed difference between experimental objects in the comprehension level; (iii) whether the benefits deriving from the use of object diagrams are preserved also for other categories of subjects (e.g., PhD students and professional). Future work will be also devoted to analyze whether the additional effort and cost to create object diagrams is paid back by an improved performance in carrying on tasks depending on the comprehension level, e.g. maintenance. In fact, from a manager point of view, the adoption of object diagrams should also take into account the costs they possibly introduce.

7. ACKNOWLEDGMENTS

We would like to thank the students that took part in the experiment.

REFERENCES

On the Effectiveness of the UML Object Diagrams
Giuseppe Scanniello et al.


A. Gross and J. Doerr. EPC vs. UML activity diagram - two experiments examining their usefulness for requirements engineering. pages 47–56, Los Alamitos, CA, USA, 2009. IEEE.


