The effects of UML modeling on the quality of software

Nugroho, A.

Chapter 5

The Impact of Level of Detail in UML Models on Comprehension

Previous studies have shown that the style and rigor used in UML models vary widely across software projects [79, 95, 96]. However, little research has been conducted to investigate the drivers and effects of using different styles and rigor in modeling on software development. In this chapter, the impact of modeling is further explored by introducing the notion of level of detail (LoD) as a form of style and rigor in modeling. In a controlled experiment, we investigate whether LoD in UML models affects the correctness and efficiency in comprehending UML models. Results show that the effect of LoD in UML models on model comprehension is significant.

5.1 Introduction

UML is the de facto standard for modeling software systems in industry. Its adoption in industry has become more prominent since the introduction of MDA (Model-driven Architecture), in which UML is one of the key foundation [100]. Apart from the use of UML for MDA or other forms of model-driven development, we have seen a great amount of evidence that UML is still prominently used in conventional manners: UML is mainly used for architecting solution, communicating design decisions, and detailed specification for implementation; and not so much for automatically generating implementation code.

While the use of UML models for automatic code generation generally requires certain level of formality in the models, in conventional model-driven development designers have

This chapter is adapted from the paper entitled "Level of Detail in UML Models and its Impact on Model Comprehension: A controlled Experiment", published in the Information and Software Technology Journal 2009.

more freedom to choose the level of formality, styles, and rigor in modeling a system. This freedom consequently leads to the use of various styles and rigor in UML modeling.

In practice, the use of different styles and rigor in modeling can manifest in various form, which includes the use of varying degree of completeness, level of detail, and disproportion in the models [79, 96]. Despite the potential effect of using different styles and rigor in modeling on model comprehension, there has been little study conducted to investigate the issue. Therefore, in this chapter we look into Level of Detail (LoD) in UML models and investigate its impact on model comprehension. LoD concerns the amount of information that is used to specify model elements. We devise the notion of LoD after observing a great deal of variation in the amount of information that is used to specify UML models [79, 95, 96].

In this chapter, we report on a controlled experiment involving two independent groups of graduate students. We explored the effects of using different degrees of LoD in a UML model on model comprehension. The results of the experiment confirm the significant effect of LoD in UML models on model comprehension.

The objective of the experiment according to the GQM template [130]:

Analyze level of detail in UML models
for the purpose of investigating its impact
with respect to model comprehension
from the perspective of the researcher
in the context of masters students at the Technology University (TU) Eindhoven

The rest of this chapter is organized as follows. In Section 2 and 3 we discuss related work and the definition of level of detail respectively. In Section 4 we discuss experiment planning. In Section 5 and 6 we discuss experiment operation and data analysis respectively. Finally, in Section 7 we provide discussion on the results, and in Section 8 we draw some conclusions and outline future work.

5.2 Related Work

In general, related work in the area of UML model comprehension studies the following aspects of UML models: diagram layout, diagram types, and modeling notations (e.g., the use of OCL and stereotypes).

Many studies that looked into UML model comprehensibility have been primarily looking at the layout or visualization aspects of UML models. The work of Purchase et al. reported in [110] for instance, investigated the impact of class diagram notations on model comprehension. In their work, two visualizations of several class diagram notations (e.g., inheritance direction, inheritance arcs, associations cardinality) were applied as treatments to a UML model. The results of their study reveal that certain visualizations are better than the other depending on the kind of comprehension tasks that need to be performed. Another work that investigated model comprehension related to diagram layout is from

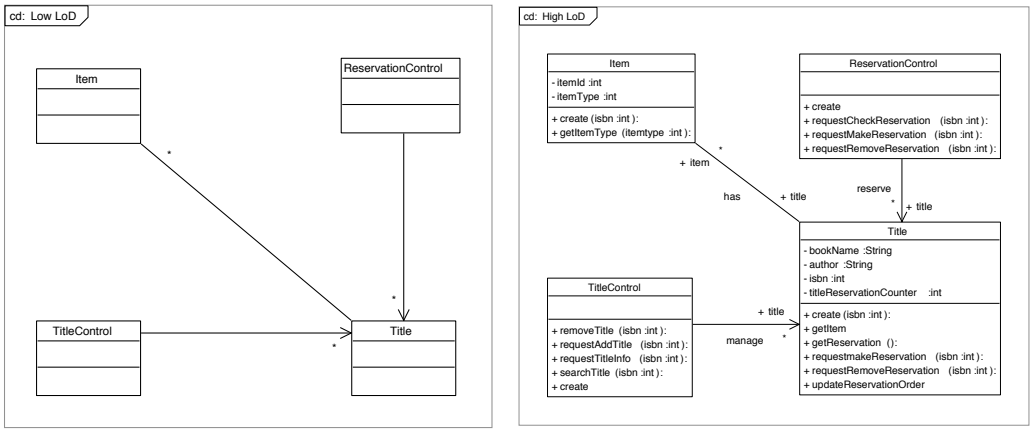
Wong and Sun [131]. In their work the authors established criteria and guidelines to create an effective layout for UML class and sequence diagrams based on perceptual theories.

Other studies that investigated UML model comprehensibility took the perspective that compares the effect of using different UML diagram types (e.g., sequence and collaboration diagrams). The work of Otero and Dolado for instance, looked into three UML diagrams types, namely sequence, collaboration, and state diagrams, and evaluated the semantic comprehension of the diagrams when used for different application domains [106]. A similar study comes from the work of Glezer et al. They evaluated the comprehensibility of sequence and collaboration diagrams, and finally concluded that collaboration diagrams are easier to comprehend than sequence diagrams in real-time systems [56]. Another study conducted by Torchiano [124] investigated the effect of object diagrams on system comprehensibility. In two of the four systems used in the experiment, the use of object diagrams to complement class diagrams was found to have significant effects on the comprehensibility of the systems.

Another line of research on UML model comprehension investigates the styles and rigor in UML models and their impact on model comprehensibility. In this line of research, the focus has been on the formality of UML models and its relation with model quality and comprehensibility. A previous study that deserves attention is the one from Briand et al. [27]. In their experimental study, Briand et al. investigated the impact of using OCL (object constraint language) in UML models on defect detection, comprehension, and impact analysis of changes. Although the overall benefits of using OCL on the aforementioned activities are significant, they have found that the benefits for the individual activities are modest. Lange and Chaudron conducted an experimental investigation into the effects of imperfections in UML models [78]. The results show that defects in UML models often remain undetected and cause misinterpretations.

Other studies in the area of modeling style looked into the effect of using stereotypes on model comprehension. The work of Staron et al. for instance, suggests that UML stereotypes with graphical representation improve model comprehensibility [120]. Ricca et al. also found that stereotypes have a positive impact on diagram comprehension [111]. However, this finding was particularly true for inexperienced subjects—the impact was not statistically significant for experienced subjects. Genero et al. studied the influence of using stereotypes in UML sequence diagrams on comprehension [53]. While this study revealed no significant impact, it suggested that the use of stereotypes in sequence diagrams was favored to facilitate comprehension. Another study was conducted by Cruz-Lemus et al. to evaluate the effect of composite states on the understandability of state-chart diagrams [40]. The authors stated that the use of composite states, which allows the grouping of related states, improves *understandability efficiency* when reading state-chart diagrams. Nevertheless, subjects' experience with state-chart diagrams was considered as a prerequisite to gain the improved understandability.

Another work worth mentioning is from Genero et al., which investigated measures that could be used to predict diagram maintainability [54]. Their study revealed that the number of associations and the maximum depth of inheritance (DIT) in class diagrams are good predictors of the time required to understand (understandability time) and modify (modifiability time) the diagrams. A study from Arisholm et al. looked at the problem from a coarser grained view: the absence/presence of UML in software maintenance [10].



(a) Class diagram with low LoD

(b) Class diagram with high LoD

Figure 5.1: Example of level of detail in a class diagram

Their experimental study using a synthetic UML model confirmed that the use of UML for maintenance significantly reduces time to make code changes in the system and increases functional correctness of the changes. However, the authors also stated that effort saving was not visible when the time required to change the UML diagrams was taken into consideration.

Similar to the work of Briand et al., we look at the effect of rigor and formality in UML models on model comprehensibility. Our approach is different from the above previous works in that we measure the rigor and formality in a UML model from the LoD used in the model. In this study our investigation is focused on LoD in UML class and sequence diagrams.

5.3 Level of Detail

LoD in UML models is defined as the amount of information that is used to represent a modeling element. For example, the modeling element 'message in a sequence diagram' may be represented by any of the following amounts of information: an informal label, a label that represents a method name, a label that is a method name plus the parameter list. Likewise, in modeling class diagrams, many syntactical features are available to increase the LoD: class attributes and operations, association names, association directionality, and multiplicity. When the LoD used in a UML model is low, it typically employs only a few syntactic features such as class-name and associations without specifying any further facts about the class. Figure 5.1 shows a fragment of a system represented using class diagram of high and low LoD.

5.4 Experiment Planning

In this section we discuss how the experiment was designed, which includes the subjects of the experiment, variables in the experiments, and the types of treatments received by the subjects. Furthermore, all instruments that were used in the experiment will be discussed.

5.4.1 Subjects in the Experiment

The subjects of the experiment were students of computer science at TU/e (Technology University Eindhoven) who were taking the Software Architecting course in their second-year of M.Sc. study. The experiment was part of mandatory assignments of the course, thus the subjects will obtain grades based on their performance in the experiment (because the treatments applied in the experiment might influence subjects' performance, we used different grading schemes for the experimental groups). Hence, we were assured that the subjects were motivated to participate in the experiment. In total, there were 53 students participating in the experiment.

5.4.2 Variables in the Experiment

The independent variable in the experiment was LoD in UML diagrams. We used two treatment levels for the independent variable, namely low LoD and high LoD. These two levels of treatments were achieved by manipulating the amount of information that was used to specify certain UML model elements in class- and sequence diagrams—that is, the UML model with low amount of information was considered low LoD and the UML model with high amount of information, high LoD. Because the independent variable was dichotomous, it was measured in nominal scale.

The dependent variable in the experiment was model comprehension. In this experiment, model comprehension was defined as the ability of the subjects to understand concepts/constructs described in a UML model. We defined two aspects of model comprehension, namely *correctness* and *efficiency*. Correctness was measured as the percentage of correctly answered comprehension questions. Similar to [40], efficiency was measured as the relation between the number of correctly answered comprehension questions and the amount of time required to complete all questions. Both correctness and efficiency were measured in a ratio scale. For the remainder of this chapter, we use the term *comprehension correctness* and *comprehension efficiency* to refer to correctness and efficiency respectively.

5.4.3 Hypotheses Formulation

With regard to LoD in UML models, we defined two hypotheses. The first hypothesis concerned the effect of LoD on comprehension correctness. This hypothesis was based on the assumption that the higher the amount of information put into a model, the more is known about the concepts/knowledge described in the model. Consider an example of a UML class that is modeled in two different ways in terms of level of detail. In the first way,

the class is modeled without any attribute or operations (methods); while in the second way, all class attributes and operations are specified. If we have two independent groups of subjects, each of which is exposed to one of the classes, we investigate whether subjects who are exposed to the later class representation will understand the class better (e.g., in terms of its role and responsibility) than those who are exposed to the former representation of the class.

In the second hypothesis we aimed at investigating whether the LoD used in UML models affects comprehension efficiency. Using the same class example presented above, we were interested in exploring whether subjects who are given a class with higher LoD would correctly understand the role of the class faster than those who are given a class with lower LoD. This is a valid assumption because humans generally infer certain phenomena faster when sufficient information is available.

Hence, the hypotheses we intended to test in the experiment were the following:

Hypothesis 1:

- $H_{1,null}$: There is no significant difference of comprehension correctness between subjects when working with UML diagrams modeled using high or low LoD.
- $H_{1,alt}$: The use of UML diagrams with high LoD significantly improves subjects' comprehension correctness.

Hypothesis 2:

- $H_{2,null}$: There is no significant difference of comprehension efficiency between subjects when working with UML diagrams modeled using high or low LoD.
- $H_{2,alt}$: The use of UML diagrams with high LoD significantly improves subjects' comprehension efficiency.

Note that we stated one-tailed hypotheses because we had prior predictions that LoD in UML diagrams will increase both comprehension correctness and comprehension efficiency.

5.4.4 Experimental Design

Experimental Design and Tasks

The experiment was designed as one factor with two treatments using a completely randomized design[130]—that is, each subject received only one treatment on one object (also known as between-subject design), and they were assigned randomly to each treatment. This experiment design was chosen mainly because we only had one session of two hours to execute the experiment.

A UML model of a library system was used in the experiment and the subjects were randomly assigned to each of the treatments—that is, low LoD and high LoD. We shall refer to subjects who received low detail model as group *L-LoD* and subjects who received high detail model as group *H-LoD*. Note, however, that we did not physically separate the two experimental groups (all subjects were in the same room without any seat arrangements related to the grouping). This way, we assured that the subjects were not aware of the grouping and treatments, which otherwise might have threatened the validity of the results.

Besides randomly assigning each subject to one of the two groups, we decided to have a balanced group size in order to simplify and strengthen data analyses [130]. To this aim, as soon as we knew the number of subjects attending the experiment, we allocated the same number of questionnaires for each group and distributed the questionnaire to the subjects in a random order. At the end, there were 27 and 26 subjects in group L-LoD and H-LoD respectively.

In the experiment, each subject received four materials, namely one UML model, one model comprehension questionnaire, one background questionnaire, and one feedback questionnaire. The first task that the subjects had to perform was to complete the model comprehension questionnaire, which contained questions about the library system specified in the UML model. Note that all subjects received the same model comprehension questionnaire regardless of their treatment groups. The subjects were also asked to register the time as they started and finished with the questionnaires. This time registration was required to measure the duration in completing the model comprehension questionnaire.

Additionally, the subjects were asked to complete background questionnaire and feedback questionnaire. The background questionnaire aimed at capturing the subject's knowledge background and experience with object-oriented analysis and design, object-oriented programming, UML, and library system. The feedback questionnaire, on the other hand, was used to get individual feedback about some aspects of the experiment such the clarity and difficulty of the tasks.

Other Factors to be Controlled

Besides the LoD treatment applied in the UML models, other factors may have influenced the subjects' performance in completing the comprehension task. We measured these factors using the background questionnaire so that it can be later verified whether they played a significant part in the differences of performance amongst subjects. The factors are listed below.

- Knowledge of and experience with object-oriented design. While some subjects may possess only theoretical knowledge about object-oriented design, others may also have industrial experience with object-oriented design.

- Knowledge of and experience with object-oriented programming. Subjects with good knowledge and experience of object-oriented programming (OOP) might outperform those with fair experience with OOP.

- Knowledge of and experience with UML. Many of the subjects might have had UML-related courses in the past. However, related to the assignment, having a theoretical knowledge may not be as helpful as having practical experience with UML.
- Experience with UML diagram types. It might be the case that some subjects had no experience with certain diagram types (e.g., sequence diagram). Those who have been using all diagrams used in this experiment might have better knowledge about the notations used in the diagrams than those who have been using only class or sequence diagram.
- Experience with a library system. Experience in designing and developing library systems might help to understand the concepts specified in the UML models better or faster.

As discussed previously, the above factors were measured using the background questionnaire discussed in section 5.4.5. Having measured those factors, we could then compare the overall knowledge and experience of the subjects in the two experimental groups. When there is no significant difference between the two groups, we should be confident that knowledge/experience differences between groups might only have a minor effect on differences of performance between groups.

5.4.5 Instrumentation

In this section each material used in the experiment will be discussed. We start with the UML model artifact, and subsequently follow with the model comprehension questionnaire, background questionnaire, and feedback questionnaire. The experiment materials can be downloaded from [93].

The UML Model

A UML model of a library system was used as the object of the experiment. The UML model was a modified version of the one that can be found in [23]. Each subject received a document containing a UML model that needed to be comprehended. Subsequently, the subjects had to answer a set of questions related to the model.

Because the objective of this experiment is to investigate the effects of LoD in UML models on model comprehension, two versions of the UML model were created. The first version of the model, hereafter referred to as model *M-Low*, depicts the library system in a low level of detail and the second version of the model, hereafter referred to as model *M-High*, depicts the library system in a higher level of detail.

Although the UML models have different levels of detail, both were modeled at the same level of abstraction. For example, both models exactly specify the same use cases, use case realizations (in terms of sequence diagrams), class structure, and architectural pattern (i.e., model-view-controller). Hence, the only difference is the amount of information that was

Table 5.1: LoD treatments in the UML model

Diagram Types	**Model Elements**	**Treatments**		**# of diagrams**
		M-Low	**M-High**	
Package diagram	Package name	Yes	Yes	1
Use case diagram	Use case name	Yes	Yes	2
	Actor name	Yes	Yes	
Class diagram	Class attributes	No	Yes	3
	Class operations	No	Yes	
	Association labels	No	Yes	
Sequence diagram	Real method names	No	Yes	17
	Message parameters	No	Yes	
	Message returns	No	Yes	

used to describe them. Further, while the model was technology-independent, we modeled it with implementation concerns in mind. Hence, both models are essentially *design models* that can be used to guide the implementation of the library system.

Four UML diagram types were used in the experiment, namely package diagram, use case diagram, class diagram, and sequence diagram. However, in the experiment the treatments were applied only to class and sequence diagrams. This decision was mainly due to the fact that class and sequence diagrams are the most commonly used UML diagrams [42]. Furthermore, use case and package diagrams have limited diagram notations; hence, the level of detail applied in use case and package diagrams is not as diverse as in class- or sequence diagrams. Nevertheless, we still used the use case and package diagrams for the sake of model completeness, as to resemble UML models used in real software projects.

Table 5.1 summarizes the two treatments applied to the class and sequence diagrams. As shown in the table, except for the package and use case diagrams, different treatments were applied to the class and sequence diagrams. Column *Model Elements* represents diagram notations to which treatments were applied. For class diagrams, treatments were applied to class attributes, operations, and association names. Thus, while model M-High specifies both class attributes and operations, neither class attributes nor operations were specified in model M-Low. However, with regard to association labels, not all associations in model M-Low were made unlabeled. Some associations that were crucial for understanding the basic concepts of the library system were kept labeled. Further, there were 20 classes modeled in the class diagrams. In model M-High, these classes contained 23 attributes and 123 operations, of which 14 percent were *getters*—mostly appeared in the entity classes. Although getters are trivial operations and may not be very useful for model comprehension, in some cases we still modeled them to make scenarios in the sequence diagrams complete and realistic.

Treatments for the sequence diagrams were mainly applied to messages. In model M-High, messages were specified exactly as they were specified as operations in the corresponding class diagrams. Hence, if an operation has a parameter and a return value, then these will also be specified as such in the corresponding messages in the sequence diagrams. On the contrary, in model M-Low dummy messages were used; they were merely text labels without parameters or return type. Note that other than the treatments discussed here, all

other aspects in the UML diagrams such as diagram layout were kept as similar as possible.

We should also underline that the decision to include/exclude certain diagram notations (e.g., class attribute, association label) for the treatment levels was based on two criteria. The first criterion was *practicality*—that is, the treatment should reflect real practices in industry. The second criterion was *information sufficiency*, which means diagram notations that are essential to the basic understanding of the UML models should not be used as treatments. For example, we did not use class name and cardinality (in class diagrams) or object type (in sequence diagrams) as part of the treatments.

As can be seen in Table 5.1, there were 23 diagrams in both versions of the UML model. The use case diagrams were used to describe the functionality and main actors in the library system, and each of the use cases was elaborated further using a sequence diagram. Moreover, one package diagram was used to specify the high level architectural layers of the system (user interface, business, and data layers). For each of the architectural layers, a class diagram was used to provide a detailed view of the static class structure. Finally, the sequence diagrams were used to describe how the functionality specified in the use case diagrams should be implemented. They capture behavioral views of the system.

In addition to the standard UML notations, we used notes to clarify certain basic concepts specific to the library system or UML notations (e.g., guards in sequence diagrams) that may not be common to every subject (these notes were put close to the diagrams that needed clarification). Hence, the notes were created to assure that the subjects would have a common basic knowledge of the system and modeling notations used. Furthermore, notes were used to annotate links between diagrams. This is mostly the case for sequence diagrams where links from one diagram to another occur very often. The notes were used in the same way in both model M-Low and M-High.

Figure 5.2 and 5.3 show the treatments applied to the class diagram. The class diagram in Figure 5.2 modeled with low LoD, whereas class diagram in Figure 5.3 modeled with high LoD. An example of treatments applied to the sequence diagrams is presented in Figure 5.4.

Model Comprehension Questionnaire

A questionnaire was designed to measure the subjects' comprehension of the UML model. The questionnaire contained questions about a library system that must be answered according to the specifications described in the UML model (the questionnaire is provided in Appendix B).

The questionnaire was designed in such a way that it covers different aspects of the systems. The themes of the questions were related to three aspects, namely domain knowledge/business rules, implementation how-to, and general architecture knowledge of a library system. However, the biggest portion of the questionnaire pertains to the implementation how-to. Furthermore, in designing the questionnaire we had to assure that regardless of the level of detail used, the questions we answerable based on the information provided in the UML models. Moreover, it was important to assure that the questions were both not trivial and too difficult to answer. We also took into account guidelines discussed in [104] in order to avoid biases and ambiguity in the questions' wording.

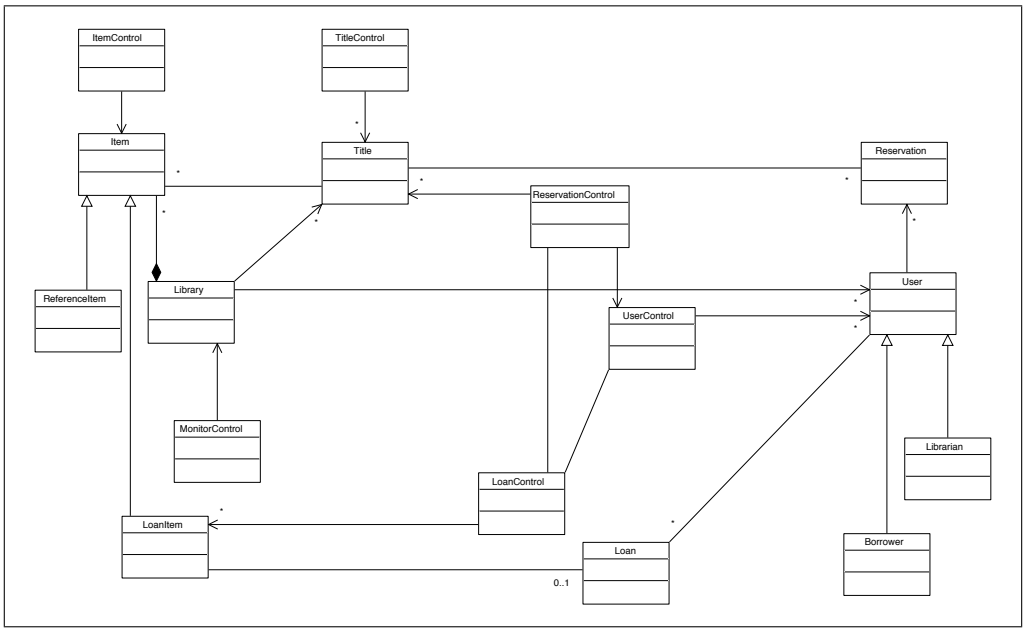


Figure 5.2: A class diagram in the UML model with low LoD (the M-Low model)

The model comprehension questionnaire that had to be completed by the subjects consisted of two parts, namely instruction- and question sheets. The instruction sheet provides information about the tasks that the subjects had to perform and how to perform them. For example, in the instruction sheet we asked the subjects not to collaborate with others or to use any books. We also asked the subjects to register the time as they started and finished with the questionnaire; furthermore, we provided a sample question.

The question sheets contain 15 multiple-choice questions, in which each question comprises three main parts, namely diagram reference, question and options, and rationale/remarks. Diagram reference provides the IDs of diagrams that were referred to by each question. Given the size of the models (23 diagrams) the references were necessary to help the subjects to quickly find the diagrams related to the question. The second part, question and options, contains the question and five available options to choose (option A up to E). The last option of every question (i.e., option E) is an option to choose where no appropriate answer could be inferred from the model. Finally, below each question there was space for the subjects to record their reasons in choosing certain answers. In fact, we encouraged the subjects to provide reasons/remarks, which turned out to be very helpful for us in understanding the reasoning of subjects' answers.

As discussed previously, we asked subjects to register the starting and ending time in completing the model comprehension questionnaire. In addition to that, we asked the subjects to register the time after answering three questions. Registering the time at certain intervals could have been valuable information to indicate difficult questions in the model comprehension questionnaire as perceived by the subjects. However, it turned out that

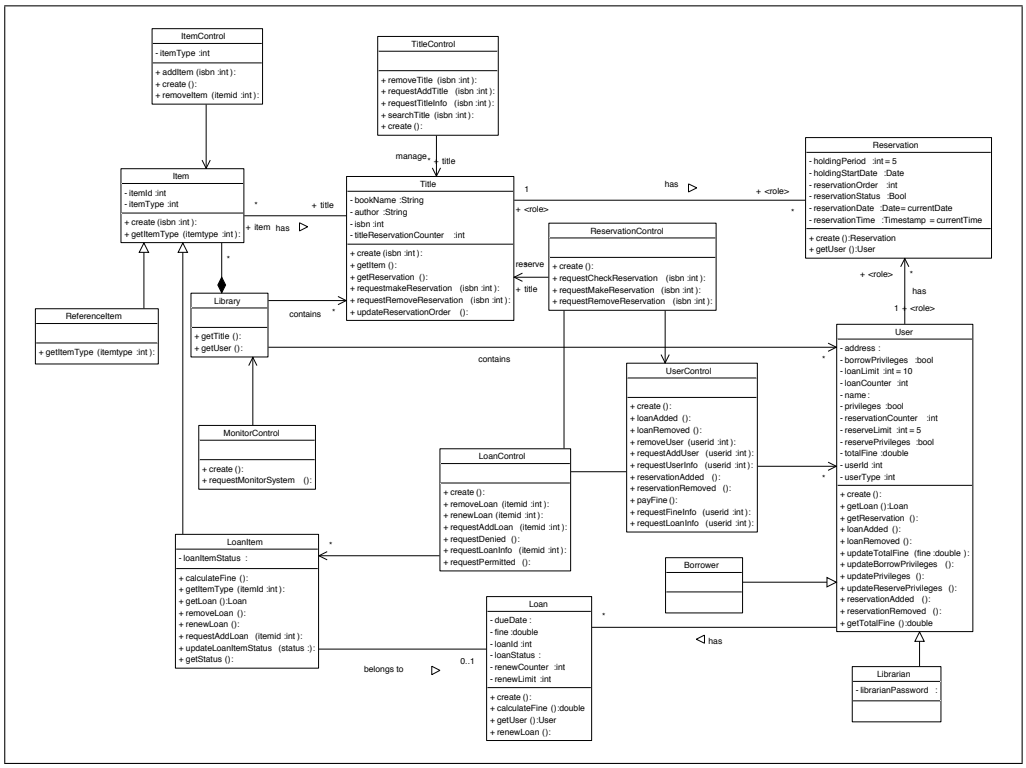


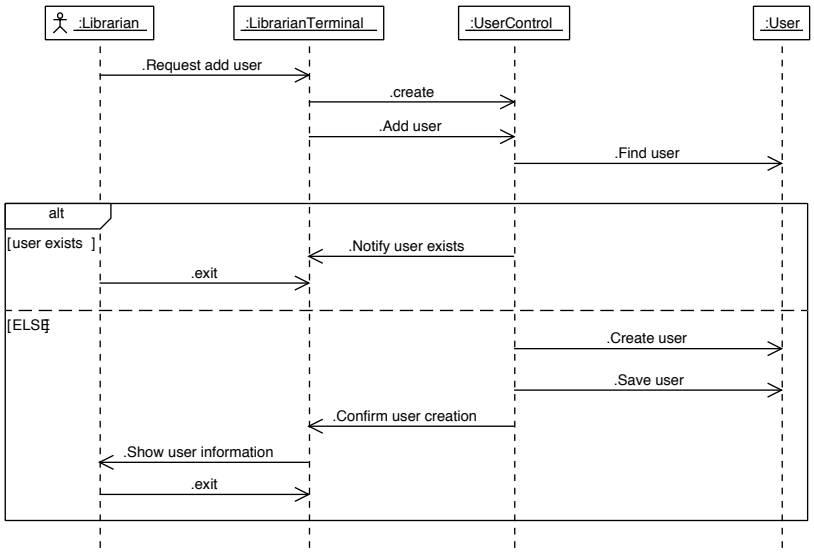
Figure 5.3: A class diagram in the UML model with high LoD (the M-High model)

many subjects did not rigorously register the time in this way; hence, it did not give usable information. Figure 5.5 shows an example of the questions asked in the model comprehension questionnaire.

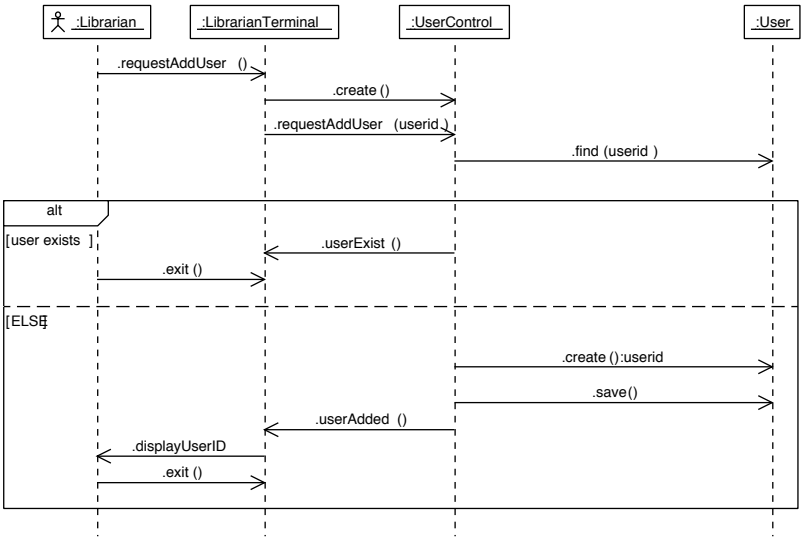
Background Questionnaire

The questionnaire to measure subjects' background knowledge and experience focused on four aspects of the subjects' knowledge/experience: object-oriented design, object-oriented programming, UML, and library systems. These knowledge/experience were thought to be influential to subjects' performance in the model comprehension task. In Section 5.4.4 we have discussed these aspects in further detail.

The aforementioned knowledge/experience were measured using 10 questions in a Likert scale. The Likert items (from 1 to 6) were: no knowledge/experience, poor, below average, above average, good, and very good. The decision for using even scales was to minimize subjects' tendency to simply choose the middle value in the scales (central tendency bias) without prior consideration. Having no middle value in the scales, we expected that the subjects would seriously consider their answers to the questions.



(a) Sequence diagram in model M-Low



(b) Sequence diagram in model M-High

Figure 5.4: Example of sequence diagrams created using different LoD treatments

Feedback Questionnaire

In addition to the background questionnaire, the subjects were asked to fill out another questionnaire about their experience with the experiment. The feedback questionnaire was

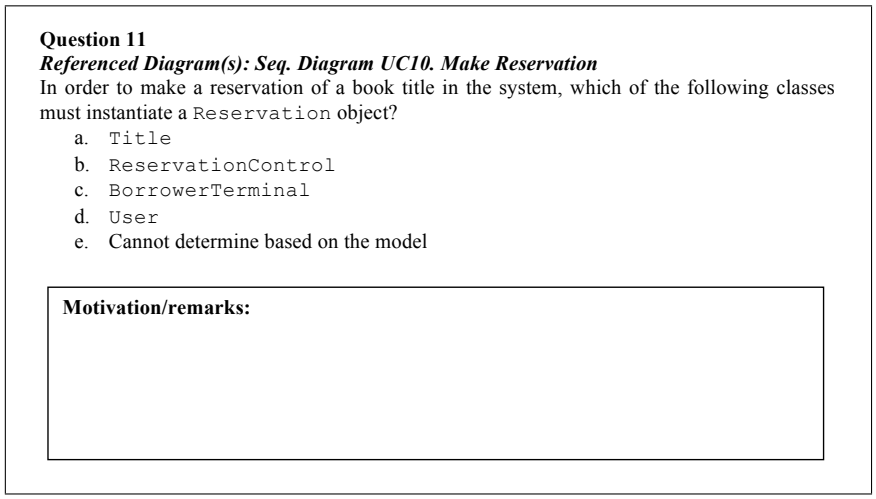


Figure 5.5: A sample question of the model comprehension questionnaire

aimed at obtaining personal feedbacks that might be useful for additional analyses or improvement of the experiment.

In the questionnaire we asked the subjects to judge the UML model they have received in terms of its complexity, comprehensibility, consistency, and clarity. We also asked the subjects to rate different aspects of the experiment from their individual perspective. Further, we asked the subjects to give any remarks/comments about the experiment. Except for the comments, the questions were all measured in a Likert scale.

5.5 Experiment Operation

The experiment was conducted in one day, and started at nine o'clock in the morning. We allocated 120 minutes to prepare and execute the experiment. Although there was no strict time limitation in performing the experiment tasks, the subjects were advised to finish the whole tasks within 90 minutes.

We started the experiment by assigning each subject to the treatment groups randomly, as discussed in Section 5.4.4. Subsequently, we briefed the subjects on how to work on the experimental tasks. The instructions were also written in the model comprehension questionnaire.

The first experiment task was to complete the model comprehension questionnaire and subsequently followed by completing the background questionnaire and feedback questionnaire respectively. In the experiment, there was no major deviation from the plan. It also turned out that all subjects could finish the entire task within the advised time duration. Once a subject handed-in the completed questionnaires, we checked the questionnaires for completeness and the subjects could only leave the premise when all the required tasks were

completed. We found this inspection very useful because we could then identified several subjects who forgot to complete the background and feedback questionnaires. In this case we asked the subjects to return to their seats and complete the tasks.

Additionally, in the inspection we immediately noticed that several subjects did not consistently register the time in the model comprehension questionnaire (recall that we required subjects to register starting time, ending time, and the time at several points during the completion of the model comprehension questionnaire). While missing starting and ending time could be easily fixed, it was tricky to fix missing time stamps during several points in the questionnaire.

After the completion of the experiment, we collected, preprocessed, and then stored the data in a spreadsheet.

5.6 Data Analysis and Interpretation

In this section we provide the results of the experiment. First, we discuss the analysis procedure. Subsequently, the results of the main statistical analyses will be discussed.

5.6.1 Analysis Procedure

The first step in the analysis was to preprocess the raw data obtained from the questionnaires. This activity was mainly related to inspecting and calculating the scores of all questionnaires. For example, for the model comprehension questionnaire, we needed to calculate the total number of correct answers for each subject. As the experiment was done on paper, the preprocessing activities were done manually. Once processed, the data was then loaded into a statistical tool for further analyses. In this respect, the statistical analysis tool SPSS [1] was used.

The next step after preprocessing the raw data was to explore the data. This step includes outlier analysis and checking whether the data sets met the assumptions required by the statistical tests. Because the hypotheses we aimed to test were differences of performance between groups of subjects, the statistical analysis techniques used were the ones for comparing means between groups (i.e., between two independent groups). Hence, the Independent t-test and Mann-Whitney test [83] were used for the parametric and nonparametric tests respectively. Note that we aimed for the parametric test because when all the assumptions are met, parametric tests (e.g., t-test) are more powerful (sensitive) than the nonparametric counterparts. Assumptions of normal data distribution and homogeneity of variance were tested using the Shapiro-Wilk and Lavene's test respectively. Data sets violating the normality assumption would be normalized using area transformation [75] prior to the analysis. For the statistical tests, we used significance level $\alpha = 0.05$ as a criterion for rejecting the null hypotheses.

In addition to the main hypothesis testing, we also performed a statistical test to investigate the influence of subjects' knowledge/experience on the dependent variables. For this analysis we used two-way ANOVA, and we used the same significance level (i.e., $\alpha = 0.05$)

Table 5.2: Descriptive Statistics of subjects' knowledge/experience, comprehension correctness, and comprehension efficiency across groups

Measures	L-LoD				H-LoD			
	N	Mdn	Mean	St.Dev.	N	Mdn	Mean	St.Dev.
Knowledge/Experience	26	31.00	31.50	8.62	26	34.50	34.07	9.57
Comp. Correctness	27	66.66	66.66	15.35	26	80.00	75.12	13.53
Comp. Efficiency	27	0.15	0.15	0.05	26	0.19	0.18	0.06

to indicate true significance. Furthermore, we were interested in the qualitative data obtained using the model comprehension questionnaire, i.e., subjects' statements to motivate each answer. For this qualitative data, the analysis was done manually by comparing differences of rationale amongst subjects and looking at trends that might appear in the provided rationale.

5.6.2 Experiment Results

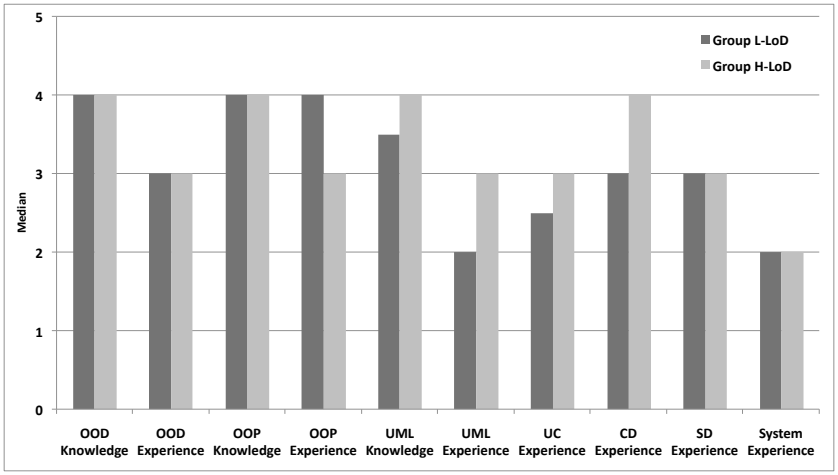
In this section we provide the results of the experiment. The results discussed here are based on the three questionnaires we have discussed previously. First, data obtained from the background questionnaire will be discussed. Subsequently, model comprehension analyses will be discussed. Finally, the results of in-depth qualitative analyses of the data will also be discussed. In Table 5.2, we provide some descriptive statistics of the results. If we consider the mean and median values in the table, we can see that there is a significant difference in comprehension correctness between the two experimental groups. However, there is little differences in knowledge/experience and comprehension efficiency between groups. We shall see from the statistical analyses whether the difference(s) are statistically significant.

Subjects' Knowledge and Experience

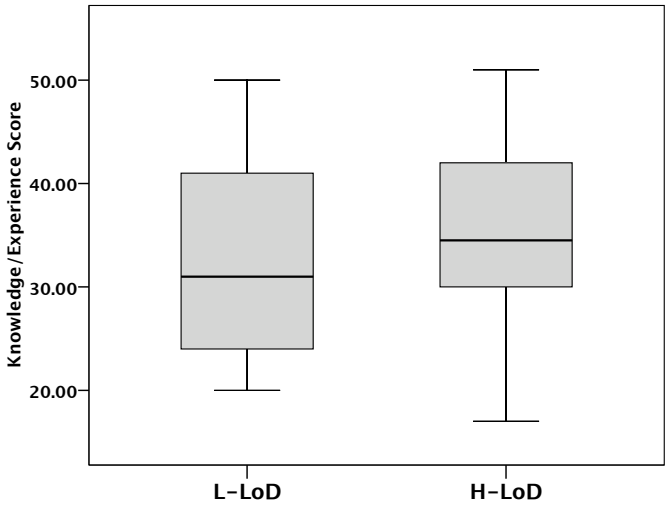
The subjects' background knowledge/experience was measured using the background questionnaire. The main aspects we measured have been discussed in Section 5.4.4.

Figure 5.6(a) shows the median values of the 10 questions asked to the subjects. As has been discussed previously, a six-points scale was used. The labels of the scales (from 1 to 6) are: no knowledge/experience - poor - below average - above average - good - very good. The figure is based on data from 52 subjects because one subject forgot to fill out the background questionnaire.

As can be seen in Figure 5.6(a), the subjects from group L-LoD and H-LoD have a reasonably good knowledge/experience (as indicated by the scale: 4 indicates above average). Furthermore, the subjects in both groups generally have equal *knowledge* about OOD, OOP, and UML. Nevertheless, we can also see in the figure that the subjects generally felt lacking of *practical experience* in the measured aspects—the lack of practical experience with UML and library system are the most prominent. However, it is interesting to note that practical experience with the UML diagrams used (i.e., use case, class, and sequence diagrams) is



(a) Median values of subjects' background knowledge/experience about different domains, across groups



(b) Box-plots summarizing subjects' knowledge/experience across groups

Figure 5.6: The profile of subjects' knowledge and experience

slightly better than general experience with UML. This phenomenon might indicate that the subjects have used the aforementioned UML diagrams and considered themselves as having more experience with those particular diagrams than with the whole concept of UML.

To investigate the extent to which the knowledge/experience of the subjects in both groups could have influenced the outcome of the experiment, we performed a statistical

Table 5.3: Ranks of knowledge/experience score across groups

Groups	N	Mean Rank	Sum of Ranks
L-LoD	26	24.90	631.50
H-LoD	26	28.71	746.50
Total	52		

Table 5.4: The results of the Mann-Whitney test showing the insignificance of the difference in knowledge/experience between the L-Lod and H-LoD groups

	Knowledge/Experience Score
Mann-Whitney U	280.500
Wilcoxon W	631.500
Z	-1.053
Asymp. Sig. (2-tailed)	.292

analysis to test whether the knowledge/experience gap between the two groups was significant. The knowledge/experience score for each subject was obtained by summing up the selected scales of all questions in the background questionnaire; thus, the minimum and maximum scores are 10 and 60 respectively. Because the knowledge/experience of the subjects was measured in ordinal scales, the Mann-Whitney test was used for the statistical test. The box-plots in Figure 5.6(b) summarize the knowledge/experience background scores of the subjects in both group L-LoD and H-LoD.

The results of the Mann-Whitney test in Table 5.3 and 5.4 confirm that there is no significant difference in knowledge/experience between the two groups—in particular, the insignificant difference is shown in Table 5.4 as the asymptotic significance (2-tailed) having the value of 0.292. A significant difference exists when the significance value is equal or lower than 0.05. Given this result, we were confident that discrepancies of subjects' knowledge/experience in the measured areas would have minor effects to the outcome of this experiment.

Testing Hypothesis 1: The Effect of LoD on Comprehension Correctness

This section provides analysis results related to the first hypothesis, which aimed at testing the effect of LoD on comprehension correctness.

For each question in the model comprehension questionnaire, there is only one correct answer. Hence, comprehension correctness represents the percentage of all correct answers. In examining the answers, the rationale/remarks for each question was also taken into account to assure that the answer was motivated properly. From our observation, most of the time subjects' answers were appropriately and sufficiently motivated, which increases our confidence on the subjects' seriousness in completing the questionnaire.

The box-plots in Figure 5.7 summarize comprehension correctness across groups. As shown in the figure, the median value of comprehension correctness of group H-LoD is higher than that of L-LoD (the median is indicated by the bold horizontal line in the rectangle).

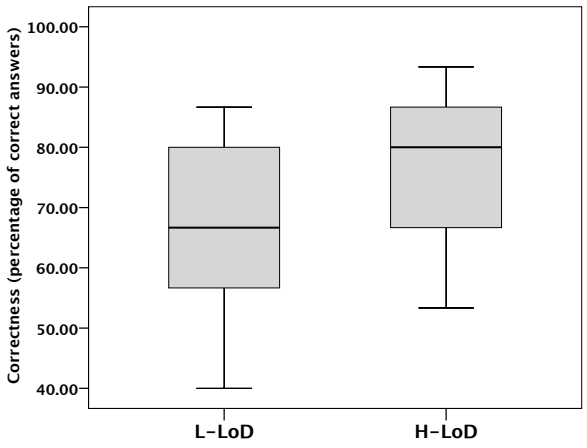


Figure 5.7: Box-plots of comprehension correctness between groups

As with the median, we can also see that the minimum value of H-LoD is also higher than that of L-LoD—the same is true for the maximum value. Overall, the box-plots suggest that the subjects in H-LoD have higher comprehension correctness than those in L-LoD.

To test whether the score difference between the two groups was significant, we performed an independent t-test. However, because the data set violated the normality assumption, we had to normalize it prior to the analysis. Once normalized, the data set met the required assumptions to perform the test (i.e., data measured at least on interval scale, normal data distribution, and homogeneity of variance). The results of the independent t-test are given in two tables. Table 5.5 summarizes the group statistics and Table 5.6 provides the main t-test result. Note that the comprehension correctness data set is a normalized data.

For this analysis we should look at row *Comprehension Correctness*. In Table 5.5 we can see that the average comprehension correctness for H-LoD was 0.297, which is higher than that of L-LoD (-0.289). Table 5.6 provides the result of the t-test to determine whether the mean difference was significant. The most important part of the table that indicates the significance is the *significance* column. As shown in Table 5.6, the mean difference of comprehension correctness between L-LoD and H-LoD was statistically significant ($p = 0.011$, 1-tailed). In other words, on average, subjects who received UML model with high LoD had higher comprehension correctness (mean=0.297, std. error mean=0.181), compared to subjects who received UML model with lower LoD (mean=-0.289, std. error mean=0.172), and this difference was statistically significant at 0.05 level ($p \leq 0.05$).

Having obtained the above results, we could reject the null hypothesis ($H_{1,null}$). Further, the H-LoD group performed significantly better than the L-LoD group in correctly comprehending the UML model, and therefore we had to accept the alternative hypothesis ($H_{1,alt}$): *the use of UML diagrams with high LoD significantly improves subjects' comprehension correctness.*

Table 5.5: Group statistics for comprehension correctness and comprehension efficiency

	Group	N	Mean	Std. Dev.	Std. Error Mean
Correctness (normalized)	L-LoD	27	-.289	.895	.172
	H-LoD	26	.297	.927	.181
Efficiency	L-LoD	27	.151	.053	.010
	H-LoD	26	.185	.060	.011

Table 5.6: The results of the independent t-test showing the significant effects of LoD on comprehension correctness and comprehension efficiency

	t	df	Sig.	Mean Diff.	Std.Error Diff.	95% Conf. Interval of the diff. Lower	Upper
Correctness (normalized)	-2.346	51	**.011***	-.587	.250	-1.090	-.084
Efficiency	-2.190	51	**.016***	-.034	.015	-.065	-.002

* indicates significance at 0.05 level (1-tailed)

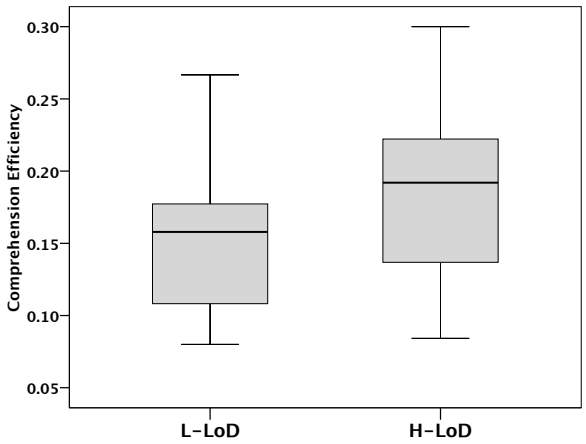


Figure 5.8: Box-plots of comprehension efficiency between groups

Testing Hypothesis 2: The Effect of LoD on Comprehension Efficiency

In the previous section we have seen that our first alternative hypothesis was confirmed. In this section we test the second hypothesis.

The second hypothesis concerns comprehension efficiency. Comprehension efficiency was measured as the relation between the number of correctly answered questions and the total amount of time spent to answer all questions. Hence, we essentially investigate whether subjects who received UML models with higher LoD required less time to correctly comprehend the model than those who received UML model with lower LoD.

Figure 5.8 shows that the comprehension efficiency of subjects in H-LoD is higher than

that of subjects in L-LoD. This can be seen from the median value of H-LoD that is higher than that of L-LoD. Furthermore, if we look at Table 5.5, which also summarizes the statistics of comprehension efficiency (see row *Comprehension Efficiency*), we see that the average comprehension efficiency of H-LoD was indeed higher than that of L-LoD. This result indicates that the comprehension efficiency of subjects in group H-LoD is higher than those in group L-LoD. Nevertheless, a statistical test had to be performed to evaluate whether the difference was statistically significant.

The same statistical test used in the earlier analysis, i.e., t-test, was used in this analysis. Unlike the previous analysis, the data set met all assumptions required by the statistical test, and therefore we could immediately run t-test on the data. The main results of the statistical test is provided in Table 5.6 (see row *Comprehension Efficiency*).

As shown in Table 5.6, the mean difference was significant, which is indicated by the significance value of 0.016 (1-tailed). Therefore, we can conclude that, on average, subjects who received UML model with higher LoD have significantly higher comprehension efficiency (mean=0.185, std. error mean=0.011) than subjects who received UML models with lower LoD (mean=0.151, std. error mean=0.010). This result has led us to accept the alternative hypothesis ($H_{2,alt}$): *the use of UML diagrams with high LoD significantly improves subjects' comprehension efficiency.*

Note that for validation purpose, we also analyzed mean differences of comprehension correctness (using the original data set) and comprehension efficiency between groups using the Mann-Whitney test. The results of the Mann-Whitney test were consistent with the results obtained using the t-test, hence their details are not reported in this chapter.

Influence of Subjects' Ability

In Section 5.6.2 we have seen that the difference of subjects' knowledge/experience between the experimental groups was not significant. However, to gain more insights about the influence of knowledge/experience on the results of the experiment, here we assess whether knowledge/experience has a significant effect of the dependent variables and whether it interacts with the LoD treatments.

We defined subjects' ability as a dichotomous variable with categories of *low* and *high*. We used a conservative criterion to determine the cut-off point: subjects with knowledge/experience scores below the mean (i.e., 33) were classified as low ability and above the mean as high ability. We subsequently used this variable as a co-factor for the LoD treatments. For the statistical test, we used two-way ANOVA because (1) it allows us to investigate interaction between ability and LoD, and (2) by introducing *ability*, the variation it explains in the data is removed, and therefore the analysis of the effect of LoD will be more powerful [10].

Table 5.7 and 5.8 show the results of two-way ANOVA tests for comprehension correctness & ability (CO-Ability) and comprehension efficiency & ability (EF-ability) respectively (as with the t-test for comprehension correctness, CO-Ability also used a normalized data set). The results show that neither ability nor the interaction between LoD and ability has a significant influence on comprehension correctness and comprehension efficiency. Further,

Table 5.7: The results of the two-way ANOVA for CO - Ability (using normalized data). The effect of LoD on comprehension correctness remains significant after the effects of Ability and LoD*Ability are accounted for

Source	Sum Sq.	df	Mean Sq.	F	Sig.
LoD	3.749	1	3.749	4.446	**.040**
Ability	.429	1	.429	.508	.479
LoD*Ability	.620	1	.620	.736	.395
Error	41.321	49	.843		

Table 5.8: The results of the two-way ANOVA for EF - Ability. The effect of LoD on comprehension efficiency is not significant after the effects of Ability and LoD*Ability are accounted for

Source	Sum Sq.	df	Mean Sq.	F	Sig.
LoD	.013	1	.013	3.969	**.052**
Ability	9.35E-04	1	9.35E-04	.281	.598
LoD*Ability	3.15E-03	1	3.15E-03	.951	.334
Error	13519.102	49	275.900		

consistent with the results in Table 5.6, the results in Table 5.7 show that LoD remains having a significant effect on comprehension correctness ($p = 0.040$).

Surprisingly, we can see in Table 5.8 that after controlling for the effects of Ability and the interaction between LoD and Ability, LoD was shown as not having significant effect on comprehension efficiency ($p > 0.05$). This result may indicate that even though ability has no significant effect on comprehension efficiency, the removal of its variation from the data set has reduced the significance of the effect of LoD on comprehension efficiency. Notice, however, that the significance value ($p = 0.052$) is only slightly higher from 0.05.

5.6.3 In-depth Analyses

In this section we further discuss the results of the experiment based on in-depth analyses of the data.

Per-question Comprehension Performance

To get a more thorough understanding concerning how the LoD treatments could have affected the subjects' performance, i.e., in terms of correctness in answering the model comprehension questionnaire, we performed an in-depth qualitative analysis of the subjects' answers.

Figure 5.9 provides a comparison of correct answers of all subjects in both group L-LoD and H-LoD. The figure shows an interesting phenomenon, in which subjects in L-LoD generally performed equally well as those in H-LoD. In some questions such as Q8 and Q15, L-LoD even outperformed H-LoD by six and four points respectively. Nevertheless, in some

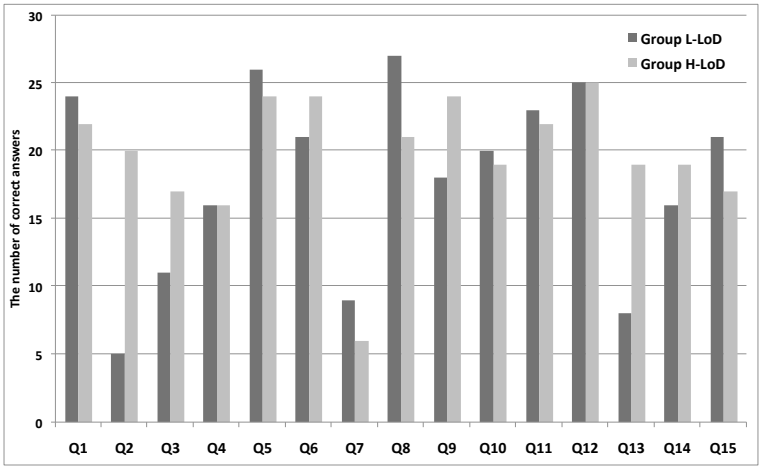


Figure 5.9: Score of all questions in both groups

other questions such as Q2, Q3, Q9, and Q13, subjects in H-LoD outperformed those in L-LoD by at least six points. Thus, these four questions might contribute largely to the difference of comprehension scores between the two experiment groups.

To understand why some questions were more often mistakenly answered, we carefully looked at subjects' rationale/reason of choosing certain answers. The information we obtained from the rationale/remarks written by the subjects shed light on how LoD in models affects model comprehension. For this analysis we chose some questions with prominent score gaps.

Questions that were poorly answered by subjects in L-LoD are Q2, Q3, Q9, and Q13 (these questions are provided in Appendix B). Except for Q13, all of the questions were related to implementation details. Furthermore, these questions mostly referred to sequence diagrams and, in general, they require subjects to answer implementation details about classes or objects. Although some subjects in L-LoD correctly answered the questions, many failed. From a careful analysis of subjects' rationales/remarks, mistakes made by subjects in group L-LoD were due to the following:

- Misinterpretation due to a missing message parameter in a sequence diagram (question Q2). Many subjects in group L-LoD failed to answer question Q2 correctly because of a missing message parameter in a sequence diagrams. Normally, missing or unclear messages in sequence diagrams can be crosschecked with the corresponding methods/operations in the class diagrams. However, because methods/operations in class diagrams were not specified in model M-Low, the subjects were forced to infer the correct answer from the information available in the sequence diagrams alone. While some subjects managed to answer correctly, many gave incorrect answers.

- Misinterpretation of a pseudo code in a sequence diagram (question Q3). Many subjects in group L-LoD misinterpreted a pseudo code *reservation.count* in a sequence

diagram. Many subjects understood *reservation* as a class, while it was actually mentioned as a concept; hence, it actually means 'the total number of reservation'. Further, we found that several subjects in group H-LoD also misunderstood the pseudo code. However, this ambiguity did not seem to confuse most subjects in group H-LoD, which was most likely due to the presence of class attributes in model M-High.

- Incorrect understanding of a class' role or responsibility. Many subjects in group L-LoD mistakenly answered question Q9 and Q13 because they misunderstood the role and responsibility of a class. We have observed that the misinterpretation was due to the absence of class attributes and operations, which indicate data contained in a class and the functionality it provides.

Apart from the misinterpretation made by group L-LoD, it is interesting to see that in some questions many subjects in group H-LoD made more mistakes than their counterparts. In this respect question Q8 is a prominent example. In question Q8 the subjects were asked to determine a class that holds certain information. Many subjects in H-LoD who answered incorrectly were confused by two classes that have an attribute with a somewhat similar name. Subjects in L-LoD, however, did not have attribute information in the class diagrams; hence, based their answers only on information in the sequence diagrams—which turned out to be leading to a correct answer.

Drawing a strong conclusion from the above observation is problematic because there seems to be more than just one factors that influence subjects' performance in answering questions—that is, question types, referenced UML diagram types, and, of course, the LoD treatments. These factors in combination might result in different degree of comprehension as compared to the individual factor in solitary. Although the design of our experiment does not allow us to further analyze the contribution of each factor to subjects' comprehension performance, the effect of low LoD in UML models seems to be more obvious when combined with questions related to implementation details. This means that the effect of LoD on comprehension is more significant in the contexts where UML models are used to guide the implementation. Hence, the impact may be less obvious when UML models are used for high-level analysis such as system architectural design.

Additionally, although UML models with low LoD generally have a negative effect on model comprehension, we have observed that its effect may be different on different people. In this regard, the effect of low LoD in UML models on model comprehensibility may be reduced under the following condition. Firstly, we have seen that a careful reading of the model artifacts to infer certain concepts is very effective to avoid misunderstanding of the model. A careful reading not only involves an in-dept examination of semantic and syntactic aspects of a diagram, but it also includes thorough assessments of a model by taking into account information from different diagrams. For example, subjects in group H-LoD who incorrectly answered question Q3 were found to be neglecting the information available in class diagrams. Secondly, we also believe that subjects' experience with or domain knowledge about the system being modeled can reduce misinterpretation of the model. Having a good knowledge of a system, a subject can make a good guess of certain concepts or business logics despite the low LoD used to model them. Finally, we also think that the awareness of

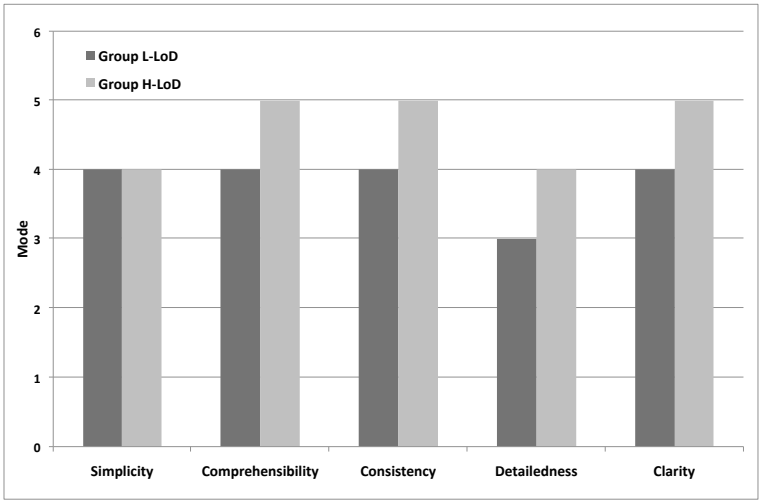


Figure 5.10: Subjects' perception on the UML model

the subjects about the quality level of the model, in this respect level of detail, also plays a role in their decision making process. That is, subjects who are aware of the fact that they receive a UML model with low LoD might be more cautious in inferring the concepts in the UML diagrams than those who are not aware of the LoD.

Subject's Feedbacks

In this section we provide an analysis of data obtained from the feedback questionnaire. As we have discussed previously, in the feedback questionnaire we asked the subjects to judge certain aspects of the UML model that comprise simplicity, comprehensibility, consistency, detailedness, and clarity of aspects of the model.

Data obtained from the questionnaire is presented in Figure 5.10. The figure shows the mode value of all questions. Except for the forth question, which was concerned with the detailedness of the UML model, the subjects in group L-LoD and H-LoD tended to rate other aspects of the UML models positively, i.e., the models were generally simple, comprehensible, consistent, and clear. In the forth question (related to model detailedness), subjects in group L-LoD generally rated the UML model somewhat negatively. This result seems reasonable given the LoD of the UML model they received in the experiment. Further, note that the subjects belonging to group H-LoD tended to rate the provided aspects, except simplicity, more positively than those in group L-LoD.

Apart from the fact that the results presented in Figure 5.10 are subjective perceptions of the subjects, it is important to note some interesting points. Firstly, it seems that model simplicity, as the subjects perceived it, was not influenced by the LoD used in the model. This finding is interesting because it poses a question whether a reader's perception about model complexity is indeed independent from the amount of information contained

in the model. Secondly, although subjects in group L-LoD tended to rate the detailedness of the model relatively low, they had a positive perception about the comprehensibility of the model. While this phenomenon may seem contradictory, it actually indicates subjects' unawareness of the effects of LoD on their real comprehension performance. In other words, the subjects were not aware that they have misunderstood the model, which was due to the limited information available in the UML models.

5.7 Discussion

In this section we provide further discussions about the results and we identify their implications on both research and practice. Additionally, we discuss validity threats to this study.

5.7.1 Reflection on the Results

Very few studies have been done to investigate the impact of UML model quality on software development. Our earlier studies reported in [95, 96] focused on explorative studies into how the quality of UML models is managed in practice and the perceived impact of styles and rigor in UML modeling on productivity and quality. Our current study extends the above studies by experimentally investigating the impact of styles and rigor, i.e., LoD in modeling, on model comprehension. One of the results of this study, i.e., the impact of LoD on model comprehension, is in line with the results of an earlier work reported in [27]. In their study, Briand et al. have found that the use of OCL in UML models improves reader's comprehension of the models. Although the treatment used was different from the one used in this study, applying OCL in UML models is essentially increasing the amount of information and rigor in specifying modeling construct; hence, it fundamentally shares the same notion as increasing LoD in UML models. We should note, however, that OCL is seen as an extension (an extra feature) that may be used to increase the rigor of UML models. Whereas LoD addresses the rigor of core UML modeling constructs that are found to vary in practice.

The results of this experiment suggest three important points. Firstly, LoD, which represents the amount of information that is used to specify UML models, is influential to a correct comprehension of the models. The amount of information in a model can be increased/reduced by being more explicit/implicit in portraying modeling constructs using UML modeling notations. Hence, it is essentially true that by increasing/decreasing LoD in a model the rigor of the model also increases/decreases. Secondly, we have observed that models with low level of detail provide potential for misinterpretation by the readers. If the readers remain unaware of the misinterpretations, they might implement models incorrectly, thus introducing defects in software—see for example our work that investigates the relation between LoD in UML models and defect density using an industrial case study [98]. Finally, on average, subjects receiving UML models with high LoD have higher comprehension efficiency.

While we observed different significance values (p-values) between the t-test and two-

way ANOVA, ANOVA reported p-value of 0.052—hence, not significant, the departure from 0.05 is almost negligible. Therefore, we remain confident that LoD has a significant effect on comprehension efficiency. What is more is the fact that we did not observe subjects' knowledge /experience as a factor or co-factor that influences comprehension correctness and comprehension efficiency. We believe this result was due to the relatively homogenous knowledge/experience of the subjects' in this experiment.

Additionally, we need to underline that certain modeling elements (e.g., class attributes, message parameters, and conditional guards) may be crucial for a good comprehension of UML models. From our observation, the presence of message parameters and conditional guards in sequence diagrams helps to better understand object interactions and logics of the application. Additionally, attributes and operations in class diagrams are primarily useful as a cross reference when information in the sequence diagrams is not clear. For classes that pertain to novel concepts, class attributes may also help to infer the roles of those classes. However, also note that the effective use of LoD in UML models may depend on the purpose of the model. For example, high LoD may be effective for UML models that are used for implementation guide, but not for models used for high-level architectural design, and vice versa.

5.7.2 Implications for Research and Practice

The significance of the results of this study for research in the area of software quality is prominent. The results of this study suggest how the amount of information used in modeling might affect the readers' comprehension of the model. In the contexts where models are used to communicate design decision or to guide implementation, model comprehensibility is crucial. These results should invite more research to further investigate aspects in UML models that can be used to improve model comprehensibility. Furthermore, the relation between the quality of UML models and the quality of the resulting software is still not well understood. Our work (discussed in Chapter 6) is one of first studies that aim to answer this important question.

With respect to modeling practice, the results of this study motivate the importance of informed decision in using style and rigor in modeling. The use of style and rigor in modeling always come at a price. Formal style and rigor lead to higher modeling effort, but might payoff in terms of more comprehensible models. On the other hand, informal modeling styles can save time and effort, but might lead to problems related to interpretations of the models. Therefore, software designers should be aware of the trade-off and subsequently make informed decisions to target the quality levels of their models. In this respect, our recommendation has been to apply more details to parts of models that are complex, critical, or pertain to important concepts that are new to the readers (e.g., developers) [96].

Given the fact that LoD in UML models is very diverse in practice, we recommend the following in order to reduce the impact of low LoD: to perform a careful reading of UML models (possibly with reading techniques), to have a sufficient domain knowledge about the system being modeled, and to be aware of the quality level of the models.

5.7.3 Threats to Validity

In this section we discuss the types of validity threats related to this experiment. We present them in the order of their priority [130].

Internal Validity

The main threat to the internal validity of this experiment comes from differences between the experimental groups, such as knowledge, experience, and motivation. Nevertheless, in Section 5.4.4 we have identified some major confounding factors that might affect subjects' performance, and in Section 5.6.2 we have also examined that in terms of the identified factors, there was no significant difference between the experimental groups. Additionally, in Section 5.6.2 we further analyzed the influence of subjects' ability on the dependent variable, but no significant influence was observed. Apart from that, we are aware that some factors were not measured in this study, such as subjects' general comprehension skill. This skill might vary amongst subjects and may subsequently affect their performance in this experiment.

External Validity

External validity threats are related to limitations to generalize the results of an experiment to industrial practice. As with other experiments with students, we must be careful in generalizing the results of this experiment to industrial professionals. In this case, because the subjects were graduate students who have sufficient knowledge about UML, we are more concerned with their experience in applying UML to real problems. However, we believe that the subjects' experience might be less of an issue for understanding UML models than it might be for creating UML models. Apart from that, a related study by Lange and Chaudron has shown that students and professionals perform equally good in reading and comprehending UML models [78]. Hence, we may expect similar results if we run the same experiment using professionals software engineers as subject. Another threats to the external validity is the use of a simplified UML model. Although, we have tried to make the model comparable to industrial models (e.g., of size and complexity), some simplifications were made to fit the experiment with the allocated time.

Construct Validity

Threats to construct validity in this study is mainly related to the extent to which the model comprehension questionnaire really measures the variables that we would like to measure, namely model comprehension. In this respect, we have carefully designed the questions in such a way that they capture different aspects of subjects' comprehension of the UML model. The questionnaire was also pilot-tested to ensure that every question is understandable to the readers. Additionally, in measuring subjects' comprehension correctness we carefully assessed subjects' rationale for each answer, thus allowing us to validate their true understanding of each answer.

Another threat to the construct validity concerns the measurement of subjects' ability. The background questionnaire that was used to measure subjects' ability may not perfectly capture subjects' real ability. This is particularly true because subjects may over/under estimate their knowledge or experience in the questionnaire. While students' grades from previous courses could have been useful as additional evidence, such information was not available for us to be used in the analysis.

Concerning comprehension efficiency, we have seen that some subjects used extra time to recheck their answers; in this respect, some of them reported the amount of time for this activity, but some other might not. There were other occasions where the subjects might have had short breaks during the execution of the experiment. Hence, the time duration may contain some noise. Nevertheless, subjects were aware of the importance of finishing the tasks timely, and thus they tried to complete the questionnaire as soon as possible. Therefore, we believe that the noise in time-duration data is not substantial.

Conclusion Validity

Threats to conclusion validity relate to ability to draw a correct conclusion from an experiment. This validity threat includes subject selection, data collection, measurement reliability, and validity of the statistical test. In this study we have addressed all factors that might have threatened the conclusion validity of this study through a careful design of the experiment and rigorous procedure in data analysis.

5.8 Conclusion and Future Work

In this chapter we report our empirical investigation into the impact of level of detail (LoD) in UML models on model comprehension, which was measured in terms of *comprehension correctness* and *comprehension efficiency*. This study was based on an experimental study using 53 M.Sc. students majoring in Computer Science at the Eindhoven University of Technology, the Netherlands. Having applied two versions of a UML model with different LoD to two independent groups, we have found that the group receiving a UML model with higher LoD comprehended the model better than those receiving a UML model with lower LoD. More specifically, the results show that applying higher LoD to a UML model significantly improves the *correctness* and *efficiency* of subjects in comprehending the UML model. Additionally, we observe that the effect of LoD on comprehension is more significant in the contexts where UML models are used to guide the implementation.

We recognize that the result of this study is not yet conclusive. Hence, further work is needed to replicate this study, particularly in settings that involve professional software engineers. Furthermore, based on our observations we think that certain modeling notations are more effective to improve model comprehension than others. In this respect, we encourage further research to empirically investigate which modeling notations are influential to improve model comprehension.