"An Experimental Evaluation of the Impact of System Sequence Diagrams and System Operation Contracts on the Quality of the Domain Model"

By

Reymes Madrazo-Rivera

A thesis submitted to

The Faculty of Graduate Studies and Research

In partial fulfillment of the requirements for the degree of

Master of Applied Science in Electrical Engineering

Ottawa-Carleton Institute of

Electrical and Computer Engineering

Department of Systems and Computer Engineering

Carleton University

Ottawa, Ontario, K1S 5B6

Canada

January 2007

Your file Votre référence
ISBN: 978-0-494-26998-5
Our file Notre référence
ISBN: 978-0-494-26998-5

Abstract

The Unified Modeling Language (UML) is an object-oriented analysis and design language widely used to describe artifacts which are created during the software system lifecycle. It standardizes several diagramming conventions associated with common design methodologies. UML must be applied in the context of a specific software engineering process. The Unified Process (UP) is one such process, extensively used by the object-oriented community, which enhances team productivity and delivers software best practices via guidelines for all software lifecycle activities. The UP suggests many artifacts to be obtained during the software lifecycle. But, nowadays many practitioners are reluctant to use those artifacts as they question their benefits. System Sequence Diagrams and System Operation Contracts are some of the key artifacts which are part of UP [1]. This thesis presents the results of controlled experiments that investigate the impact of using System Sequence Diagrams and System Operation Contracts on software development. One way to do that is to study the extent to which those artifacts improve the quality of the Domain Model or reduce the effort to obtain this Domain Model.

Results show that the use of those System Sequence Diagram and System Operation Contracts significantly improves the quality of the Domain Model. On the other hand, there is no noticeable evidence that those two artifacts reduce the time to obtain the Domain Model.

Acknowledgements

This thesis would not have been possible without the tremendous support of many people.

This thesis is complete because of Dr. Lionel Briand and Dr. Yvan Labiche who not only served as my supervisors by providing me great guidance and support but also encouraged and challenged me throughout my studies.

I would also like to thank everyone at the Squall Lab for being supportive and friendly. Throughout this time, I have had a great time and enjoyed every moment of my time in the lab and had the chance of working with (in alphabetical order): Zaheer Bawar, Micheal G. Bowman, Hongyan Chen, Bin Dong, Maged Elaasar, Vahid Gharousi, Xuetao Lie, Yanhua Liu, Samar Mouchawrab, Marwa Shousha, Alex Sauve, Mike Sowka, Tao Yue, and Gregory Zoughbi.

Thanks to all my Cuban friends (Maite, Yaquelin, Dianita, Niurvis, Irene, Mimi, Lourdes, Liverkis, Virgy, Daimi, Belkis, Reinel, Ivan, Wisdenilderd, Mae, Joel, Yoemil, Frank, Luis Enrique, Thomas, and many others whose names will fill pages) for their remarkable support.

Special thanks to Dr. Sorina, Dr. Mihai-Raul Gherase, Dr. William Bill Mac Kenzie, Prof. Shirley E. Mills, and Prof. Kim Davis for providing extensive feedback on my thesis.

Last but not least, I would like to particularly thank my parents, Ana Carina Rivera and Jorge Madrazo and my brothers and sister (Jorge, David, Reyjor, Allen, and Tanya) for their continuous and generous encouragement.

Content:

List of tables

List of Figures

1 Introduction

According to Bruegge and Dutoit [2], a software life cycle is the "set of activities and work products necessary to develop a software system." Common activities are: Requirements Elicitation, Analysis, Design, Implementation, Testing, Installation and Maintenance [21]. Each one is usually carried out by different stakeholders, in the framework of a specific software process development, and involves the production of substantial documentation.

The overall picture for software practitioners has become complex given the large number of development methodologies and notations. Examples include Unified Process (UP), Rapid Prototyping, Extreme Programming and Object-Oriented Process Environment and Notation (OPEN). Each one has its own characteristics and set of steps to follow as a way of properly addressing the user requirements by ensuring a high quality level of the final product while accounting for budget and time constraints. An exception is the Unified Modeling Language (UML) notation, which has become a de facto standard for the modeling of software systems [1].

Victor Basili [21] pointed out the dramatic growth over the past thirty years of the information which has to be accounted for when developing software. And one just could wonder whether all that information is really necessary. There is the tendency for stakeholders to avoid any form of formal documentation because they typically question whether it is worth the time and offer tangible benefits [10, 23].

Unfortunately, there is a lack of empirical studies to support the usefulness of formal or semi-formal documentation during software development. Despite the quick pace with which Software Engineering technologies change, experimentation is considered one of the adequate ways to evaluate how well new artifacts, methods and techniques work [7]. This thesis is a small step in addressing the issue of determining what UML artifacts are necessary to ensure proper modeling. It reports on the results of four experiments, performed in a university setting, on the impact of using specific UML diagrams during UP-based development. Concretely, we examine three artifacts that Larman [1] recommends in the context of the UP: The Domain Model (DM), System Sequence Diagrams (SSD) and System Operations Contracts (SOC). The thesis goal was to answer two questions: 1) whether the use of System Sequence Diagrams and System Operation Contracts improves the quality of the Domain

Model, and 2) whether using those artifacts leads to a decrease of the effort invested in obtaining the Domain Model.

By analyzing the results of the experiments two conclusions were reached. First, the use of System Sequence Diagrams and System Operations Contracts improves significantly the correctness of the Domain Model. Second, by using those artifacts one does not save effort invested in obtaining the Domain Model. Even though, as a whole, slightly less time was required when these artifacts were used than when they were not used.

The thesis structure is as follows. Chapter 2 provides some background information for the research that was conducted, which involves topics related with software engineering, experimentation, and statistical testing. Chapter 3 presents the methodological aspects of the experiments and reports on possible threats to validity. All analysis results are presented in Chapter 4. Finally, Chapter 5 concludes and discusses opportunities for future investigations.

2 Background

This chapter can be used by readers as a reference to get familiar with the different terms used throughout the rest of the thesis. Section 2.1 presents an overview of the Unified Process (UP). In Section 2.2 we discuss some important and general concepts associated with empirical studies. An overview of all statistical tests used to analyze the collected data is introduced in Section 2.3.

2.1. The Unified Process

As it is well known, there exist many object-oriented software process models to guide the entire software life cycle [1, 3, 4]. Nevertheless, this study only focuses on Unified Process as it is a process suggested by one of the prominent authors in the field [1].

2.1.1 Basic Principles of the Unified Process

The Unified Process (UP) is an iterative use-case driven and object-oriented process that uses the UML notation to describe its models [1, 3, 7]. According to Larman [1], UP is an iterative approach to software development where the entire system is subdivided into sub-systems called iterations. Each iteration follows the usual development steps: requirement elicitation, analysis, design, implementation and testing. Each subsequent iteration augments over the previous one, with more details and improvement taken into consideration [36].

Usually, the functional and non-functional requirements of an application are documented during the Requirements Elicitation. These requirements are used to write the Uses Cases [1]. Functional requirements refer to the set of functionalities/operations supported by the system [2]. Non-functional requirements cope with the set of non-functional constraints and properties of the system such as performance [2, 36]. Uses Cases are carefully described usage scenarios, describing the main functionalities of the system and its interactions with external actors [1, 26]. UML has extended this definition to diagrams that show relationships between those use cases, and their interaction with the actors of the application [1], where actors are roles adopted by external entities that interact directly with the system under development.

The next step in the software development lifecycle is usually the Domain Model definition. The Domain Model is a description of the problem domain under study where the main components or

concepts of that domain problem are identified and defined, and then classified as conceptual classes. For the most important conceptual classes, the main attributes as well as their associations are also usually identified and described [1, 3].

2.1.2 Larman's extensions to the UP

One of Larman's extensions to the UP is the introduction of two artifacts at the end of Requirement Elicitation, and beginning of Analysis: System Sequence Diagrams (SSD) and System Contracts (SC)—also called System Operation Contracts. They will be referred to from now on as "the artifacts", unless otherwise specified. Both artifacts were added to the Unified Process as a way of refining the resulting Use Cases [1].

Larman defines System Sequence Diagrams in [1] as artifacts that specify input and output events related to the system under study. It is also an explicit graphic representation of the main scenarios of a Use Case, which shows an exchange of events between the system and the actors. A System Sequence Diagram "is a way of expressing what the system does without explaining how it does it" [1] and Larman recommends their use for the most complex alternative Use Case Scenarios.

Figure 1 shows the main elements which compose a System Sequence Diagram. Specifically, a System Sequence Diagram (SSD) is a sequence diagram (one can use the UML sequence diagram notation) that shows the system as a whole, as an object, that receives messages from actors and responds to them. No detail is shown as to how the system actually responds using for instance boundary, control, and entity classes. Figure 2 illustrates through a SSD the main success scenario of the VideoRental Use Case. This is one of the Use Cases modeled for the Video Store software system used in our experiments.

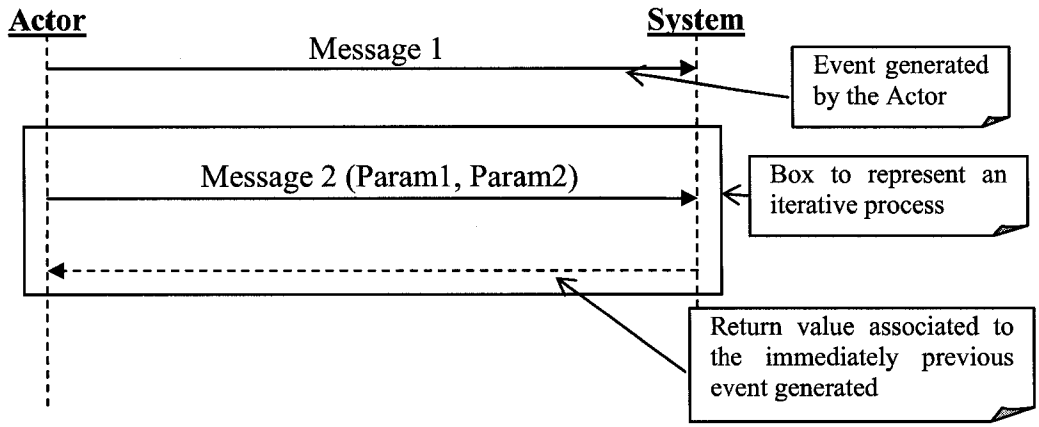

Figure 1 Main elements to represent a System Sequence Diagram

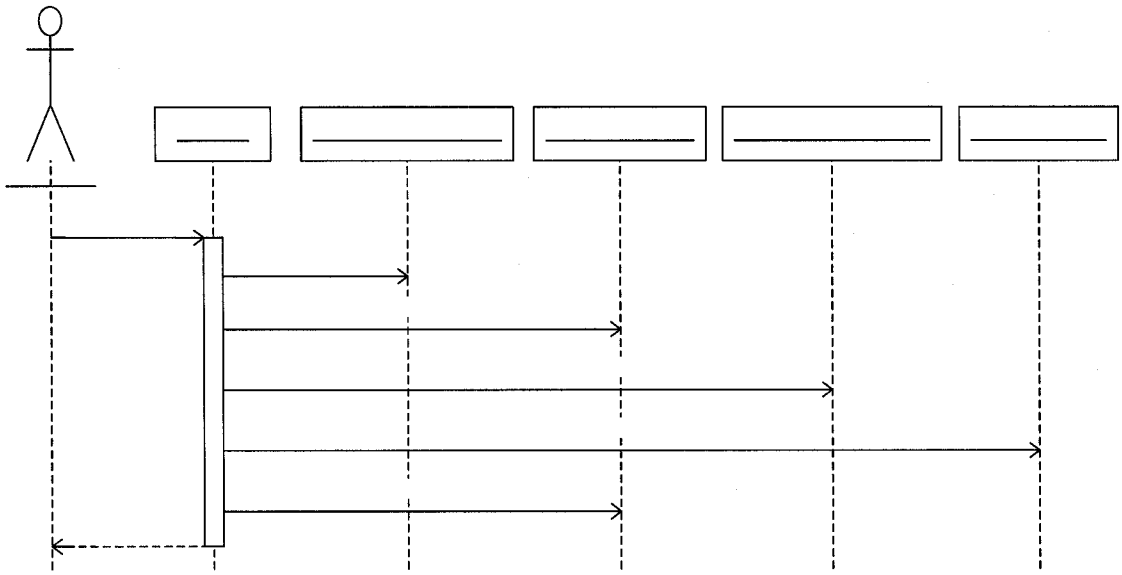

Figure 2 SSD for Rental Video scenario

The intent of the second artifact, System Contracts, is to illustrate those changes that occur in the state of the system, i.e., changes to the objects defined in the Domain Model as a result of the arrival of some system event defined in the System Sequence Diagram. Figure 3 shows the template taken from [1], which is used to describe a System Contract. Figure 4 shows a concrete example using this template. This example refers to the same VideoRental Use Case previously mentioned. The pre and post-conditions are defined using the Object Constraint Language (OCL) that is part of the UML standard.

- **Operation**: System event
- **Cross references**: Use Cases where the operation can occur
- **Pre-conditions**: Sates of objects in the Domain Model before operation execution
- **Post-conditions**: State of objects in the Domain Model after an operation completing. Instance creation or association formed.

Figure 3 Template for System Contracts

- **Operation**: rent(): Boolean
- **Cross references**: VideoRental
- **Pre-conditions**:
- **Post-conditions**:

```
let barcode:String = BC_Scanner.scanBarCode()
let id:Integer = CS_Reader.RaedCardStrip()
let toady:Date = CCS_Interface.getCurrentTime()
in
Rental.allInstances->exists(r| r.copy.barCode=barCode and r.member.id=id
and r.whenMade=today and r.copy.state=#Rented
and not Rental.allInstances@pre→includes(r)) and result=true
```

Figure 4 Example System Operation Contract

It is argued that those two artifacts provide a better understanding of the system under study and that they could improve the quality of the next step in the software life cycle, that is the Analysis phase, and more specifically the construction of the Domain Model. Larman discusses for instance how the use of those elements leads to the detection of new design element into the Domain Model. But he also recommends their use exclusively for the most complex system operations. Other authors (e.g., [5]) concur, indicating that using System Sequence Diagrams and System Contracts leads to a much more complete and precise Class Diagram being built in the Analysis phase. Recall that building the Domain Models is the first step towards building the Class Diagram during Analysis.

Empirical studies are necessary to investigate the above statement by getting quantitative data to evaluate the actual influence of those two artifacts. That is why a controlled experiment was designed in order to verify if there exists a significant improvement in quality and decrease in effort invested in the Analysis stage when those artifacts are used.

2.2. Designing experiments

There are essentially three strategies to follow when carrying out empirical studies: Surveys, Case Studies and Experiments [7]. The last one was used for this thesis. An experiment is a controlled study, normally performed in a laboratory environment, which explores the effect of some input—independent variable or factor—to the environment or the output—dependent variable or response variable [7, 27]. Controlled experiments allow checking over extraneous variables in such a way that the experimenter is able to eliminate their effects on the dependent variable so that the isolated effect of changes in the independent variable on the dependent variable can be evaluated. In general, any experiment strictly involves the following steps: definition, planning, operation, analysis & interpretation, and presentation of results [7, 33].

When performing an experiment, the experimenter holds the belief that there exists a cause-effect relationship between some inputs and outputs. That is why the starting point of any experiment is the hypothesis, which expresses this relationship. The hypothesis is the basis for further statistical analyses which quantify how well the observed data stands in agreement with a given predicted probability [28]. Two hypotheses have to be stated and they are essentially defined as follows [7, 13, 27, 28]. One of them is referred to as the *null hypothesis* (H_0), and it is the hypothesis that the experimenter wants to reject. It usually states the absence of patterns in the experiment settings. The other one is the alternative hypothesis (H_a), and it is the hypothesis in favor of which the null hypothesis is rejected. The experimenter is interested in finding enough evidence suggesting that H_0 is false.

When testing hypothesis, two types of risks are involved with H_0; Type_I_Error and Type_II_Error. The first error has to do with the probability of rejecting H_0, even though H_0 is true, whereas, the second error is the probability of not rejecting H_0, even though H_0 is false. The probability of rejecting H_0 if H_0 is true (Type_I_Error) is denoted with α. α is called the significant level of the test [27, 28]. This α value, as it is expected, has to be as small as possible (a value of 0.05 or 0.01 is usually used as a threshold) and it means that if H_0 is true and the same test procedure is performed for different samples in the same population, then H_0 could be incorrectly rejected only 5% or 1% of the time, respectively.

To decide between H_0 and H_a, a statistical test is constructed and computed, as a function of the measured variables. Commonly, a p-value—also called observed significance level—is calculated. This p-value is the probability, assuming H_0 is true, of obtaining a test statistic value at least as contradictory

to H_0 as the actual situation [27] (i.e. there is a statistical association between the variables that were examined). That is why, a small p-value means that there is a small chance that H_0 is true and as a consequence the null hypothesis should be rejected if the p-value is too small. How small? That is determined by a predefined α value. H_0 is rejected if p-value $\leq \alpha$.

When designing an experiment there are three principles of strict fulfillment that limit the bias, the experimental errors and the effect of external factors [7, 16, 27]: Randomization; Blocking; Balancing. The first principle refers to randomly assigning the subjects of the experiment to the different treatments under study and the objects to use in order to avoid biased results. The Blocking principle is a way of eliminating undesired effects in the experiment by screening out the suspicious external effects. The Balancing principle refers to assigning each treatment with the same number of subjects. Note that some statistical tests can be used even when the design is unbalanced, which is the case of our experiment (See Sections 2.3), without introducing major threats to validity.

2.3. Review of relevant statistical tests

There are many statistical analysis techniques available. Which one is most appropriate closely depends on the experiment design being used. Making the right choice is paramount as the selection of an inappropriate technique may lead to an erroneous data interpretation. In this section, only the statistical tests that were used for the experiment are reviewed.

Firstly it has to be mentioned that in general statistical tests are categorized as parametric and non-parametric [7]. Parametric tests are those dealing with data samples that follow a normal distribution and whose groups of subjects have the same size as well as a homogeneous variance. One-Sample *t*-test and Paired *t*-test are examples of parametric tests. Non-parametric tests, also called distribution free tests, are those that do not make any assumption about the data under analysis. Mann-Whitney test and Wilconxon Signed Rank test are examples of non-parametric tests. One should use a parametric test instead of a non-parametric test when parametric conditions apply because non-parametric tests use a very conservative approach when significance is computed. The use of non-parametric tests when the data meets parametric conditions may lead one to inappropriately accepting the null hypothesis.

Statistical tests are also classified as univariate and multivariate. Univariate tests, such as Independent *t*-test and Paired *t*-test, are used when only one independent variable is studied at a time. On the contrary, multivariate test are those where two or more independent variables are analyzed. For

example, Two-Way ANOVA and Mixed Model Repeated Measures ANOVA tests are multivariate tests.

A number of statistical test are used in this thesis, namely One-Sample *t*-test, Two-Sample Independent *t*-test, N-Factorial Way ANOVA (e.g., Two-Way ANOVA and Three-Way ANOVA) and Repeated Measures ANOVA (e.g., Simple Repeated Measures ANOVA and Mixed Repeated Measures ANOVA). They are now described.

One Sample *t*-test is used to compare the mean score of a given sample to a known value [13]. For example, it could be used to compare a group of teenagers' weight to a known value representing the whole teenager population in the city. We used this test to analyze the subjects' comprehension level of the systems for which they are supposed to obtain their Domain Model (see Section 3).

Two-Sample Independent *t*-test is applied to compare the mean outcome variable between two independent groups [16, 27]. We use it to analyze how different was the performance of those subjects who used the artifacts with respect to those who did not use the artifacts, by independently considering measures of each attempt of obtaining the Domain Models.

N-Factorial Way ANOVA is a statistical test that also allows detecting differences between several populations or processes means, by basing its algorithm on a comparison of variances [7, 16, 27]. Two-Way ANOVA and Three-Way ANOVA are the variants used in our experiment. Those tests allow designing experiments where more than two factors of interest are simultaneously compared. That is, those two tests allow characterizing populations based in two and three factors (i.e. independent variables) respectively, where the hypothesis testing situation involves k-independent samples [41]. A special case of N-Factorial Way ANOVA analysis is that one with a randomized block design. In this kind of design there exist one single factor of primary interest and there are some others factors that are conveniently introduced to control for the influence of the extraneous factors on subjects taking part in the experiment (e.g., variations in subjects' expertise) [13]. (See Section 3 for additional details).

Conducting an experiment with repeated measurements (i.e., where subjects perform a task several time) is considered a proper procedure to reduce errors introduced by outside factors that may bias the final results [27]. Using N-Factorial Way ANOVA test to analyze the data corresponding to such an experiment is not appropriate because the data violate the ANOVA assumption of independence (the same subjects perform the tasks several times and therefore the data are not obtained in an independent

way). For those cases, specific statistic tests should be used, such as, Simple Repeated Measures ANOVA and Mixed Repeated Measures ANOVA tests.

Repeated Measures ANOVA tests in general are used to check for significant differences for those experimental designs where the subjects, that are randomly assigned to a control or a treatment, have an outcome measured two or more times during a longitudinal research [16, 18, 19, 20, 29], each time under different experimental conditions. Those tests are popular for many reasons. Such experiments are useful because they do not require a large number of subjects to draw a conclusion about the treatment effect (because subjects perform tasks several times). Authors suggest, when applying any version of Repeated Analysis ANOVA tests, to use a sample size larger than 30 subjects [30]. These tests also provide a reliable result even when the experiment design consists of an unequal number of subjects (i.e. unbalanced designs), by assuming that such values are randomly missing [37]. They are able to deal with unbalanced data designs because they estimate linear combinations of those missing level combination values by using the Method of Restricted Maximum Likelihood (REML) (also known as Residual Maximum Likelihood) when trying to fit the data to the selected Covariance Structure [25]. These statistical tests are also stronger than the previous two statistical tests when looking to determine if there exists a significant difference between Method's treatments because subjects act as their own control in this kind of experiments. When subjects serve as their own control, variability between subjects is isolated, which allows the analysis to better focus on treatments effects [16, 17, and 18].

Two types of Repeated Measures ANOVA tests were considered for this experiment: Single Repeated Measures ANOVA and Mixed Model Repeated Measures ANOVA. The first one is used to check a set of dependent samples where those samples represent populations with different means [41]. Mixed Model Repeated Measures ANOVA is a kind of N-Factorial ANOVA test with the addition that it is applied to study the significance of main factor effects and of interaction between factors when dealing with dependent samples [39]. It is worth remembering that "dependent samples" refers to measuring the same subjects more than once, for each of the experimental conditions when executing the experiment.

3 Experiment planning

The steps to follow when conducting a controlled experiment, according to [7, 34], are depicted in Figure 5. The next sub-sections walk the reader through the first steps of the experiment process. The last two steps are simply implemented in section 4.

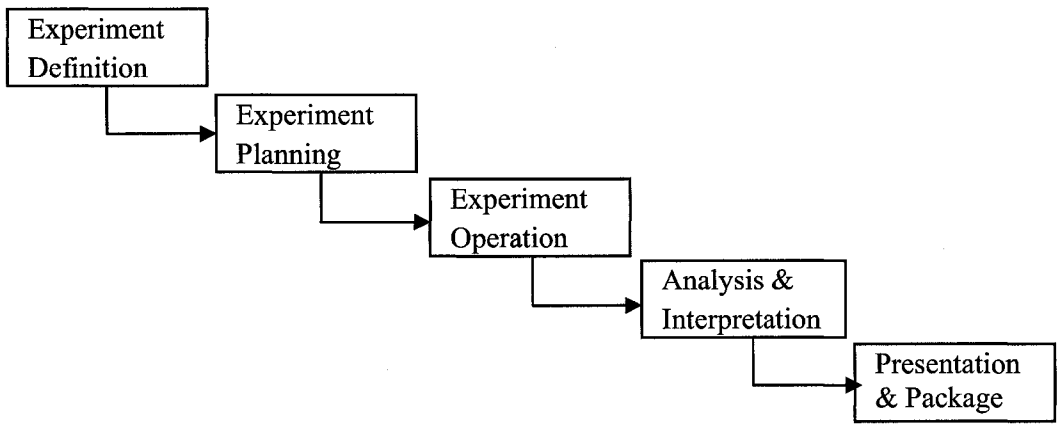

Figure 5 Steps to conduct an experiment

Four experiments were conducted. The first during Summer 2005, the second during Fall 2005, the third during Summer 2006 and the fourth one during Fall 2006. All experiments are referred to as Experiment I, Experiment II, Experiment III and Experiment IV respectively hereafter.

Section 3.1 provides the experiment definition, where the goals of the experiment are clearly stated. Section 3.2 provides the experiment planning, which constitutes the base of any experiment, describing such things as the context and validity of results.

3.1. Experiment definition

This experiment intended to evaluate the impact of using System Operation Contracts and System Sequence Diagrams on obtaining a much more complete Domain Model. The independent variable or factor of interest was defined as Method. This variable refers to how the subjects will be obtaining the Domain Model, some of them by using the two artifacts and some of them without using them. As a consequence two treatments were defined for Method factor. One expresses that the artifacts—System Contracts and System Sequence Diagrams)—were used to support the identification of the Domain Model (referred to as SSD/SC), and the other one to express that such artifacts were not used at all (referred to as No_SSD/SC).

Two dependent variables were defined: "Correctness" of the Domain Model and "Effort" invested in obtaining the Domain Model. The ways of measuring each of those variables are further discussed in Section 3.2.2.

3.2. Experiment planning

So far, we have provided the reasons for conducting this experiment. The next step is to describe how the experiment was conducted and this is what is presented in the next height subsections.

3.2.1 Context selection

The experiment was framed in the context of a fourth year course in Computer Systems Engineering and Software Engineering bachelors at Carleton University. All the subjects had been previously registered in full-term course on UML-based object-oriented software modeling. At the moment of the experiment all the subjects were registered for a course on software engineering, more specifically on UML-based software development. Prior to the experiment, the students were actually prepared with a set of lectures and laboratories and assignments where they had to deal with Domain Models, System Sequence Diagrams and System Contracts.

Two systems analysis documents were the objects used for this experiment: Video Store System (VS) and Car Part Dealer System (CPD). The first document describes the process of renting and selling video copies and the second one describes the process of selling car parts. Both systems have a proper level of complexity and they represent a domain problem with which the subjects are familiar with. For

each system was provided a verbal description of the problem domain, and a document containing a Use Case Diagram and use case descriptions.

3.2.2 Hypothesis formulation

Two hypotheses were evaluated in this experiment. The first one checked for significant difference between the Domain Model Correctness obtained by those subjects that used the artifacts and those that did not use the artifacts. To perform that evaluation, six dependent variables or correctness sub-features were defined: Number of missing classes (MC); Number of useless classes (UC); Number of missing relationships (MR); Number of wrong relationships (WR); Number of missing attributes (MA); Number of wrong attributes (WA). These six metrics were chosen since the classes, relationships and attributes are the elements used to properly describe Domain Models. By applying statistical test for each of those sub-features separately, one can have a first idea about the way independent variables are influencing the Domain Model Correctness. Because of the fact that two different software systems were used for the study, these sub-feature values were normalized before being used for the statistical tests. In this way, results associated with each system would be comparable to each other. Nevertheless, a more reliable result could be obtained by averaging the outcome of those six sub-features and applying the statistical tests to that result. In that regard a new variable was inserted in the model: Averaged Correctness (AC). That variable was computed as the sum of all sub-features values divided by the total number of sub-features (six).

Table 1 shows the hypotheses definition for those six sub-features describing Domain Model Correctness and the one that computes the average correctness value.

Dependent variable	Null hypothesis (H_0)	Alternative hypothesis (H_a)
Missing Classes (MC)	MC(SSD/SC) = MC(No_SSD/SC)	MC(SSD/SC) < MC(No_SSD/SC)
Useless Classes (UC)	UC(SSD/SC) = UC(No_SSD/SC)	UC(SSD/SC) < UC(No_SSD/SC)
Missing Relationships (MR)	MR(SSD/SC) = MR(No_SSD/SC)	MR(SSD/SC) < MR(No_SSD/SC)
Wrong Relationships (WR)	WR(SSD/SC) = WR(No_SSD/SC)	WR(SSD/SC) < WR(No_SSD/SC)
Missing Attributes (MA)	MA(SSD/SC) = MA(No_SSD/SC)	MA(SSD/SC) < MA(No_SSD/SC)
Wrong Attributes (WA)	WA(SSD/SC) = WA(No_SSD/SC)	WA(SSD/SC) < WA(No_SSD/SC)
Averaged Correctness (AC)	AC(SSD/SC) = AC(No_SSD/SC)	AC(SSD/SC) < AC(No_SSD/SC)

Table 1 Hypotheses definition associated to Domain Model Correctness

The second hypothesis checked for significant differences between the Effort in obtaining the Domain Model invested by those subjects who used the artifacts and those subjects who did not use the artifacts. In this case a single dependent variable was measured, the Time (T). That time was computed

by checking the time at which the students sent their Domain Model solution by email. Table 2 shows the hypotheses definition for Time.

Dependent variable	Null hypothesis	Alternative hypothesis
Time (T)	T(SSD/SC) = T(No_SSD/SC)	T(SSD/SC) < T(No_SSD/SC)

Table 2 Hypothesis definitions associated to Effort in obtaining Domain Model

3.2.3 Selection of subjects

Selection of subjects is a crucial issue when preparing an experiment that is supposed to be generalized to a desired population. This can be fulfilled by guaranteeing a random sample selection, a balanced data set and an adequate sample size.

As it has been already mentioned, subjects used for the experiments were students taken their last year in either Computer Science or Software Engineering. All of them were registered for the same Software Engineering course. At this level of study, students are supposed to have a strong background in UML-based object -oriented software development. This thesis is additionally supported by the fact that students have been already enrolled in courses related with object-oriented programming and UML-based modeling. That leads to think that any result reached with those experiments is applicable into an industry environment since those students constitute a representative sample of that vast mass of professionals starting working at an entry level.

3.2.4 Experiment design

As discussed in [7], a poor design may ruin a well-intended study. In that regard section 3.2.4.1 explains all the tasks that were performed as well as the time allocated for each one of them. Section 3.2.4.2 accounts for extraneous factors, which even though were not of crucial interest for the experiment, were taken into consideration because of their potential effects on the measures. Sections 3.2.4.3 and 3.2.4.4 consider two others issues, namely Learning/Fatigue Effect and Cross-Over Effect, whose presence could reduce credibility of any final result reached. Last, Section 3.2.4.5 describes the experiment design properly.

3.2.4.1 Experimental tasks and time allocation

The experiment was composed of four labs. The two first labs were used to get the subjects familiar with both toy-systems—VS and CPD—and to obtain their corresponding System Sequence Diagram

respectively. During the last two labs the main task for the subjects was to obtain the Domain Model corresponding to a given system description and define the System Operation Contracts. The experiment was performed in four weeks, so that once the experiment started; each lab was one week apart from the previous one.

3.2.4.2 Other factors to control

The purpose of the experiment was to measure the effect of changing the Method's treatments. Nevertheless, some other factors affecting the subjects were considered because they could have confounding effects with the Method effect. Those effects must be minimized as much as possible by blocking any known secondary factor as a way of increasing the experiment effectiveness [7].

For this experiment two extraneous factors were identified:

1. Ability: It was said that subjects received all the same training in UML and OCL notation. Nonetheless, subjects have varying talents to perform a task. This factor considered two levels, high and low. Subjects' classification into those two levels was given by their grades in previous Software Engineering courses.

2. System: Two systems with a similar level of complexity are used for the experiment (Table 3), but for some subjects it could be difficult to understand and capture all domain problems' specificities of a specific application. In addition, in other experiments where two apparently similar objects were used, a difference in subjects' performance when using them was revealed [10, 23]. So, it will be good to also consider this element as extraneous factors for the experiment. This factor had two levels, CPD and VS, which refer to Car Part Dealer system and Video Store system respectively.

A randomized block design that accounted for those two previous factors' levels was selected. This kind of design helps to reduce noise or variance in the data because the variability within each block is supposed to be less than the variability of the entire sample. As a consequence any estimate of the treatments effect within a block is more efficient than those estimates performed across the entire sample. In a randomized block design, subjects within each block are randomly assigned in equal proportion to a given Method treatment.

Aspect	System	
	CPD	VS
# of actors	5	7
# of use cases	12	7
# of classes	13	7
# of associations	16	10
# of attributes	32	24

Table 3 Systems complexity

3.2.4.3 Learning or Fatigue Effect

Learning Effect, also called Practice Effect, is one issue that researchers should take care of when carrying any experiment. This phenomenon has to do with the fact that a person learns more and more about a problem after dealing for a while with the same problem. As a result any significant change observed could be an effect of being tested a second time, instead of an effect of the treatment of interest [18, 39].

On the other hand, the Fatigue Effect is described as a decrease in subjects performance or effectiveness after doing the same activity for a while [39].

In this experiment subjects were asked to do the same task twice. In the first attempt, some of the subjects obtained the Domain Model by using the artifacts and the rest of the subjects obtained the Domain Model without using the artifacts. Later in the experiment, the treatments were crossed over between the two groups of subjects. In this way the Learning/Fatigue Effect can be avoided (or minimized) because the tasks to perform were not exactly the same. Another way of avoiding that Learning/Fatigue Effect was to provide in each attempt a completely different software system description to the subjects. It is good to remember once more that the experiment was implemented with two systems, CPD and VS. This last issue is better explained in Section 3.2.4.5.

3.2.4.4 Carry-over Effect

Carry-over Effect is another problem to pay attention when designing experiments. A Carry-over Effect would be present in any longitudinal study if the response of a measurement depends upon the treatment previously applied [18, 20].

In this case it is advised to have a "wash out" period in the experiment, so that any adaptation produced by the experimental training program disappears before subjects get the control program [18]. The uses

of two software system descriptions as well as the one-week separation between each task attempt were used as a way of decreasing the Carry-over effect in this experiment.

3.2.4.5 Experiment structure design

An explanation of how the experiment was structured is provided in the context of Experiment I. Experiments II, III and IV were very similar, but there were many more available subjects for experiments II and IV. Section 3.2.8 is more explicit about what was different in the other experiments.

It is worth remembering that four labs were designed for the whole experiment. The two first labs were supportive tasks to help subjects to understand both the software system descriptions under consideration (lab 1) as well as to start obtaining the design artifacts whose influence in the Domain Model was to be studied (lab 2). The two last labs were the core of the experiment, where the Domain Models for a given system were obtained. At the end of this section a table with detailed information about each lab is provided (Table 6).

The experiment was initially designed by considering a single factor: Method. This factor was conceived with two levels or treatments: SSD/SC and No_SSD/SC. As a consequence the experiment was designed with two groups of subjects. Because subjects were required to obtain the Domain Model twice, a Cross-Over design was applied with respect to both Method factor's treatments.

In this regard two groups of subjects were created, GroupA and GroupB. In lab 3, GroupA obtained the Domain Model with the support of the artifacts and GroupB obtained the Domain Model but without using any of the artifacts. In lab 4, the two treatments were crossed over between the two groups (Table 4).

Lab	**Task**	**Groups**	
		GroupA	**GroupB**
Lab3	Domain Model	SSD/SC	No_SSD/SC
Lab4	Domain Model	No_SSD/SC	SSD/SC

Table 4 Design by considering the Method factor's treatments

Because of the fact that a longitudinal study was considered, where the Domain Model was obtained twice, it made no sense to obtain the Domain Model for the same software system twice. That is why each group was further subdivided into two groups, so that in each attempt of obtaining the Domain Model, each level of Method factor interacted with both Systems under consideration (Table 5).

Lab	Task	Groups			
		GroupA		GroupB	
		Group1	Group2	Group3	Group4
Lab3	Domain Model	CPD (SSD/SC)	VS (SSD/SC)	CPD (No_SSD/SC)	VS (No_SSD/SC)
Lab4	Domain Model	VS (No_SSD/SC)	CPD (No_SSD/SC)	VS (SSD/SC)	CPD (SSD/SC)

Table 5 Design by considering interactions of between levels of Method factor and System factor

This design allowed carrying out different statistical tests of interest. It was possible to consider tests such as Two-Way ANOVA, Three-Way ANOVA and Mixed Repeated Measures ANOVA, which permit studying the real effect of certain confounding factors, if any. It also allows performing Simple Repeated Measures ANOVA tests that permit considering data of both attempts of obtaining the Domain Model when analyzing if using the artifacts had a significant positive effect (the more data available for any study, the more reliable the results will be). This last statistical test, also allows verifying the influence of the order in which Method's treatments were applied (Carry-over effect) and for significant differences between each attempt of obtaining the Domain Model (Practice/Fatigue effect). More explicit explanations about these statistical tests are available in Section 3.2.6.

Given the above information, it is possible to describe the entire experiment design in a table (Table 6). There was a first lab where all the subjects (the four groups) were supposed to get familiar with both software system descriptions. Two groups of subjects were considered for the second lab. In that lab half of the subjects obtained the System Sequence Diagram for CPD system and the other half for VS system.

Lab	Task	Groups			
		GroupA		GroupB	
		Group1	Group2	Group3	Group4
Lab1	Software systems comprehension	VS and CPD systems			
Lab2	System Sequence Diagram	CPD		VS	
Lab3	Domain Model and System Contracts	CPD (SSD/SC)	VS (No_SSD/SC)	CPD (SSD/SC)	VS (No_SSD/SC)
Lab4	Domain Model and System Contracts	VS (No_SSD/SC)	CPD (SSD/SC)	VS (SSD/SC)	CPD (No_SSD/SC)

Table 6 Experiment design

In lab three, half of the subjects that worked with CPD system at lab two were further subdivided in equal parts. One half was asked to obtain the Domain Model and System Contracts for CPD system

(they were also given the System Sequence Diagram obtained at lab two). The other half of subjects was asked to just obtain the Domain Model for VS systems, that is, for that software system description for which they did not deal with at lab two. A similar procedure was followed for the other half of subjects that worked with VS system at lab two. That is, the group was further subdivided so that half of the subjects obtained the Domain Model for VS system, the system description that they used at lab two, and the other half obtained the Domain Model for CPD system, a system description that they did not use in lab 2.

Lab 4 was essentially the same as lab 3, but with a cross over of systems descriptions and Method treatments between subgroups. For example, those subjects that at lab 3 obtained the Domain Model for CPD systems by using the artifacts (Group1), were asked at lab 4 to obtain the Domain Model for VS system (a different system), for which the did not use the corresponding artifacts (different treatment). On the other hand, the other half of subject that at lab 3 obtained the Domain Model for VS system without using any artifact (Group2), had to obtain at lab 4 the Domain Model for CPD system with the support of the artifacts.

3.2.5 Instrumentation

Whenever an experiment is planned, researchers have to be very careful about choosing and having available the proper objects to be used by subjects and about determining the measurement mechanism [7]. In this experiment, by objects are interpreted any document or tool required to carry out this Software Engineering study.

Labs were performed in a computer laboratory where each subject had a computer to accomplish the requested tasks. At each lab, subjects had on their computers all the materials that they will need to use. Those materials are described next.

Subjects had to fill in questionnaires at the end of each lab as a way of measuring their perception about the task performed. Those questionnaires were also useful to support some of the conclusions associated to the statistical hypotheses. They were designed by following guidelines provided in [22] to avoid bias and optimize reliability. For lab 3 and 4 there were some differences in the questionnaires we gave to subjects, depending on whether the subject had to deal with the artifacts or not. Those questionnaires were essentially answered by using Likert scales, which measure the degree of

agreement or disagreement of subjects with each statement on the questionnaires. See Appendix F from Figure 34 to Figure 39 for the detailed questionnaires.

At each lab, subjects were also provided with three documents according to the software system that they were supposed to work with:

1. A verbal description of the software system
2. Use Case Diagram
3. Use Case Description

In lab 1, subjects were asked to carefully read each system and answer a comprehension questionnaire that helped us to measure their level of understanding of each software system description. The questionnaires were also an incentive for them to carefully understand the systems.

For labs 2, 3 and 4 students had to download templates from the course web site to accomplish the tasks requested. In those templates, subjects were asked to provide some specific information with a particular format (Appendix F, Figure 40 and Figure 41).

From Lab 2 to Lab 4 subjects were also provided with materials corresponding to another Case Study (Cab Dispatching system), a system used for dispatching cabs to customers. The subjects were also familiar with this system. It was used as an example of the kind of results we were expecting. For example, during Lab 3, during which subjects had to build System Sequence Diagrams, all subjects were provided with a description of the Cab Dispatching system, with the corresponding Use Case Diagram and System Sequence Diagrams.

Three others tools were used during the experiment. JMP and SAS were two statistical packages used to analyze the collected information. Both tools were created by SAS Institute Inc.. Visio 2003 was the third tool, which was used by subjects to easily draw the requested diagrams (Domain Model and System Sequence Diagrams).

It was already explained in section 3.2.2 that the Domain Model Correctness was measured through six sub-features. Reference models developed by the experimenters (one for each software system) were used to equally evaluate each subject solution, based on a unique solution. The following guidelines were considered when evaluating each Domain Model:

- Only classes were accounted for, not interfaces.
- Associations of type Aggregation and Composition were indistinctly accepted
- Association Classes can be substituted by relations involving three classes and vice versa [31, 32].
- When dealing with an inheritance structure, if the superclass is missing in the subjects' design, but the attributes of such a superclass are defined in the corresponding children classes, then those attributes do not count as missing or wrongly-defined attributes.
- In case that there is a missing multiplicity, it is assumed that it has a value of one [3, 4].
- If one or both multiplicities of an association are wrongly specified, then that would account as only one wrong association.
- All the relationships a missing class is involved in, in the correct model, are considered as missing relationships.
- Associations involving classes that are identified as not necessary (useless) do not account as wrong associations.
- Attributes defined for useless classes do not count as wrongly defined attributes.
- Attributes that were defined in the wrong class account as wrongly-defined attributes.
- In general, a mistake is only accounted for once.

The students' models were not expected to necessarily closely resemble the reference model. For example, we expected the class and attribute names to differ. As a result a careful comparison had to be performed to differentiate real mistakes from differences of no consequence.

Before analyzing the score reached by each subject for those six metrics, their values were conveniently normalized according to each system parameter. That is, the right number of classes, attributes and relationships between classes that were supposed to be defined in each Domain Model were used for that normalization. This way, the outcome will be much more appropriate for any statistical analysis. Additionally, an average value was computed with those normalized data, which was finally used to carry out the statistical study.

To measure Invested-Effort, it was already mentioned that it was measured the time invested by the subjects to accomplish their tasks. That time will not include the time invested by subjects to answer

the questionnaires. That is why those questionnaires will be answered at the end of the corresponding lab.

3.2.6 Data analysis procedure

The analysis of the date, as already mentioned, considered the two dependent variables defined for the experiment (Correctness of the Class Diagram and Time invested in lab tasks) as well as a single independent variable (Method), with two separate treatments (SSD/SC and No_SSD/SC).

The data analysis was divided in three parts. The first one involved descriptive statistical tests on the data. The second one was further subdivided in two parts: Univariate Analysis and Multivariate Analysis of the data, both of them are used to test the equality of means. The third part consisted of Univariate Analysis on questionnaires answered by the subjects. For those Univariate and Multivariate analyses, a level of significance (α) of 0.05 was used for each hypothesis test. As a consequence, a p-value less than 0.05 meant that the alternative hypothesis could be accepted. It is worth also mentioning that all Univariate and Multivariate statistic tests were run with the SAS package.

3.2.6.1 Descriptive statistical analysis

Before going further with much more complex statistic tests, an investigator is supposed to perform descriptive statistic. In that way it is possible to summarize and easily describe some interesting issues embedded in the data. This statistics test was applied in all four experiments for all data collected for each one of the dependent variables in each task.

3.2.6.2 Univariate analysis

A One-Sample *t*-test was used to analyze data regarding lab 1. This statistical test is used to compare the mean of a sample to a known value. The goal of this test was to measure the subjects' level of understanding of each system that they were supposed to deal with during the following 3 labs.

A One-Sided Two-Sample *t*-test was another statistical test used. It was used for each task related to obtaining a Domain Model by considering the dependent variables. The goal was to identify whether the two samples' means, which refer to those subjects who used the artifacts (SSD/SC) and those who did not use them (No_ SSD/SC), were equal or not. In the case that they were not equal, we then checked if that inequality was really significant. A p-value less than 0.05 indicated, according to the

hypothesis formulated, that subjects using the artifact performed significantly much better than those that did not use the artifacts.

That test was not definitive to reach a conclusion about a significant effect of using the artifact to obtain the Domain Model. The reason is that this test does not consider all the observations of both attempts at once—remember that the more observations are available, the more reliable the results are. There exist other tests that can account for all the available data of both experiments attempts, by removing individual differences. For example, Simple Repeated Measures ANOVA test (see below). It was decided to use a One-Sided Two-Sample *t*-test anyway to have another insight into the data behaviour in each lab separately. Actually, this approach helped to support some of the final conclusions (see section 4.1). Besides, results of this statistical test could have been really useful in case that a Carry-over or Practice Effect could have been detected as a result of carrying out a longitudinal study.

A one-sided approach was always considered because it was expected that those subjects using artifacts had a better performance than those not using them.

As mentioned above, Simple Repeated Measures ANOVA, also called Single Repeated Measures ANOVA, was another Univariate analysis used to statistically analyze the data observations. With this statistical test, all the observations from lab 3 and 4 were used together because the test accounts for standard deviations and correlations between observations [37, 38]. That is something that the regular ANOVA test does not permit. One could think that a Paired *t*-test would be feasible for this experiment because it was a longitudinal study where the same subjects are measured twice under different conditions. But it can not be applied in this case because the design of this experiment considered a control group which is an aspect that is not modeled by paired *t*-test. Repeated Measures ANOVA designs account for subjects organized in control group and experimental group. In this case, both groups perform the same activities, but the first one is not receiving a treatment at all. Control groups are necessary as a way of separating the effect of the experimental stimulus from the effect of the test itself. For example, in this experiment, the subjects could have obtained a much better Domain Model simply because they just have learned how to do the task and not because they responded to the treatment at all. The aim of this test was to check if there was a significant difference between the subjects that used artifacts to obtain a Domain Model and those who did not use such artifacts. In addition, this kind of analysis also allowed checking Method Order and Learning/Practice Effect.

In each experiment, during the first attempt of obtaining the Domain Model (lab 3), half of the subjects designed the Domain Model for a system description for which they previously obtained its System Sequence Diagrams and System Operations Contracts (SSD/SC). The other half did similarly, but for a system description for which they did not obtain the corresponding artifacts (No_SSD/SC). During the second attempt (lab 4), subjects obtained a new Domain Model, but this time Method's treatments (SSD/SC and No_SSD/SC) were applied in reverse order. With a Simple Repeated Measures ANOVA test, it was possible to compare the subjects' performance when obtaining the Domain Model and to take into consideration at the same time two other factors: Method Order and Time factor.

The Method Order factor effect is also called Cross-Over Effect. This factor has to be introduced into the model and studied because it may negatively impact the final results if it is present. This effect appears when a new treatment is applied (second attempt at the task in our experiment) before the effect of a previous applied treatment has disappeared (first attempt at the task). Additionally, it is expected that the subjects' performance get better from one task to another because they are supposed to gain in experience. Conversely, that performance could get worse if subjects do not feel really motivated with the task to perform (especially the second time). Both situations have also a negative effect on results. This is commonly called Learning/Practice Effect and Boredom Effect, respectively, and these effects can be also taken into account by using Repeated Measures ANOVA. In that regard a "Time" factor was introduced into the model with level values of 1 and 2. The goal was to determine if subjects' response improved or worsen from lab 3 to lab 4 (where they obtained the Domain Model).

3.2.6.3 Multivariate analysis

Two approaches were undertaken using multivariate analysis. N-Factorial Way ANOVA test, also referenced to as regular ANOVA, was used to analyze independently data collected from labs 3 and 4. The second approach was Mixed Repeated Measures ANOVA, which was used for the combined analysis of data from labs 3 and 4.

Applying Two-Sample Independent *t*-test could be considered as a sufficient test to analyze whether there exist significant differences in results between the two Method treatments. Nevertheless, in the experiment there were some other elements involved that could have strongly affecting the results (i.e., the additional factors we already mentioned). That is why it was necessary to use a test that could help to isolate those confounded factors and provide a much more reliable conclusion. Two/Three Way

ANOVA is one such test, which in addition allowed studying for possible interactions between the factor of interest (Method) and those confounded effects. It is said that factors interact when the effect of changing the levels of one factor depends on the particular level of some other factor [13, 27].

Two extraneous factors were identified as part of the experiment. Subjects' ability and the system used to obtain the Domain Model were those two confounded factors for this experiment. The goal of considering those two new factors was to improve the data analysis by studying the effect of those two variables and their interactions with the main independent variable of the experiment (Method).

As a result, three factors were taken into account when trying to apply N-Factorial Way ANOVA test: Method, Ability and System. One can infer that Three-Way ANOVA was the proper statistic analysis to apply. However, for Experiments I and III, the small amount of data available did not allow for applying Three-Way ANOVA because that test could lead to erroneous results. That is why results of Experiments I and III were analyzed by using Two-Way ANOVA where Method & Ability and Method & System factors combinations were considered. For Experiments II and IV it was possible to apply a Three-Way ANOVA because the sample size was large enough.

Recall that regular ANOVA tests were implemented for each individual lab where Domain Models were obtained. This way, the most significant assumption imposed by ANOVA about independence between observations was not violated. Those tests were used as an alternative, in case a Practice effect or Carry-Over effect were detected when analyzing each dependent variable. If this were to happen, we would have only considered the measures associated to the first attempt at the task (lab 3).

The second part of the analysis considered the combined data from labs 3 and 4. Mixed Repeated Measures ANOVA test was applied to study the influence of confounded factors previously mentioned, but this time by considering data observations of both attempts of obtaining the Domain Model. In previous section (3.2.6.2), as well as section 2.3, we mentioned the advantages of using Repeated Measures ANOVA tests in general and Mixed Repeated Measures ANOVA tests.

The data were analyzed for both Simple Repeated Measures ANOVA and Mixed Repeated Measures ANOVA, by using Mixed procedure in SAS package. That procedure makes it possible to analyze repeated measures data efficiently by first modeling the variance and correlation structure of the repeated measures as it is suggested in [20, 30].

3.2.6.4 Analysis of survey data

We used survey data to support the results computed by the different statistical tests mentioned in sections 3.2.6.2 and 3.2.6.3. A One-Sample *t*-test was implemented on the Comprehension questionnaire performed during the first lab. This test measured the subjects' level of understanding of each system description considered for the experiment.

One-sided, two-sample *t*-tests were used on the other questionnaires, for each lab independently.

3.2.7 Threat of validity

One very important aspect to consider while designing an experiment has to do with accounting for the right validity of the results that will be obtained with that experiment. One has to ensure that all factors affecting the experiment are taken into consideration, that all means of measurements are accurate enough, that those means are actually measuring what they really intended to measure and that any result can be extrapolated to the whole population of interest.

3.2.7.1 Internal Validity:

This threat has to do with the fact that in the experiment there exist a real cause-effect relationship between the independent and dependent variables by considering some extraneous variables affecting the dependent variables. A controlled setting like the one used for this experiment is efficient for controlling those extraneous effects. For example, for this experiment it is expected that the Correctness factor increases and Invested-Effort factor decreases when using SSD/SC to obtain the Domain Model. We discuss below the characteristics of the experiment that ensure internal validity:

- Tasks were assigned randomly to subjects and at the same time all possible confounding effects such as subject's ability, and system used were considered.
- Given that a longitudinal study was implemented, some event could have influenced subjects during the experiment. There was specially a concern about a possible exchange of information between subjects. That was faced in different ways:
    - Subjects could not talk to each other while they were performing a task. Any doubt would have to be discussed with someone in charge of the lab and we ensured that any explanation would not provide too many details about the expected result of the task.

- Subjects were not allowed to keep any material handed in during the labs.
- Student never had an idea about the experiment design, that is, about what aspects were intended to be studied or about what task to carry out next time in the lab.

- Subjects applied consistently any technique involved with obtaining the Domain Model, the System Sequence Diagram or the System Operations Contracts. That is why they took courses were they were taught how to develop those artifacts. They were additionally provided with a toy-project (Cab Dispatching system) with similar documentation as the one that they were provided and the expected kind of result.

- Fatigue and learning effect was another issue of concern. They were avoided using different means:
  - The fact that subjects are immersed in some others activities, that is, taking some other courses. That helps with the learning effect concern.
  - Labs were one week apart. Consequently it is hard to believe that they will get bored or tired for doing the same activity twice. Besides, there were some differences between activities to carry out at each lab when obtaining the domain Model, that is, they did not have to do exactly the same activity in both attempts of the experiment.
  - We assigned different software systems to subjects at each lab to avoid both effects.

- The methodology or instruments used to measure the dependent variable are reliable. Some metrics were defined to properly measure the Domain Models Correctness (see section 3.2.2). Those metrics were computed for each subject's Domain Model through a reference model that exists for each system. That guaranteed that all subjects were equally measured. In that way, the whole evaluation process was repeatable and easy to carry out, which is a desirable feature for any experiment [7].

- The number of subjects taking part in the experiment was the same during the whole experiment. This is hard to control because a subject can drop out the course or just be absent during a lab because she/he got sick. Actually, that happened a couple of times but it did not affect the experiment progress at all (especially the balancing of groups). Repeated Measures ANOVA is the only test that could have been affected with such a situation, but it did not happen because subjects

in that situation never missed both attempts of obtaining the Domain Model. It was pointed out that this statistical test can deal with missing data.

- A last aspect to be considered is related with the Invested-Effort factor. As it was stated, this factor measured how long a subject takes to perform her/his task. As a consequence it will be important to avoid as much as possible the number of interruptions during the experiment as well as subjects' break times. It has to be mentioned that each lab was three hours long.

3.2.7.2 External Validity

This threat to validity has to do with the real possibility of generalizing any result that is obtained during the experiment. Two elements worth mentioning are:

- As it was mentioned, in this experiment the subjects were fourth-year students of Computer System Engineering and Software Engineering bachelors. The software industry is normally composed of junior professionals with not too much practice in real work environments, especially because of the high demand of those kinds of professionals nowadays. That is why it is thought to be proper to generalize the result with respect to this issue.

- For this experiment, two toy-systems were selected whose complexity is not too high so that tasks can be performed within a laboratory environment (three hours laboratories). In an industrial environment, much more complex software systems can be usually found, but the real question to be answered is whether it will be equally valuable or useless to use Systems Sequence Diagrams and Systems Operations Contracts to obtain Domain Models for those much more complex systems. Since those systems are more complex, with a large number of classes, attributes and class relationships, it is thought that there could be a great possibility for those two artifacts to help to device new elements in the Domain Model as well as to refine it.

3.2.7.3 Construct validity

This threat to validity is about the connection that has to exist between the concepts that are being studied and the measurements that are being used. For this experiment there is no doubt that the time taken by each subject to build the Domain Model is a way to measure the invested Effort factor. Six metrics were defined to objectively measure the Correctness of the Domain Model (see section 3.2.2).

Those six metrics guaranteed to cover all elements of the Domain Model and how well-defined they were.

3.2.7.4 Conclusion validity

This threat has to do with considering all possible issues that can lead toward a wrong conclusion about the relationship between the independent variable and the dependent variables. Addressing this issue could be achieved as follows:

- By performing the proper statistical tests that accurately assess the relationships between those variables under study.
- By conducting a power analysis before executing each experiment to make sure that there was a large enough data set to have a high probability of finding a statistically significant difference when there was one. We consistently obtained a power above 0.8.
- By using an appropriate level of significance (0.05).
- By guaranteeing on each experiment that subjects will be under the exact same conditions and remained undisturbed from any extraneous factor. It was a controlled experiment, carried out in a laboratory setting where the subjects had all the resources needed to perform the tasks.
- Model correctness measures were precisely defined and justified. Questionnaires were designed according to state-of-the-art guidelines.
- The experiment was replicated four times as a way of assessing the repeatability of the findings.

3.2.8 Others experiment trials

The experiment was repeated during the Fall/2005 (Experiment II), the Summer/2005 (Experiment III) and the Fall/2006 (Experiment IV) terms. All those experiments were carried out under almost the same conditions as for the first experiment. We describe the changes in this section.

First, a new variable was introduced to measure the time invested to obtain the Domain Model: "Time obtaining Domain Model". "Time in lab" variable was the one initially considered for Experiment I. This variable did not provide a reliable measure of the time invested by students to obtain de Domain Model because during the lab the students had to deal with some other task than just obtain the Domain

Model. But, it was decided to keep the "Time in Lab" independent variable as part of the experiment design as a way of checking if similar results were obtained when analyzing this variable. If so, then one could extrapolate any result obtained for "Time obtaining Domain Model" variable in Experiment II, III and IV to Experiment I

Second, for Experiments II and IV there was twice the number of subjects. That change led to some others changes to the experiment design. It was not possible to have all the subjects at the same time taking a given task and as a consequence they were divided in two groups. With this new distribution, each group was attending the labs every two weeks and as a result the experiment had to be extended to eight weeks. The larger number of subjects allowed the use of Three Way ANOVA analysis and Mixed repeated Measures ANOVA analysis by considering a three variable (Method, Ability and System) all together.

Another change had to do with adding some new questions to questionnaires in those labs where the Domain Model was obtained: this applies to Experiments II, III, and IV. The goal of those questions was to have a better idea of how much time the subjects invested in obtaining the Domain Model.

In all replications the students were encouraged to do their best during the labs by suggesting that lab evaluation would be a consensus of attendance to the labs and amount of work performed during the labs.

A last change only involved Experiment IV. For this experiment, the students were more exposed to the concepts and ways of dealing with the software engineering artifacts that they needed to use during the labs experiment. The use of those artifacts was evaluated through an assignment, a quiz and a midterm exam.

4 Experiment results and analysis

This section presents the results of applying the statistical tests announced in the previous section as well as the discussion of those results.

Results are presented for all experiments together, grouped by statistical test. Those results presentation followed the same structure explained in Section 3.2.6: First, the descriptive statistic (section 4.1), second the univariate (section 4.2) and multivariate (section 4.3) analysis, and finally the analysis of the questionnaires (section 4.4).

4.1. Descriptive Statistic

As already mentioned, the descriptive statistics were useful to create a preliminary impression. In that regard the minimum, maximum, mean and standard deviation values were computed. Descriptive statistics for the four experiments are presented next in sections 4.1.1, 4.1.2, 4.1.3, and 4.1.4, respectively.

4.1.1 Experiment I (Summer 2005)

Table 7 shows the Descriptive Statistic results for Experiment I. For Lab-3, a considerable difference is seen in one of the sub-features related to the Domain Model Correctness factor, specifically "Useless Classes" (0.46 for SSD/SC-subjects and 0.25 for No_SSD/SC-subjects). It seems that subjects who used the artifacts defined on average much more useless classes than subjects who did not use such artifacts. In addition, for the "Missing Relationships" sub-feature, there exists a difference, not as large as the previous one. A quantitative analysis of this difference is tested and explained in the following sections. It should be also noticed that, opposite to what was expected, the mean values computed for those subjects not using the artifacts are slightly smaller that those for subjects using the artifacts. The mean time invested for each group to obtain the Domain Model was essentially the same (176.9 min for SSD/SC-subjects and 174.5 min for No_SSD/SC-subjects).

No big differences seem to exist between the two groups of subjects under study when analyzing Lab-4, specifically when analyzing the sub-features related to the Domain Model Correctness factor. The mean values computed for subjects who used the artifacts and those who did not use them are very similar. Again, "Useless Classes" is the only one important difference between both groups (0.28 for

SSD/SC-subjects and 0.41 for No_SSD/SC-subjects). The direction is however the opposite as what was observed in Lab 3: the subjects who used the artifacts created much fewer "Useless Classes" than those who did not use the artifacts. The average times for both groups are noticeably different in Lab 4 (173.3 min for SSD/SC-subjects and 153 min for No_SSD/SC-subjects). This time, the subjects who did not use the artifacts were faster than those who used the artifacts. Notice that the fastest subjects in Lab 4 were the ones who used the artifacts in Lab 3 (recall that they used two different systems in labs 3 and 4). Whether this means that there was a Learning Effect (with the task at hand) will be answered in section 4.2.3.1

In summary, descriptive statistic results over the two attempts do not show a remarkable difference between both groups of subjects (in terms of correctness and time). One counter-intuitive result occurred in Lab-3: subjects not using the artifacts produced a better Domain Model.

			Lab 3					Lab 4				
Dependent variable		**Group**	**Obs.**	**Min**	**Max**	**Mean**	**Std. Dev.**	**Obs.**	**Min**	**Max**	**Mean**	**Std. Dev.**
Correctness	Missing classes	SSD/SC	16	0.15	1.00	0.37	0.18	16	0.14	0.69	0.39	0.14
		No_SSD/SC	16	0.23	0.53	0.35	0.10	16	0.14	0.53	0.37	0.11
	Useless classes	SSD/SC	16	0.23	1.14	0.46	0.28	16	0.00	0.85	0.28	0.23
		No_SSD/SC	16	0.00	0.57	0.25	0.15	16	0.00	0.85	0.41	0.26
	Missing relationships	SSD/SC	16	0.33	1.00	0.62	0.17	16	0.33	0.81	0.59	0.14
		No_SSD/SC	16	0.33	0.77	0.56	0.16	16	0.31	1.00	0.59	0.20
	Wrong relationships	SSD/SC	16	0.00	0.25	0.11	0.07	16	0.00	0.22	0.08	0.06
		No_SSD/SC	16	0.00	0.27	0.08	0.08	16	0.00	0.27	0.11	0.06
	Missing attributes	SSD/SC	16	0.37	1.00	0.62	0.16	16	0.40	0.86	0.58	0.14
		No_SSD/SC	16	0.35	0.86	0.61	0.15	16	0.35	0.89	0.55	0.17
	Wrong attributes	SSD/SC	16	0.00	0.55	0.18	0.17	16	0.00	0.75	0.22	0.16
		No_SSD/SC	16	0.03	0.44	0.16	0.13	16	0.06	0.45	0.25	0.11
	Averaged correctness	SSD/SC	16	0.26	0.56	0.37	0.08	16	0.23	0.51	0.36	0.06
		No_SSD/SC	16	0.19	0.47	0.34	0.07	16	0.22	0.54	0.38	0.09
Time in lab		SSD/SC	16	156.0	186.0	176.9	6.98	16	146.0	188.0	173.3	11.51
		No_SSD/SC	16	156.0	187.0	174.5	9.73	16	131.0	176.0	153.0	14.8

Table 7 Summary of descriptive statistic by considering Method's levels (Experiment I)

4.1.2 Experiment II (Fall 2005)

Table 8 shows the descriptive statistic results for Experiment II. Notice that two different times are computed in Table 8. It is worth remembering that from Experiment II on two time variables were

measured: "Time obtaining the Domain Model" and "Time in lab". Recall from section 3.2.8 that, it was realized that the last variable was not really useful to check whether the use of the artifacts contribute to reduce the time in obtaining the Domain Model. But, it was decided to keep it as part of the experiment design as a way of checking if similar results were obtained when analyzing this variable. If so, then one could generalize any result obtained for "Time obtaining Domain Model" variable in Experiment II, III and IV to Experiment I. For lab 3 all the Correctness sub-features results were slightly different, always favoring those subjects who used the artifacts. The only difference that seems to be significant is the one for the "Useless Classes" sub-feature (0.20 for SSD/SC-subjects and 0.29 for No_SSD/SC-subjects). Additionally, one worrying result that is important to point out is the high score (close or equal to 1) reached for those sub-features measuring missing elements in the Domain Model ("Missing Classes", "Missing Relationships" and "Missing Attributes"), indicating that some students did not identify a single correct class. For "Time obtaining Domain Model", subjects who used the artifacts seem to require less time than the others (108.9 min for SSD/SC-subjects and 120.7 min for No_SSD/SC-subjects). For "Time in lab", the results seem to favor to those subjects not using the artifacts (159.3 min for SSD/SC-subjects and 153.9 min for No_SSD/SC-subjects). However, remember that subjects using the artifacts had to also produce contracts in addition to the domain model.

In Lab-4, the trend for the Domain Model correctness sub-features changed. The magnitudes were very similar to the ones computed for Lab-3, but they surprisingly favored many times those subjects not dealing with the artifacts. A Learning Effect could explain this situation. The "Time obtaining Domain Model" variable favored those subjects not using the artifacts, but this time the difference between the two groups of subjects was less marked than in Lab 3(104.1 min for SSD/SC-subjects and 106.5 min for No_SSD/SC-subjects). This again could be explained by a Learning Effect. This will be investigated in section 4.2.3.2. Overall, the descriptive statistic results for Experiment II are very similar to the ones obtained for Experiment I. Similar to Experiment I, the ranges of values and the variations lead us to doubt that differences are statistically significant. This will be answered later.

Dependent variable		Group	Lab 3 Obs.	Min	Max	Mean	Std. Dev.	Lab 4 Obs.	Min	Max	Mean	Std. Dev.
Correctness	Missing classes	SSD/SC	32	0.15	1.00	0.45	0.18	30	0.14	0.76	0.46	0.13
		No_SSD/SC	30	0.14	0.92	0.45	0.14	32	0.00	0.71	0.45	0.18
	Useless classes	SSD/SC	32	0.00	0.71	0.20	0.21	30	0.00	0.61	0.22	0.17
		No_SSD/SC	30	0.00	0.85	0.29	0.21	32	0.00	1.00	0.22	0.25
	Missing relationships	SSD/SC	32	0.50	1.00	0.71	0.13	30	0.50	1.00	0.76	0.13
		No_SSD/SC	30	0.43	1.00	0.72	0.14	32	0.37	1.00	0.71	0.18
	Wrong relationships	SSD/SC	32	0.00	0.50	0.17	0.14	30	0.00	0.31	0.19	0.07
		No_SSD/SC	30	0.00	0.50	0.19	0.10	32	0.00	0.50	0.20	0.11
	Missing attributes	SSD/SC	32	0.37	1.00	0.70	0.15	30	0.41	0.87	0.69	0.11
		No_SSD/SC	30	0.56	0.87	0.72	0.09	32	0.41	0.96	0.68	0.12
	Wrong attributes	SSD/SC	32	0.00	0.58	0.17	0.13	30	0.03	0.54	0.17	0.10
		No_SSD/SC	30	0.00	0.40	0.17	0.10	32	0.00	0.71	0.19	0.16
	Averaged correctness	SSD/SC	32	0.28	0.61	0.40	0.07	30	0.25	0.50	0.41	0.05
		No_SSD/SC	30	0.32	0.56	0.42	0.06	32	0.20	0.56	0.41	0.08
Time obtaining Domain Model		SSD/SC	32	27.5	159.3	108.9	31.6	30	35.5	173.7	104.1	33.9
		No_SSD/SC	30	51.0	170.1	120.7	30.6	32	47.5	158.4	106.5	38.1
Time in lab		SSD/SC	32	89.0	183.0	159.3	24.6	30	104.0	193.0	154.5	23.1
		No_SSD/SC	30	102.0	189.0	153.9	24.2	32	76.0	176.0	136.6	31.4

Table 8 Summary of descriptive statistic by considering Method's levels (Experiment II)

4.1.3 Experiment III (Summer 2006)

Table 9 shows some descriptive statistic for Experiment III. When looking at Lab-3 results, one can realize that some of the sub-features describing Domain Model correctness favor the subjects using the artifacts, while others favor subjects no using the artifacts. Despite those variations, the general Domain Model correctness ("Average Correctness") favored those subjects using the artifacts (0.35 for SSD/SC-subjects and 0.38 for No_SSD/SC-subjects). The difference seems to be no large enough to find a significant difference associated to "Averaged Correctness" dependent variable. For "Time obtaining Domain Model" the panorama is very similar. The mean value favors subjects using the artifacts, but the difference between the two groups (SSD/SC and No_SSD/SC) does not seem to be significant enough (105.7 min for SSD/SC-subjects and 108.9 min for No_SSD/SC-subjects). "Time in Lab" dependent variable follows the same trend so far described: subjects using the artifacts spent more time to solve the lab tasks (171.4 min for SSD/SC-subjects and 149.3 min for No_SSD/SC-subjects).

			Lab 3					Lab 4				
Dependent variable		**Group**	**Obs.**	**Min**	**Max**	**Mean**	**Std. Dev.**	**Obs.**	**Min**	**Max**	**Mean**	**Std. Dev.**
Correctness	Missing classes	SSD/SC	10	0.14	0.46	0.32	0.11	11	0.14	0.61	0.38	0.13
		No_SSD/SC	11	0.28	0.57	0.38	0.10	10	0.28	0.53	0.40	0.09
	Useless classes	SSD/SC	10	0	0.57	0.21	0.17	11	0.07	0.71	0.39	0.23
		No_SSD/SC	11	0	1.14	0.30	0.32	10	0	0.71	0.34	0.26
	Missing relationships	SSD/SC	10	0.40	0.75	0.58	0.14	11	0.50	0.90	0.63	0.12
		No_SSD/SC	11	0.40	0.90	0.57	0.13	10	0.50	0.75	0.62	0.10
	Wrong relationships	SSD/SC	10	0	0.37	0.18	0.12	11	0.06	0.50	0.19	0.12
		No_SSD/SC	11	0.12	0.50	0.28	0.13	10	0.06	0.40	0.16	0.11
	Missing attributes	SSD/SC	10	0.33	0.87	0.61	0.16	11	0.50	0.70	0.61	0.06
		No_SSD/SC	11	0.37	0.71	0.57	0.10	10	0.62	0.93	0.70	0.10
	Wrong attributes	SSD/SC	10	0.06	0.41	0.18	0.11	11	0	0.29	0.16	0.08
		No_SSD/SC	11	0.03	0.54	0.17	0.13	10	0	0.34	0.20	0.11
	Averaged correctness	SSD/SC	10	0.27	0.44	0.35	0.05	11	0.28	0.48	0.39	0.06
		No_SSD/SC	11	0.31	0.51	0.38	0.06	10	0.31	0.49	0.40	0.06
Time obtaining Domain Model		SSD/SC	10	62.5	141.7	105.7	29.6	11	43.7	137.2	101.4	28.6
		No_SSD/SC	11	68.5	146.2	108.9	23.3	10	46.0	129.6	89.0	25.3
Time in lab		SSD/SC	10	125.0	190.0	171.4	21.4	11	103.0	183.0	165.2	25.0
		No_SSD/SC	11	113.0	195.0	149.3	21.3	10	91.0	144.0	121.4	21.0

Table 9 Summary of descriptive statistic by considering Method's levels (Experiment III)

The situation for Lab-4, at first glance, seems to be very similar to what happened at Lab-3. For Domain Model correctness sub-features values, we sometimes observe a better score for subjects using the artifacts, and other times for subjects not using the artifacts. In addition, the variations of those correctness sub-features did not follow the same pattern seen for Lab-3. As a consequence, fluctuations of values are once more present (it also happened at Lab-3, but this time are less marked). As a whole, the Domain Model Correctness hardly favored subjects using the artifacts (0.39 for SSD/SC-subjects and 0.40 for No_SSD/SC-subjects).

It is interesting to note that mean values related with Correctness sub-features increased in most of the cases from lab 3 to lab 4. That situation did not happened during the two previous experiments. That could be an indication of Fatigue Effect, which should be double-checked through the corresponding statistical tests that were carried out (section 4.2.3.3).

For the first time, "Time obtaining Domain Model" dependent variable seems to favor considerably subjects not using the artifacts (101.4 min for SSD/SC-subjects and 89.0 min for No_SSD/SC-subjects). Regarding the "Time in lab" independent variable, the usual pattern happened again. It seems again that subjects not using the artifacts needed significantly less time to complete the lab tasks (165.2 min for SSD/SC-subjects and 121.4 min for No_SSD/SC-subjects).

In summary, similarly to previous experiments, we observe some differences between the groups of subjects, but the differences are too small and it could be really hard to find a significant difference when analyzing the dependent variables with respect to Method independent variable's levels.

4.1.4 Experiment IV (Fall 2006)

Table 10 shows some descriptive statistic for the Fall/2006 experiment.

			Lab 3					Lab 4				
Dependent variable		Group	Obs.	Min	Max	Mean	Std. Dev.	Obs.	Min	Max	Mean	Std. Dev.
Correctness	Missing classes	SSD/SC	31	0.07	0.69	0.35	0.17	29	0.14	0.61	0.38	0.10
		No_SSD/SC	31	0.14	0.85	0.42	0.14	31	0	0.61	0.37	0.15
	Useless classes	SSD/SC	31	0	1.14	0.25	0.24	29	0	0.85	0.24	0.19
		No_SSD/SC	31	0	1.28	0.29	0.27	31	0	0.92	0.40	0.22
	Missing relationships	SSD/SC	31	0.25	0.90	0.59	0.19	29	0.40	0.81	0.62	0.12
		No_SSD/SC	31	0.40	1.20	0.72	0.16	31	0.20	0.87	0.65	0.18
	Wrong relationships	SSD/SC	31	0	0.25	0.10	0.07	29	0	0.30	0.13	0.08
		No_SSD/SC	31	0	0.50	0.18	0.11	31	0	0.50	0.16	0.12
	Missing attributes	SSD/SC	31	0.33	1.00	0.64	0.14	29	0.14	0.91	0.62	0.13
		No_SSD/SC	31	0.37	1.33	0.66	0.17	31	0.33	0.93	0.64	0.14
	Wrong attributes	SSD/SC	31	0	0.41	0.17	0.12	29	0	0.29	0.14	0.07
		No_SSD/SC	31	0	0.41	0.17	0.10	31	0	0.33	0.15	0.09
	Averaged correctness	SSD/SC	31	0.18	0.51	0.35	0.09	29	0.23	0.48	0.36	0.06
		No_SSD/SC	31	0.25	0.82	0.41	0.09	31	0.17	0.55	0.40	0.09
Time obtaining Domain Model		SSD/SC	31	67.0	165.6	120.8	28.2	29	20.0	166.5	98.8	41.3
		No_SSD/SC	31	73.0	170.1	128.9	26.4	31	55.0	55.0	106.5	31.9
Time in lab		SSD/SC	31	120.0	190.0	170.5	16.4	29	80.0	188.0	157.3	34.3
		No_SSD/SC	31	120.0	189.0	156.8	22.6	31	102.0	189.0	151.8	27.2

Table 10 Summary of descriptive statistic by considering Method's levels (Experiment IV)

For Experiment IV the results seem to be in better accordance with the expected values. For Lab-3 subjects using the artifacts performed better than the other subjects for every single Correctness measure (0.35 for SSD/SC-subjects and 0.41 for No_SSD/SC-subjects for "Averaged Correctness"), especially for "Missing Classes", "Missing Relationships" and "Wrong Relationships". For the "Time obtaining Domain Model" dependent variable it is shown a similar result to the ones so far commented for most of the labs in other experiments: Subjects dealing with the artifacts required less time to obtain the Domain Model (120.8 min for SSD/SC-subjects and 128.9 min for No_SSD/SC-subjects).

At Lab-4 the behavior was very similar to the one at Lab-3. The only difference was that this time the difference between both groups of subjects (SSD/SC and No_SSD\SC) was less pronounced. This was

perhaps due to a Practice Effect, especially due to the lower time that the groups required to complete the lab tasks.

Overall, results show that the use of the artifacts seems to positively influence the construction of the Domain Model from both the points of view of Quality and Time needed to obtain it.

4.2. Univariate analysis

As it was already mentioned, the univariate analysis involved three statistical tests: One-Sample *t*-test (section 4.2.1), Two-Sample *t*-test (section 4.2.2), and Simple Repeated Measures ANOVA test (section 4.2.3).

4.2.1 One-Sample *t*-test

Recall that a One-Sample *t*-test compares the mean score of a data sample to a known value. For this experiment, this test was used to check the level of understanding reached by the subjects about the software system description used (VS and CPD). The known value used for that comparison was 70 %, which corresponds, according to the standards evaluation used in Canada, to a B- grade, which is considered an average grade.

4.2.1.1 Experiment I (Summer 2005)

Table 11 shows that subjects reached a considerable understanding for both systems. The results also show that for the subjects it was easier to deal with CPD system. For this experiment no outliers were detected. 33 subjects were used for the study.

System	Mean	Standard deviation	H0	p-value (*t*-test)	p-value (Sign Rank)
Car Part Dealer	83.165	11.577	70	**<.0001**	**<.0001**
Video Store	73.844	9.9762	70	**0.0341**	**0.0380**

Table 11 One-Sample *t*-test to evaluate subjects' level of understanding of each system (Experiment I)

4.2.1.2 Experiment II (Fall 2005)

Table 12 shows, once again, that all the subjects involved in the study got a very good understanding of both systems. This time, no significant system difference is observed, though the score for CPD is

higher than the one for VS. No outliers were detected for this experiment. However, only 73 of the 101 subjects were used as a number of them had a course conflict with the lab or missed both labs.

System	Mean	Standard deviation	H0	p-value (*t*-test)	p-value (Sign Rank)
Car Part Dealer	81.05	16.892	70	**<.0001**	**<.0001**
Video Store	78.082	14.221	70	**<.0001**	**<.0001**

Table 12 One-Sample *t*-test to evaluate subjects' level of understanding of each system (Experiment II)

4.2.1.3 Experiment III (Summer 2006)

Table 13 shows again a good level of systems understanding. It is not as good as the two previous experiments though. The corresponding non-parametric statistical test (Sign Rank) for VS system does not show a significant result. Such a test is usually used when the data sample does not have a normal distribution (which is a hypothesis for using a *t*-test). We therefore checked whether we had a normal distribution, using a Kolmogorov-Smirnov test (available in SAS). Table 14 shows the results of testing for sample data normal distribution, showing that the normal distribution hypothesis holds. We can therefore rely on the *t*-test results.

System	Mean	Standard deviation	H0	p-value (*t*-test)	p-value (Sign Rank)
Car Part Dealer	79.722	13.877	70	**0.0055**	**0.0104**
Video Store	76.111	12.37	70	**0.0396**	0.0666

Table 13 One-Sample *t*-test to evaluate subjects' level of understanding of each system (Experiment III)

System	p-value (Shapiro-Wilk)	p-value (Kolmogorov-Smirnov)
Video Store	**0.0053**	**<0.0100**

Table 14 Test for Normality

Three subjects data info was discarded from the study. One of them missed two of the four labs. The others two subjects got scores for the last two labs which were definitely far away from the median. They were considered outliers.

4.2.1.4 Experiment IV (Fall 2006)

Table 15 shows a situation similar to the previous experiments. Subjects got a high level of understanding of each system. That value was significantly superior to the 70% hypothesized value

used for the test. Subjects seemed to have a better understanding of the CPD system than the VS system.

System	Mean	Standard deviation	H0	p-value (*t*-test)	p-value (Sign Rank)
Car Part Dealer	83.165	13.625	70	**<.0001**	**<.0001**
Video Store	73.224	12.567	70	**0.0496**	**0.0184**

Table 15 One-Sample *t*-test to evaluate subjects' level of understanding of each system (Experiment IV)

Only one subject's data were discarded. That subject's performance during Lab-1 was very bad, and was considered an outlier.

4.2.1.5 Summary

A One-Sample *t*-test was executed to check the subjects' levels of understanding of both systems involved in the study, CPD and VD (Table 16).- Overall, a significant level of understanding of both systems was obtained for the four experiments. In all cases the computed value was significantly superior to 70 percent. It has to be pointed out that subjects' level of understanding of CPD system was for most of the experiments superior to the one computed for VS system, thus suggesting that CPD is simpler to understand than VS. Some subjects' data were discarded from some experiments because they clearly did not reach a proper understanding of the software systems during Lab-1 or missed some of the labs.

Experiment	CPD	VS
Experiment I	<0.0001	0.03
Experiment II	<0.0001	<0.0001
Experiment III	0.005	0.03
Experiment IV	<0.0001	0.04

Table 16 Summary of One-Sample *t*-test for the four experiments

4.2.2 Two-Sample *t*-test

A one-side, Two-Sample *t*-test is implemented to compare the means of the dependent variables for both groups of subjects during the summer experiment. This test was performed for both attempts of obtaining the Domain Model. As mentioned previously, this test had two goals: First, to provide

another insight into the data behavior, similar to how Descriptive Statistics did; Second, to serve as an alternative solution in case that a Learning/Fatigue or a Carry-Over Effects were detected when statistically analyzing the data.

4.2.2.1 Experiment I (Summer 2005)

In Appendix A, Table 62 shows the results of applying the one-side, Two-Sample *t*-test for each of the dependent variables at Labs 3 and 4. As discussed below, the table confirms the lack of significant difference between the two groups (see descriptive statistics analysis performed in Section 4.1.1).

In Lab 3, the Correctness dependent variable is far from revealing a significant difference between the two groups of subjects (p-value=0.2310). Only one of the sub-features, namely "Useless Classes" shows a significant difference (p-value=0.0170). According to that result, in general there is no significant difference between subjects who used the artifacts and those who did not use them with respect to the Correctness of the Domain Model obtained. A similar conclusion is reached with respect to the other dependent variable, "Time in lab". There is no significant difference in the time invested to obtain the Domain Model between those subjects who used the artifacts and those who did not. Subjects using the artifacts spent slightly more time at the task than the others. Subjects using the artifacts were complaining much more about the lack of time to complete the whole task, which is shown in Appendix E, Table 165.

The situation is relatively similar for Lab-4. No significant difference is observed between the two groups of subjects for any of the sub-features describing Correctness. As a result no significant difference was found for Lab-4 when analyzing the Domain Model Correctness variable as a whole (p-value=0.4825). Contrary to what was concluded for Lab-3, in Lab-4 we observed a significant difference between the times to complete the tasks (p-value=0.0002). Actually, both groups used less time during Lab-4 compared to Lab-3, but the reduction shown by subjects who did not use the artifacts was really drastic. This severe reduction could be explained by the presence of a Learning Effect, whereby those subjects gained experience after using the artifacts during Lab-3. Furthermore, the subjects not using the artifacts may have felt they had ample time to perform their tasks during Lab-3 and consequently did not feel the same pressure to work fast. Note that the detected difference during Lab-4 favors subjects not using the artifacts. It seems that subjects using the artifacts spent more time to obtain the Domain Model. However, remember that subjects who used the artifacts were required to

perform the additional task of defining contracts, which was also time-consuming. The collected time measured how long the subjects spent in the lab without distinguishing the time needed to obtain the Domain Model. This is the reason why we introduced a second time measure in Experiments II, III, and IV to specifically capture the time subjects spend on domain modeling.

4.2.2.2 Experiment II (Fall 2005)

The results observed for Experiment II were very similar to the ones observed for Experiment I: detailed results are shown Appendix B, Table 93.

A look at Lab-3 results shows a small but insignificant difference for the "Averaged Correctness" variable between the two groups of subjects (p-value=0.1208). The only sub-feature close to showing a significant difference was "Useless Classes" (p-value=0.00913). It has to be pointed out that in general the result favors the subjects dealing with the artifacts. For the same lab, no significant difference was found for the "Time in lab" dependent variable between the subjects who used the artifacts and those who did not.

In Lab-4, no significant difference is shown between SSD/SC subjects and No_SSD/SC subjects for any of the six sub-features describing Correctness. As a consequence, a similar result is observed for the "Averaged Correctness" variable (p-value=0.7374). What is observed is that subjects not using the artifacts in Lab 4 improved from Lab 3 to Lab 4, and that the performance of subjects using the artifacts in Lab 4 was not as good as in Lab 3, thus resulting in a small difference overall between the two groups. A similar situation was pointed out for Experiment I. We can conjecture that using the artifacts in Lab 3 helped the subjects during their second attempt at the task. Again, the small detected difference favored subjects dealing with the artifacts. Statistic tests also show that the difference between the two groups of subjects for "Time obtaining the Domain Model" was not significant (p-value=0.7915), and is much less significant than the one previously observed for Lab-3 (especially because of drastic reduction for subjects not using the artifacts). On the other hand, there is a statistically significant difference between the two groups for "Time in lab" (p-value=0.0117): subjects not dealing with the artifacts needed less time to fulfill the lab tasks.

4.2.2.3 Experiment III (Summer 2006)

Table 112 in Appendix C shows the results of applying Two-Sample Independent *t*-tests for Experiment III.

In Lab-3, no significant difference is shown for any of the Correctness sub-features. The sub-features mean values for both No-SSD/SC subjects and SSD/SC subjects reflect similar level of mistakes. Consequently, not significant difference is observed. As for previous experiments, the "Time obtaining Domain Model" is not significantly different for the two groups (p-value=0.7865): Subjects using the artifacts spent a little less time to obtain their Domain Model than the other subjects.

Observations for Lab-4 are similar to the ones of Lab-3. No significant "Averaged Correctness" difference is observed (p-value=0.7719), though contrary to Lab 3, the "Missing Attributes" values were significantly different (p-value=0.00328). That difference slightly favored subjects using the artifacts. Similarly to previous experiments, the difference between the two groups is much less marked in Lab-4 than in Lab-3, which could be due to a Learning/Fatigue Effect. The "Time obtaining Domain Model" results changed from previous experiments. This time the difference between the two groups of subjects increased, but the difference is still not significant (p-value=0.3225). The only significant difference was once again related to the "Time in Lab" variable (p-value=0.0006). The analysis shows that subjects using the artifacts stayed longer in the lab than the others. In addition to what was previously observed, it is worth noting a decline in the quality of the Correctness variables, which could be due to a Fatigue Effect.

4.2.2.4 Experiment IV (Fall 2006)

Table 143 in Appendix D shows the results of applying Two-Sample *t*-test for Experiment IV. At first glace many significant differences are noticed in that table. It seems that the new strategy followed for this fourth version of the experiment is paying off.

In Lab-3 we observe a significant difference for "Averaged Correctness" between the two groups of subjects (p-value=0.0175). Subjects using the artifacts made fewer mistakes than those not using the artifacts. This is particularly true for the "Missing Classes", "Missing Relationships" and "Wrong Relationships" sub-features. That result was already foreseen from the Descriptive Statistic. Regarding the "Time obtaining the Domain Model" variable, even though subjects using the artifacts needed less

time to obtain the Domain Model, the difference with the other group was not significant (p-value=0.2504).

In Lab-4 we observed the same quality (correctness) reduction we noticed in all previous experiments. This time the "Averaged Correctness" variable was close to show a significant difference between the two groups of subjects, where the one using the artifacts were favored again (p-value=0.0641). The "Time obtaining Domain Model" variable also favored subjects dealing with the artifacts, but the difference was no significant enough, once again (0.4327).

By looking at this experiment as a whole, one can realize that this time there was not an evident data fluctuation from Lab-3 to Lab-4. Despite the data consistency, there was once more a slightly decrease of quality of mean values sub-features describing the correctness of the Domain Model.

4.2.2.5 Summary

Two-Sample Independent *t*-tests were performed for Lab-3 and Lab-4 in the four experiments (Table 17). Recall that the goal was mainly to obtain another insight into the data before getting into Repeated Measures ANOVA.

Experiment	Lab-3	Lab-4
Experiment I	Correctness: 0.23 Time in lab: 0.43	Correctness: 0.48 Time in lab: 0.00
Experiment II	Correctness: 0.12 Time in lab: 0.37 Time obtaining DM: 0.14	Correctness: 0.73 Time in lab: 0.01 Time obtaining DM: 0.79
Experiment III	Correctness: 0.25 Time in lab: 0.03 Time obtaining DM: 0.78	Correctness: 0.77 Time in lab: 0.00 Time obtaining DM: 0.32
Experiment IV	Correctness: 0.01 Time in lab: 0.00 Time obtaining DM: 0.25	Correctness: 0.05 Time in lab: 0.49 Time obtaining DM: 0.43

Table 17 Summary of Two-Sample *t*-test for the four experiments

Only one significant difference between subjects using the artifacts and those not using the artifacts was detected when analyzing the "Averaged Correctness" dependent variable. That significant difference was observed for Experiment IV. Subjects using the artifacts had a better performance than those not using the artifacts. It has to be mentioned that even though no significant difference was observed for the other experiments, the difference was overall always in favor of the subjects who used the artifacts. Experiment I was the only exception, likely due to the fact that the subjects were not

familiar enough with building and using the artifacts under study. This can be checked through questionnaires analysis (Table 162, Page 113) where subjects appeared not to have achieved a good understanding regarding these artifacts.

No significant difference was detected in any of the experiments for the "Time obtaining Domain Model" dependent variable. It is worth mentioning that for all the experiments, subjects using the artifacts needed less time to obtain their Domain Model.

For the "Time in lab" dependent variable, we detected many significant differences for all the experiments. For the four experiments, subjects not using the artifacts were always favored.

4.2.3 Simple Repeated Measures ANOVA test

This test, also called Single or One-Way Repeated Measures ANOVA test, is used to analyze observations taken from the same subjects over time, each time under different conditions (in our case, in Lab 3 and Lab 4). The goal is to check for differences among means of groups. Groups refer to the different conditions under which the subjects will be measured. One particularity of this test is that the repeated measures of only one factor are considered. That is why it is a univariate analysis.

This test not only checks for differences between those subjects using or not the artifacts, but also for the presence of a Learning/Fatigue Effect and Carry-Over Effect [35]. Carry-Over Effect was measured provided that a Cross-Over design was implemented in the experiment. For that, a variable was defined (MethodOrder) to control the order in which Method's treatment was applied by the subjects. This involves two choices: 1) subjects using the artifacts in the first attempt and not using the artifacts in the second attempt, or 2) subjects not using the artifacts in the first attempt and using the artifacts in the second attempt. The Learning/Fatigue Effect was controlled by introducing a "Attempt" variable that defined if the subjects were obtaining their first or their second Domain Model.

It has to be mentioned that in case a Learning/Fatigue Effect or Carry-Over Effect is detected, the data corresponding to the second attempt (i.e., Lab 4) will have to be discarded. In such a case, the Simple Repeated Measures ANOVA will therefore not be useful to check for the "Method" independent variable. In this case we have to resort to using the corresponding results obtained when applying Two-Sample *t*-test for Lab-3.

Results for Simple Repeated Measures ANOVA test on each experiment version are presented next.

4.2.3.1 Experiment I (Summer 2005)

Table 18 shows the statistical results for the three factors in association with "Averaged Correctness" dependent variable. No significant difference is shown between any of the three factors' levels.

Effect	DF	F-Ratio	Prob > F
Method	1	0.34	0.5673
Method Order	1	1.19	0.2840
Attempt	1	0.34	0.5673
Covariance Structure	Toeplitz		

Table 18 Simple Repeated Measures ANOVA analysis for "Averaged correctness" feature (Experiment I)

No significant difference in Domain Models quality was identified between the two groups (Prob>F=0.5673). This is in accordance with results (i.e., no significant differences) shown for each of the sub-features describing the Correctness dependent variable (Appendix A, from Table 63 to Table 68). Such a result can be better understood after checking the mean values provided by Table 7 (Page 32) and Table 62 (Appendix A, Page 80). Data scores in both groups were very similar.

Three reasons could explain the above observation. First, there could be an interaction of the "Method" factor with subjects' ability or with systems' level of complexity. Addressing this issue has to be postponed until we consider multivariate analysis. A second reason could be the presence of a Learning Effect. However, Table 18 clearly shows that no Learning Effect was detected for the "Averaged Correctness" variable (Prob>F=0.2840). (This also applies to all the Correctness sub-features, as illustrated in Appendix A, from Table 63 to Table 68.) A third reason could be the poor subjects' understanding of the artifacts. This is a highly plausible explanation, confirmed by the analysis of a questionnaire (Table 162, Page 113): a number of questions (specifically questions 2, 3, and 4) clearly show a poor understanding of SSDs and SOCs, although they have been trained in a classroom setting.

The situation was completely different for the "Time in lab" dependent variable. Table 19 shows a significant difference for the three analyzed factors. A significant difference exists between the time required by subjects who used the artifacts and those who did not ("Prob > F"= 0.0002). That result is in agreement with what was partially obtained with descriptive statistics and Two-Sample *t*-test (Appendix A, Table 62).

Effect	DF	F-Ratio	Prob > F
Method	57	15.77	**0.0002**
Method Order	57	10.28	**0.0022**
Attempt	57	19.26	**<0.0001**
Covariance Structure	Toeplitz		

Table 19 Simple Repeated Measures ANOVA analysis for "Time in lab" feature (Experiment I)

The same test also suggests that a Practice Effect was present in the experiment in association with the "Time in lab" variable ("Prob>F" < 0.0001). There exists also a significant influence of the order in which methods were applied ("Prob>F" = 0.0022). Remember that in this first experiment we only measured the time the subjects spent in the lab, without considering the specific time needed to obtain the Domain Model.

The Simple Repeated Measures ANOVA cannot be used to explain the behavior of the "Method" independent variable with respect to the "Time in lab" dependent variable, since we have observed a practice effect. Data of Lab-4 have to be discarded. As a result, the "Time in lab" variable behavior has to be explained through the Two-Sample *t*-test performed for Lab-3 (Table 15, Page 39), which did not reveal any significant difference between the time needed by both groups of subjects to complete the lab's tasks.

4.2.3.2 Experiment II (Fall 2005)

Table 20 confirms what was observed through the Descriptive Statistic regarding the lack of apparent difference between the groups of subjects with respect to Domain Model Correctness ("Prob>F"=0.4095). This time the difference between the two groups of subjects was larger (i.e. a smaller p-value) than the one computed for Experiment I, and this could be due to the larger sample size in Experiment II. The table additionally shows that no Carry-Over or Learning/Fatigue Effect was detected ("Prob>F"=0.1897 and "Prob>F"==09797, respectively). We observed the same results, i.e., no significant difference, for every sub-feature of Correctness, as reported in Appendix B (Table 94 to Table 99, Page 90). Recall that we reached the same conclusion for Experiment I. The lack of significant difference can , once again, be explained by insufficient knowledge of the subjects, although they have been trained: the questionnaire (Table 169, Page 116) results, reported in Table 171 and Table 173 (Pages 117 and 118), show that subjects dealing with the artifacts strongly complained about

needing more time to complete the lab tasks. It is worth mentioning that subjects using the artifacts did slightly better than subjects not using them.

Effect	DF	F-Ratio	Prob > F
Method	125	0.69	0.4095
Method Order	125	1.74	0.1897
Attempt	125	0.00	0.9797
Covariance Structure	Toeplitz		

Table 20 Simple Repeated Measures ANOVA analysis for "Averaged Correctness" feature (Experiment II)

From Table 21 we see that only one factor showed a significant difference when analyzing the "Time Obtaining Domain Model" dependent variable ("Prob>F"=0.0337). This suggests a Learning Effect because corresponding Descriptive Statistics (Table 8, Page 34) show that the quality of solutions in both labs were almost the same, and only the mean time of both groups was reduced. As a consequence the dependent variable behaviour has to be explained by just using Two-Sample *t*-test results for Lab-3 (Table 93, Page 90). Results showed that subjects using the artifacts appear to need less time to obtain the Domain Model than the rest of the subjects. That difference was not significant ("p-value"=0.1400).

Effect	DF	F-Ratio	Pros > F
Method	57	3.12	0.0799
Method Order	57	0.01	0.9310
Attempt	57	4.61	**0.0337**
Covariance Structure	Toeplitz		

Table 21 Simple Repeated Measures ANOVA analysis for "Time obtaining Domain Model" feature (Experiment II)

Table 22 shows a significant difference for "Method" when analyzing the "Time in lab" dependent variable ("Prob>F"=0.0015). That result was expected as Descriptive Statistics showed an important difference between the two groups at both labs (Table 8, Page 34). That difference, which always favored subjects not using the artifacts, was significant only for Lab-4.

Effect	DF	F-Ratio	Pros > F
Method	125	11.07	**0.0015**
Method Order	125	1.40	0.2405
Attempt	125	11.25	**0.0013**
Covariance Structure	Toeplitz		

Table 22 Simple Repeated Measures ANOVA analysis for "Time in lab" feature (Experiment II)

The "Time" factor also shows a significant difference ("Prob>F"=0.0013). This result suggests the presence of Learning Effect. As in Experiment I, similar levels of solutions quality were observed from Lab-3 to Lab-4, with less time needed to complete the lab tasks in Lab 4. This Learning Effect could be associated to a better understanding by subjects of the tasks to carry out. That theory is supported by the score improvement observed for questions 1 and 2 of the questionnaires used in Lab-3 to Lab-4 (Appendix E, Table 171 and Table 173, Page 117). That Learning Effect leads us to use the Two-Sample *t*-test performed for Lab-3 to check for significant differences between the two "Method" levels (Section 4.2.3.2, Page 46). No significant difference was found between the time required by subjects using the artifacts to obtain the Domain Model and those others no using the artifacts

4.2.3.3 Experiment III (Summer 2006)

Table 23 shows that no significant difference in quality was detected between subjects using the artifacts and those who did not ("Prob>F"=0.3138.). Similarly to Experiment II, Experiment III showed a larger difference than Experiment I. That difference favored subjects using the artifacts. In Experiment III, subjects seemed to have a better understanding of the artifacts under study than in previous Experiments (questionnaire answers in Table 173, Page 118). This better understanding could have influenced that increment in Domain Model correctness difference (though still not significant) between subjects using the artifacts and subjects not using them.

Effect	DF	F-Ratio	Prob > F
Method	1	1.07	0.3138
Method Order	1	0.26	0.6149
Attempt	1	3.50	0.0778
Covariance Structure	Toeplitz		

Table 23 Simple Repeated Measures ANOVA analysis for "Averaged Correctness" sub-feature (Experiment III)

Results associated with the six correctness sub-features are completely in accordance with the result previously discussed: no significant difference was detected (Appendix C, from Table 113 to Table 118, Page 97).

Table 24 shows the results of the Repeated Measures ANOVA test for the "Time obtaining Domain Model" dependent variable. This time the difference between the two "Method" levels was not as pronounced as what was observed so far. That decrease was mainly influenced by an abrupt reduction of time needed by subjects not using the artifacts in Lab-4 (Appendix C, Table 112, Page 97), which could be associated to a Learning Effect (levels of quality from Lab-3 to Lab-4 were essentially the same). But no Learning Effect is visible in Table 24 ("Prob>F"=0.1404).

Effect	DF	F-Ratio	Prob > F
Method	1	0.35	0.5629
Method Order	1	0.79	0.3868
Attempt	1	2.38	0.1404
Covariance Structure	Toeplitz		

Table 24 Simple Repeated Measures ANOVA analysis for "Time obtaining Domain Model" feature (Experiment III)

Table 25 shows results for the "Time in lab" dependent variable. Results are the same as in previous experiments. There is a significant difference between the times spent by the two groups to accomplish the lab's tasks ("Prob>F"<0.0001). Subjects not using the artifacts, once again, required less time to complete the task. The table also shows a Learning Effect ("Prob>F"=0.0176). As a result, the "Method" factor for the "Time in lab" dependent variable has to be analyzed by a Two-Sample *t*-test that only considered the data for Lab-3 (Table 112, Page 97). That test showed also a significant difference for the time needed by both groups of subjects to finish the lab's tasks. Again, that difference favored those subjects not using the artifacts. That longer time needed by subjects using the

artifacts could explain the lack of significant difference in Domain Model correctness with respect to subjects not using the artifacts.

Effect	DF	F-Ratio	Prob > F
Method	1	23.59	**<.0001**
Method Order	1	2.51	0.1290
Attempt	1	6.70	**0.0176**
Covariance Structure	Toeplitz		

Table 25 Simple Repeated Measures ANOVA analysis for "Time in lab" feature

4.2.3.4 Experiment IV (Fall 2006)

Table 26 shows a significant difference in Domain Model Correctness between subjects using the artifacts and subjects not using them ("Prob>F"=0.0018). This is consistent with what was observed in the corresponding Descriptive Statistics (Table 10, Page 36) and Two-Sample *t*-test (Table 143, Page 106). Subjects using the artifacts were favored, even though questionnaires show, as in previous experiments, that subjects strongly complained about time constraints to complete the lab tasks (questionnaire answers in Table 183, page 122). Recall that for the previous experiments, subjects using the artifacts did slightly better in general, but no significant result was observed. This significant result, which was not observed in previous experiments, can be explained by the better background understanding of the subjects, than in previous experiments, as illustrated by the questionnaire answers (Table 180, Page 121).

Effect	DF	F-Ratio	Prob > F
Method	1	10.66	**0.0018**
Method Order	1	0.34	0.5619
Time	1	0.03	0.8600
Covariance Structure	Toeplitz		

Table 26 Simple Repeated Measures ANOVA analysis for "Averaged Correctness" feature (Experiment IV)

We also observe similar results for the Correctness sub-features (Appendix D, from Table 144 to Table 149, Page 106). The sub-features "Useless Classes" ("Prob>F"=0.0264), "Missing Relationships" ("Prob>F">0.0058) and "Wrong Relationships" ("Prob>F"=0.0011) were especially involved in the overall result of "Averaged Correctness".

Table 27 shows results for the "Time obtaining Domain Model" dependent variable, which are similar to previous experiments: no significant differences, and mean values favoring the subjects using the artifacts. The table also shows a Learning Effect ("Prob>F"<0.0001). This leads us to use results already computed for the Two-Sample *t*-test for Lab-3 (Table 143, Page106). That test revealed a no significant difference between the two groups of subjects ("p-value"=0.2504).

Effect	DF	F-Ratio	Prob > F
Method	1	2.15	0.1482
Method Order	1	0.01	0.9047
Attempt	1	20.05	**<0.0001**
Covariance Structure	Toeplitz		

Table 27 Simple Repeated Measures ANOVA analysis for "Time obtaining Domain Model" feature (Experiment IV)

Table 28 shows that for the "Time in lab" dependent variable there is a significant difference that again favored subjects not using the artifacts ("Prob>F"=0.0048). The table also shows a significant difference between the "Time" factor's levels ("Prob>F">0.0064). That difference is an indicator of a Learning Effect. As a consequence, the result of the Two-Sample *t*-test for Lab-3 is the result to be considered (Table 143, Page 106): A significant difference was detected ("p-value"=0.0087).

Effect	DF	F-Ratio	Prob > F
Method	1	8.62	**0.0048**
Method Order	1	0.57	0.4547
Attempt	1	8.02	**0.0064**
Covariance Structure	Toeplitz		

Table 28 Simple Repeated Measures ANOVA analysis for "Time in lab" feature (Experiment IV)

4.2.3.5 Summary

Simple Repeated Measures ANOVA tests were used to check for Learning/Fatigue Effect, Carry-over Effect, and significant difference between "Method" independent variable's levels. Table 29 shows a summary of results obtained for the four experiments.

In some cases we detected a Learning/Fatigue or a Carry-Over Effect when analyzing a given dependent variable. In those cases, we therefore decided not to use the result of the Simple Repeated Measures ANOVA test involving the "Method" factor. Instead, a conclusion regarding the "Method"

factor was reached through the corresponding Two-Sample Independent *t*-test described in Section 4.2.2.

Overall, when analyzing the "Averaged Correctness" dependent variable there was a trend favoring subjects who used the artifacts. That trend increased from Experiment II to Experiment IV. For the last experiment, the difference in Domain Model quality between subjects using the artifacts and those not using them was statistically significant. That behavior could have been related to the facts that the subjects understood better the benefits of performing this exercise and also they were more familiar with the artifacts that they dealt with. Experiment I did exhibit the same trend; possibly because subjects' understanding of how to use the artifacts and the advantage of using them was not sufficient (Table 162, Page 113). In general, we can conclude, in terms of Domain Model Correctness, that a much better Domain Model can be obtained in practice when dealing with the two artifacts (SSD and SC) if practitioners are familiar enough with using such artifacts.

Simple Repeated Measures ANOVA			
Experiment	**Method**	**Method Order**	**Attempt**
Experiment I	Correctness: 0.57 Time in lab: 0.0002	Correctness: 0.28 Time in lab: 0.002	Correctness: 0.57 Time in lab: 0.0001
Experiment II	Correctness: 0.41 Time in lab: 0.001 Time obtaining DM: 0.08	Correctness: 0.19 Time in lab: 0.24 Time obtaining DM: 0.93	Correctness: 0.98 Time in lab: 0.0013 Time obtaining DM: 0.03
Experiment III	Correctness: 0.31 Time in lab: 0.0001 Time obtaining DM: 0.56	Correctness: 0.61 Time in lab: 0.13 Time obtaining DM: 0.39	Correctness: 0.07 Time in lab: 0.01 Time obtaining DM: 0.14
Experiment IV	Correctness: 0.0018 Time in lab: 0.0048 Time obtaining DM: 0.15	Correctness: 0.56 Time in lab: 0.45 Time obtaining DM: 0.90	Correctness: 0.86 Time in lab: 0.006 Time obtaining DM: 0.0001

Table 29 Summary of Simple Repeated Measures ANOVA for the four experiments

For the "Time obtaining Domain Model" the results were exactly the same for the three experiments were this dependent variable was considered (i.e., Experiments, II, III, and IV). No significant difference was found between subjects using the artifacts and those not using them. It is worth highlighting that even though that difference was never significant, those subjects using the artifacts spent always less time obtaining the Domain Model than subjects not using the artifacts. However, it cannot be concluded that with more training subjects using the artifacts to obtain the Domain Model would need much less time than subjects not using the artifacts. Subjects already received much more training during Experiment IV, and the time value ranges were very similar to the ones computed for the previous two experiments.

Regarding the "Time in lab" dependent variable the trend was exactly the same for the four experiments. Subjects not using the artifacts completed the tasks in less time than subjects using the artifacts. Though this may seem to be expected since subjects not using the artifacts had less tasks and issues to deal with (e.g., they did not have to create contracts), the question was whether the gain in common tasks would somehow compensate.

4.3. Multivariate analysis

A multivariate analysis has to do with performing a statistical analysis where the behaviour of more two or more independent variables is analyzed at once. Two approaches were followed for all the experiments: N-Way ANOVA and Mixed Repeated Measures ANOVA tests. In general, both tests help finding out whether data from several groups have a common mean or not by considering more than two categories (independent variables). Those tests are useful to study the influence of each main effect (independent variable) independently as well as the interactions among them.

Those two statistical tests will be used to study the influence of two factors that could be confounded with "Method" factor: "System" and "Ability". The first one refers to the two systems used to carry out the experiment: CPD and VS. The second one refers to the subjects' skills in Software Engineering.

N-Way ANOVA test is used when the assumptions of normal distribution and independence among the samples are satisfied. A Mixed Repeated Measures ANOVA test is used when the second assumption is not fulfilled. In addition, deciding among the two tests depends on whether there is Learning/Fatigue or Carry-Over Effect when analyzing the results of a Simple Repeated Measures ANOVA test. For example, for our experiment, if for a given dependent variable a Learning Effect was detected, then an N-Way ANOVA test has to be used because only data from Lab-3 can be taken into consideration. If no such effects are detected, then the Mixed Repeated Measures ANOVA test can be used because data from the two labs can be used for the analysis. Those tests are accompanied by some Descriptive Statistics that allow a better understanding of the results.

Last, recall that when the sample size is large enough, we can consider the three factors ("Method", "System", and "Ability") at once. This happened for Experiment II and Experiment IV. Instead, when the sample size is too small (i.e., Experiment I and Experiment III), then pairs of factors are considered ("Method" & "System", and "Method" & "Ability").

Note that the results for the "Method" main effect are not discussed in the following sub-sections because they are exactly the same as the ones we obtained from Simple Repeated Measures ANOVA tests.

4.3.1 Experiment I (Summer 2005)

A Mixed Repeated Measures ANOVA test was used to analyze "Correctness" (section 4.3.1.1) and a Two-Way ANOVA test was used to analyze "Time in lab" (section 4.3.1.2).

4.3.1.1 Correctness

Table 30 and Table 31 show the results of analyzing the combination of the "Method" and "Ability" factors for the "Averaged Correctness" dependent variable. Results show no significant difference for any of the two main effects. There is no significant difference between high-ability subjects and low-ability subjects ("Prob>F"=0.3863). Table 30 also shows that there is no interaction between "Method" and "Ability" ("Prob>F"=0.6049).

Source	DF	F-Ratio	Prob > F
Method	1	0.00	0.9918
Ability	1	0.76	0.3863
Method * Ability	1	0.27	0.6049

Table 30 Mixed Model Repeated Measures ANOVA (Method & Ability) analysis for "Averaged Correctness" feature (Experiment I)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.34	12	0.33	4	0.34	16
		SSD/SC	0.36	11	0.39	6	0.37	17
		All Methods	0.35	23	0.36	10		
	Lab 4	No_SSD/SC	0.36	10	0.43	5	0.38	15
		SSD/SC	0.36	12	0.36	4	0.36	16
		All Methods	0.36	22	0.37	9		

Table 31 Descriptive statistics for the test reported in Table 30 (Experiment I)

Once again, no significant main and interaction effects were obtained when independently analyzing the Correctness sub-features (Table 69 to Table 80), with the exception of the "Missing Association" sub-feature (Table 75, Table 76, and Figure 17) ("Prob>F"=0.0046). This one significant difference

associated to the "Ability" main effect did not lead to similar result for Domain Model Correctness as a whole ("Averaged Correctness").

Table 32 and Table 33 show results associated to the "Method" and "System" main effects combination for the "Averaged Correctness". No significant main effects were obtained. There is no significant difference in Domain Model Correctness between subjects dealing with CPD system and those dealing with VS system. No interaction between factors was detected.

Source	DF	F-Ratio	Prob > F
Method	1	0.06	0.8019
System	1	0.25	0.6185
Method * System	1	1.53	0.2216

Table 32 Mixed Model Repeated Measures ANOVA (Method & System) analysis for "Averaged Correctness" feature (Experiment I)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.38	8	0.30	8	0.34	16
		SSD/SC	0.36	8	0.39	9	0.37	17
		All Methods	0.37	16	0.34	17		
	Lab 4	No_SSD/SC	0.38	8	0.39	7	0.38	15
		SSD/SC	0.37	8	0.36	8	0.36	16
		All Methods	0.37	16	0.37	15		

Table 33 Descriptive statistics for the test reported in Table 32 (Experiment I)

Different trends are observed when looking at the Correctness sub-features (Appendix A, from Table 81 to Table 92, Page 85). A significant main effect of the "System" factor is observed for the following sub-features:

- "Useless Classes" (Appendix A, Table 83, Table 84 and Figure 18) ("Prob>F"=0.0029)
- "Missing Associations" (Appendix A,, Table 85, Table 86 and Figure 19) ("Prob>F"=0.0288)
- "Wrong Attribute" (Appendix A,, Table 89, Table 90 and Figure 20) ("Prob>F"=0.0146)
- "Missing Attributes" (Appendix A,, Table 91, Table 92, and) ("Prob>F"=0.0021)

Those four significant main effects on Correctness sub-features had no repercussion at all on the "Averaged Correctness" dependent variable because of the differences in results across the two

systems. Half of the time the CPD system was favored whereas the VS system was favored the rest of the time.

4.3.1.2 Time in lab:

Table 32 and Table 33 show the results of applying a Two-Way ANOVA test for the "Time in lab" dependent variable by considering the "Method" and "Ability" factors. There is no significant main effect for the "Ability" factor ("Prob>F"=0.3843). High-ability subjects and low-ability subjects stayed approximately the same time in the lab. There is no interaction between the two factors ("Prob>F"=0.6152).

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob > F
Method	1	46.52268519	46.52268519	0.63	0.4340
Ability	1	57.68935185	57.68935185	0.78	0.3843
Method * Ability	1	19.07824074	19.07824074	0.26	0.6152
Residual	28	2067.683333	73.845833		
Total	31	2198.000000			

Table 34 Two-Way ANOVA (Method & Ability) analysis at task 3 for "Time in lab" feature (Experiment I)

		Ability					
		High		Low		All Abilities	
		Mean	Size	Mean	Size	Mean	Size
Method	No_SSD/SC	174.25	12	175.50	4	174.56	16
	SSD/SC	175.20	11	179.83	6	176.94	17
	All Methods	174.68	23	178.10	10		

Table 35 Descriptive statistics for the test reported in Table 34

The same situation was observed when considering the "Method" and "System" factors for the "Time in lab" variable (Table 36 and Table 37). Subjects dealing with both software system stayed about the same amount of time in the lab ("Prob>F"=0.3343). No interaction is observed between the "Method" and "System" factors ("Prob>F"=0.9676).

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob > F
Method	1	45.12500000	45.12500000	0.61	0.4424
System	1	72.00000000	72.00000000	0.97	0.3334
Method * System	1	0.12500000	0.12500000	0.00	0.9676
Residual	28	2080.750000	74.312500		
Total	31	2198.000000			

Table 36 Two-Way ANOVA (Method & System) analysis at task 3 for "Time in lab" feature (Experiment I)

		System					
		CPD		VS		All Systems	
		Mean	Size	Mean	Size	Mean	Size
Method	No_SSD/SC	176.00	8	173.13	8	174.56	16
	SSD/SC	178.50	8	175.38	9	176.94	17
	All Methods	177.25	16	174.25	17		

Table 37 Descriptive statistic associated to the previous Two-Way ANOVA analysis

4.3.2 Experiment II (Fall 2005)

A Mixed Repeated Measures ANOVA was applied for the "Correctness" dependent variable (section 4.3.2.1). A Three-Way ANOVA test was used to analyze the "Time obtaining Domain Model" and "Time in lab" dependent variables (section 4.3.2.2 and 4.3.2.3, respectively).

4.3.2.1 Correctness

Table 38 and Table 39 show the results of applying a Mixed Repeated Measures ANOVA test for the "Averaged Correctness" dependent variable by considering the "Method", "System" and "Ability" factors at the same time. There is no significant main effect for factors "Method" ("Prob>F"=0.4902), "System" ("Prob>F"=0.0557) and "Ability" ("Prob>F"=0.1727). Results show that the "System" factor was close to reveal a significant main effect that favored subjects working on CPD (Figure 6).

Similar results were obtained for the Correctness sub-features (Appendix B, Table 100 to Table 111, Page 92). However, "Ability" had a significant main effect for the "Missing Classes" and "Wrong Associations" sub-features. For "Missing Classes" sub-feature, high-ability subjects were favored ("Prob>F"= 0.0312). For "Wrong Associations" low-ability subjects were favored ("Prob>F"= 0.0273). Only two sub-features were affected and in each one a different subject ability level was favored. That is why no significant main effect of "Ability" factor was revealed for the "Averaged Correctness" variable.

Source	DF	F-Ratio	Prob > F
Method	1	0.48	0.4902
System	1	3.73	0.0557
Method * System	1	0.11	0.7368
Ability	1	1.88	0.1727
Method * Ability	1	0.03	0.8640
System * Ability	1	2.37	0.1262
Method * System * Ability	1	0.24	0.6220

Table 38 Mixed Model Repeated Measures ANOVA (Method, System & Ability) analysis for "Averaged Correctness" feature (Experiment II)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.43	18	0.43	15	0.43	33
		SSD/SC	0.39	19	0.42	16	0.40	35
		All Methods	0.41	37	0.42	31		
	Lab 4	No_SSD/SC	0.39	16	0.43	19	0.41	35
		SSD/SC	0.42	15	0.42	18	0.42	33
		All Methods	0.41	31	0.42	37		

Table 39 Descriptive statistic associated to the previous Mixed Model Repeated Measures ANOVA analysis

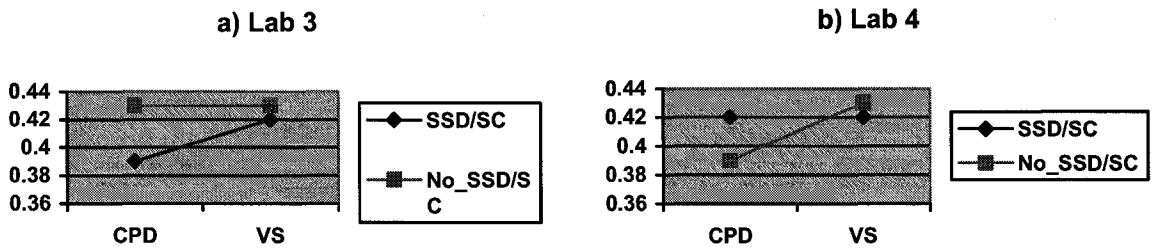

Figure 6 Graph of means (Method & System & Ability) for "Averaged Correctness" feature

Only the "Useless Classes" sub-feature shows a significant main effect for the "System" factor ("Prob>F"<0.0001). Additionally, low values computed for the same factor for "Wrong Association", "Wrong Attributes" and "Missing Attributes" sub-features contributed to that almost significant main effect of "Averaged Correctness" dependent variable for "System" factor.

4.3.2.2 Time obtaining Domain Model

Using a Three-Way ANOVA test, Table 40 and Table 41 show that no significant main effect was detected, except for the "Method" factor. In section 4.2.3.2 we concluded that subjects dealing with the artifacts consumed slightly less time to obtain the Domain Model.

For this analysis were considered only measures from Lab-3 because of the Learning Effect detected through the Simple Measures Repeated ANOVA (section 4.2.3.2). Nevertheless, this main effect is not supported by the corresponding Two-Sample *t*-test performed for the same independent variable at Lab-3 (section 4.2.2.2).

Source	DF	Sum of Squares	Mean Square	F-Ratio	Pros > F
Method	1	4462.132886	4462.132886	4.68	**0.0349**
System	1	331.497261	331.497261	0.35	0.5578
Method * System	1	1520.154327	1520.154327	1.60	0.2120
Ability	1	348.747382	348.747382	0.37	0.5477
Method * Ability	1	1856.281367	1856.281367	1.95	0.1685
System * Ability	1	1031.037891	1031.037891	1.08	0.3029
Method * System * Ability	1	994.310664	994.310664	1.04	0.3115
Residual	1	51451.51148	952.80577		
Total	1	60347.21274			

Table 40 Three-Way ANOVA (Method & Ability & System) analysis at task 3 during the fall experiment for "Time obtaining Domain Model" feature (Experiment II)

		System					
		CPD		VS		All Systems	
		Mean	Size	Mean	Size	Mean	Size
Method	No_SSD/SC	122.92	18	118.35	15	120.79	33
	SSD/SC	116.38	19	99.40	16	108.95	35
	All Methods	119.46	37	108.88	31		

Table 41 Descriptive statistic for the test reported in Table 40

a) Lab 3

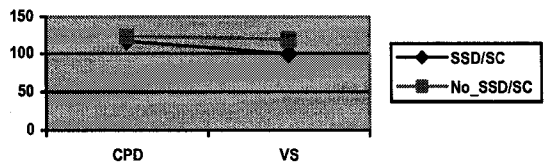

Figure 7 Graph of means (Method & System & Ability) for the "Time obtaining Domain Model" feature

4.3.2.3 Time in lab

Table 42 and Table 43 show the result of applying a Three-Way ANOVA test. No significant main effect was detected for any of the factors. An important interaction was found between the "Method" and "Ability" factors ("Prob>F"=0.0232). According to the Descriptive Statistic (Table 43), high-ability subjects did better than low-ability subjects when dealing with the artifacts to obtain the Domain Model.

Source	DF	Sum of Squares	Mean Square	F-Ratio	Pros > F
Method	1	44.8170	44.8170	0.09	0.7655
System	1	56.9430	56.9430	0.11	0.7367
Method * System	1	708.7371	708.7371	1.42	0.2383
Ability	1	193.1575	193.1575	0.39	0.5363
Method * Ability	1	2717.8174	2717.8174	5.45	**0.0232**
System * Ability	1	3318.5307	3318.5307	6.65	**0.0125**
Method * System * Ability	1	6.6831	6.6831	0.01	0.9083
Residual	1	28441.0277	498.9654		
Total	1	38167.5384			

Table 42 Three-Way ANOVA (Method & Ability & System) analysis at Lab-3 during the fall experiment for "Time obtaining Domain Model" feature (Experiment II)

		Ability					
		High		Low		All Abilities	
		Mean	Size	Mean	Size	Mean	Size
Method	No_SSD/SC	149.67	26	168.57	7	153.94	33
	SSD/SC	163.11	27	144.86	8	159.35	35
	All Methods	156.78	53	156.71	15		

Table 43 Descriptive statistic for the test reported in Table 42

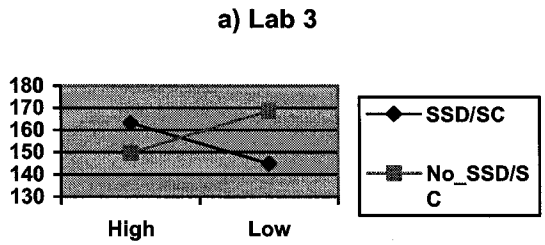

Figure 8 Graph of means (Method & System & Ability) for the "Time obtaining Domain Model" feature

4.3.3 Experiment III (Summer 2006)

A Mixed Repeated Measures ANOVA test was used to analyze the "Correctness" and "Time obtaining Domain Model" dependent variables (section 4.3.3.1 and 4.3.3.2, respectively) and a Two-Way ANOVA test was used to analyze the "Time in lab" dependent variable (section 4.3.3.3).

4.3.3.1 Correctness

Table 44 and Table 45 show the results of applying a Mixed Repeated Measures ANOVA test for the "Averaged Correctness" dependent variable by considering the "Method" and "Ability" factors. No significant main effect was found for any of the factors. No significant interaction was either detected between the two factors. Similar results were obtained when applying the same test for each of the sub-features describing Correctness (Appendix C, from Table 119 to Table 130, Page 99).

Source	DF	F-Ratio	Prob > F
Method	1	0.13	0.7184
Ability	1	1.80	0.1880
Method * Ability	1	0.02	0.9031

Table 44 Mixed Model Repeated Measures ANOVA (Method & Ability) analysis for "Averaged Correctness" feature (Experiment III)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.39	10	0.32	1	0.38	11
		SSD/SC	0.37	8	0.29	2	0.35	10
		All Methods	0.38	18	0.30	3		
	Lab 4	No_SSD/SC	0.41	8	0.39	2	0.41	10
		SSD/SC	0.39	10	0.46	1	0.40	11
		All Methods	0.40	18	0.42	3		

Table 45 Descriptive statistic for the test reported in Table 44

Table 46 and Table 47 present the results of applying the same test for the "Averaged Correctness" dependent variable, but this time by considering the "Method" and "System" factors. Once again, no significant main effect or interaction between factors was detected. Results also show that the "System"

factor was close to having a significant main effect. Similarly to what was observed for Experiment II, the subjects dealing with the CPD system did better than those using the VS system (Figure 9).

Source	DF	F-Ratio	Prob > F
Method	1	1.08	0.3064
System	1	3.27	0.0792
Method * System	1	1.29	0.2639

Table 46 Mixed Model Repeated Measures ANOVA (Method & System) analysis for "Averaged Correctness" feature (Experiment III)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.35	6	0.42	5	0.38	11
		SSD/SC	0.36	5	0.34	5	0.35	10
		All Methods	0.36	11	0.38	10		
	Lab 4	No_SSD/SC	0.39	5	0.44	5	0.41	10
		SSD/SC	0.38	5	0.42	6	0.40	11
		All Methods	0.38	10	0.42	11		

Table 47 Descriptive statistic for the test reported in Table 46

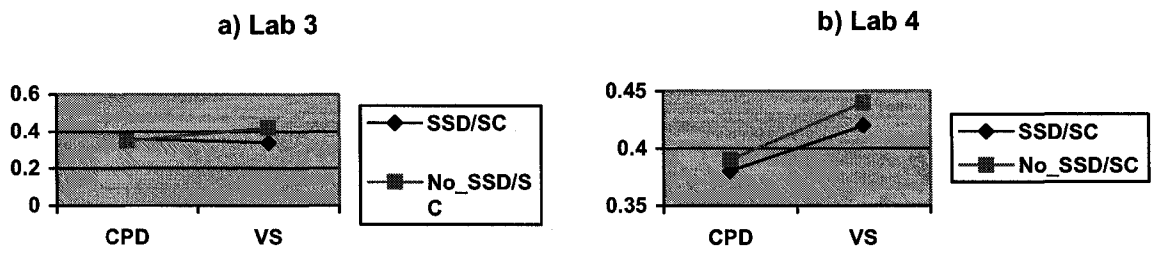

Figure 9 Graph of means (Method & System & Ability) for the "Averaged Correctness" feature

Appendix C, from Table 131 to Table 142 (Page 101), shows that similar results were computed for most of the sub-features describing Correctness. Some of the sub-features show or were close to show a significant main effect for "System" factor: "Missing Classes", Useless Classes", "Wrong Attributes" and "Missing Attributes".

4.3.3.2 Time obtaining Domain Model

Table 48 and Table 49 show the results of applying a Mixed Repeated Measures ANOVA test and considering the combination of factors "Method" and "Ability". No significant main effect of the factors or interaction among them was detected.

Source	DF	F-Ratio	Prob > F
Method	1	0.13	0.7254
Ability	1	0.01	0.9093
Method * Ability	1	0.02	0.8979

Table 48 Mixed Model Repeated Measures ANOVA (Method & Ability) analysis for "Time for Domain Model" feature (Experiment III)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	109.23	10	106.50	1	108.98	11
		SSD/SC	104.64	8	109.50	2	105.72	10
		All Methods	107.34	18	108.5	3		
	Lab 4	No_SSD/SC	88.58	8	92.70	2	89.03	10
		SSD/SC	101.67	10	99.75	1	101.50	11
		All Methods	95.85	18	96.23	3		

Table 49 Descriptive statistic for the test reported in Table 48

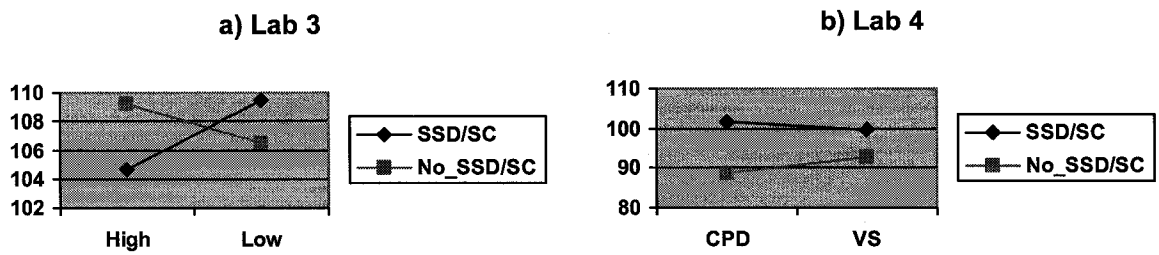

Figure 10 Graph of means (Method & Ability) for "Time obtaining the Domain Model" feature

The situation was different when considering the "Method" and "System" factors. Table 50 and Table 51 show a significant main effect of "System" ("Prob>F"=0.0034). Subjects dealing with the VS system needed in general less time than subjects dealing with the CPD system.

Source	DF	F-Ratio	Prob > F
Method	1	0.37	0.5449
System	1	9.89	**0.0034**
Method * System	1	1.47	0.2330

Table 50 Mixed Model Repeated Measures ANOVA (Method & System) analysis for "Time for Domain Model" feature (Experiment III)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	122.5	6	92.75	5	108.98	11
		SSD/SC	126.00	5	80.38	5	105.72	10
		All Methods	124.09	11	87.25	10		
	Lab 4	No_SSD/SC	86.61	5	92.06	5	89.03	10
		SSD/SC	115.15	5	90.12	6	101.50	11
		All Methods	100.88	10	90.90	11		

Table 51 Descriptive statistic for the test reported in Table 50

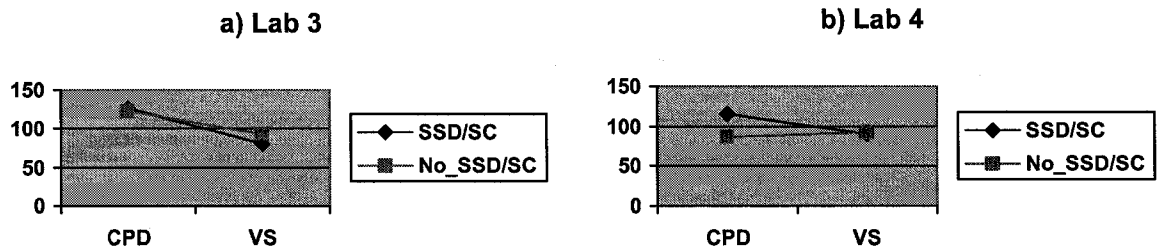

Figure 11 Graph of means (Method & System) for "Time obtaining Domain Model" feature

4.3.3.3 Time in lab

Table 52 and Table 53 show results of a Two-Way ANOVA test for the "Time in lab" dependent variable by considering the "Method" and "Ability" factors. No significant main effect or interaction was detected.

Source	DF	Sum of Squares	Mean Square	F-Ratio	Pros > F
Method	1	1675.7715	1675.7715	3.30	0.0882
Ability	1	2.8863	2.8863	0.01	0.9409
Method * Ability	1	111.7715	111.7715	0.22	0.6455
Residual	16	8133.7571	508.35982		
Total	19	10660.2000			

Table 52 Two-Way ANOVA (Method & System) analysis at Lab-3 during the summer06 experiment for "Time in lab" feature (Experiment III)

		Ability					
		High		Low		All Abilities	
		Mean	Size	Mean	Size	Mean	Size
Method	No_SSD/SC	150.10	10	142.00	1	149.36	11
	SSD/SC	170.14	8	176.00	2	171.44	10
	All Methods	158.35	18	164.67	3		

Table 53 Descriptive statistic for the test reported in Table 52

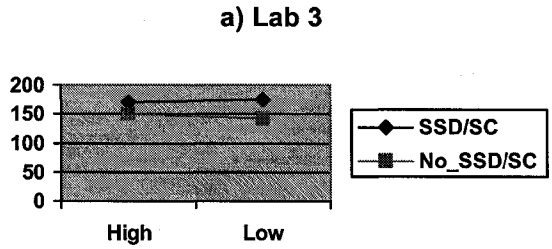

Figure 12 Graph of means (Method & Ability) for the "Time obtaining Domain Model" feature

Table 54 and Table 55 show a significant main effect of "Method" when considering the "Method" and "System" factors ("Prob>F"=0.0381). That significant result favors subjects not using the artifacts. No significant main effect was detected for the others factors. No interaction was detected.

Source	DF	Sum of Squares	Mean Square	F-Ratio	Pros > F
Method	1	2265.8370	2265.8370	5.11	**0.0381**
System	1	1115.5513	1115.5513	2.51	0.1323
Method * System	1	84.6942	84.6942	0.19	0.6680
Residual	1	7097.2833	443.5802		
Total	1	10660.2000			

Table 54 Two-Way ANOVA (Method & System) analysis at Lab-3 during the summer06 experiment for the "Time in lab" feature (Experiment III)

		System					
		CPD		VS		All Systems	
		Mean	Size	Mean	Size	Mean	Size
Method	No_SSD/SC	154.33	6	143.40	5	149.36	11
	SSD/SC	180.00	5	160.75	5	171.44	10
	All Methods	166.00	11	151.11	10		

Table 55 Descriptive statistic for the test reported in Table 54

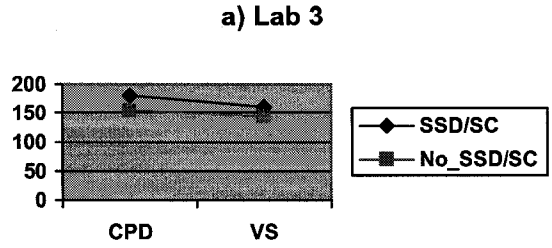

Figure 13 Graph of means (Method & System & Ability) for the "Time obtaining Domain Model" feature

4.3.4 Experiment IV (Fall 2006)

A Mixed Repeated Measures ANOVA test was used to analyze the "Correctness" dependent variable (section 4.3.4.1) and a Three-Way ANOVA test was used to analyze the "Time obtaining Domain Model" and "Time in lab" dependent variables (section 4.3.4.2 and 4.3.4.3, respectively).

4.3.4.1 Correctness

Table 56 and Table 57 show the results of applying a Mixed Repeated Measures ANOVA test for the "Averaged Correctness" dependent variable. A significant main effect is observed for the "Method" factor only ("Prob>F"=0.0048). This significant main effect favors subjects using the artifacts to obtain the Domain Model. No significant interaction between the factors was detected. Notice that the "Method" and "System" factors are close to show a significant interaction ("Prob>F"=0.0676). What is interesting about this issue is the fact that once more, similarly to what we observed for Experiments II and III, subjects did better when dealing with the CPD system. For this experiment the questionnaire for Lab-4 included questions regarding the perceived systems' level of complexity (Appendix E, Table 179, Page 120): There was a trend to consider the VS system more complex than the CPD system.

Source	DF	F-Ratio	Prob > F
Method	1	8.27	**0.0048**
System	1	0.34	0.5628
Method * System	1	3.40	0.0676
Ability	1	2.66	0.1057
Method * Ability	1	0.01	0.9231
System * Ability	1	0.18	0.6719
Method * System * Ability	1	3.46	0.0655

Table 56 Mixed Model Repeated Measures ANOVA (Method, System & Ability) analysis for "Averaged Correctness" feature (Experiment IV)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.38	14	0.44	16	0.41	30
		SSD/SC	0.33	15	0.38	16	0.35	31
		All Methods	0.35	29	0.41	32		
	Lab 4	No_SSD/SC	0.44	15	0.36	14	0.40	29
		SSD/SC	0.36	16	0.37	15	0.36	31
		All Methods	0.40	31	0.36	29		

Table 57 Descriptive statistic for the test reported in Table 56

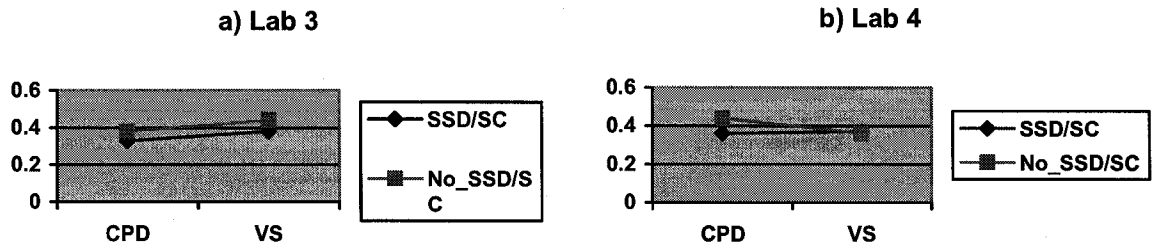

Figure 14 Graph of means for "Averaged Correctness" feature

4.3.4.2 Time obtaining Domain Model

A Three-Way ANOVA test for the "Time obtaining Domain Model" does not show any significant result (Table 58 and Table 59): no significant main effect, and no significant interaction between factors. However, although the "System" factor is not even close to show a significant main effect, similarly to what happened for the previous two experiments, it seems that subjects dealing with the VS system required less time to obtain the Domain Model.

Source	DF	F-Ratio	Prob > F
Method	1	0.81	0.3734
System	1	0.61	0.4397
Method * System	1	0.97	0.3288
Ability	1	1.28	0.2633
Method * Ability	1	0.02	0.8878
System * Ability	1	0.08	0.7799
Method * System * Ability	1	0.05	0.8316

Table 58 Three-Way ANOVA (Method, System & Ability) analysis for the "Time obtaining Domain Model" feature (Experiment IV)

		System					
		CPD		VS		All Systems	
		Mean	Size	Mean	Size	Mean	Size
Method	No_SSD/SC	127.81	14	127.91	16	127.86	30
	SSD/SC	128.47	15	113.69	16	120.84	31
	All Methods	128.15	29	120.80	32		

Table 59 Descriptive statistic for the test reported in Table 58

a) Lab 3

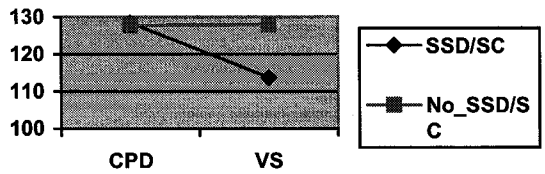

Figure 15 Graph of means (Method & System & Ability) for the "Time obtaining Domain Model" feature

4.3.4.3 Time in lab

Table 60 and Table 61 show the results of applying a Mixed Repeated Measures ANOVA test for the "Time in lab" dependent variable. "Method" is the only factor with a significant main effect ("Prob>F"=0.0371). No other main effect or interaction is observed.

Source	DF	F-Ratio	Prob > F
Method	1	4.58	**0.0371**
System	1	0.45	0.5070
Method * System	1	0.48	0.4899
Ability	1	0.10	0.7584
Method * Ability	1	0.83	0.3672
System * Ability	1	1.24	0.2714
Method * System * Ability	1	0.37	0.5435

Table 60 Mixed Model Repeated Measures ANOVA (Method, System & Ability) analysis for "Time in lab" feature (Experiment IV)

		System					
		CPD		VS		All Systems	
		Mean	Size	Mean	Size	Mean	Size
Method	No_SSD/SC	155.43	14	156.81	16	156.17	30
	SSD/SC	172.60	15	168.56	16	170.52	31
	All Methods	164.31	29	162.69	32		

Table 61 Descriptive statistic for the test reported in Table 60

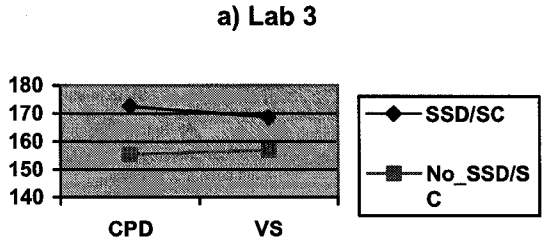

Figure 16 Graph of means (Method & System & Ability) for "Time in lab" feature

4.3.5 Summary

The goal of the multivariate analysis was to provide information about some extraneous factors that could have been potentially acting as confounding or interaction factors for the study. "System" and "Ability" were the two factors considered. Results specific to the "Method" factor are not commented as they were previously explained in Section 4.2.3.5.

Overall, we did not find any significant relationship between the "Ability" factor and the two dependent variables. On the other hand, the "System" factor seems to be a variable yielding visible results. Its influence was noted when analyzing the "Averaged Correctness" variable for the last three experiments (results were close to show a significant difference). On average, subjects performed better

when dealing with the CPD system. This is in accordance with what we observed when the subjects' level of system understanding was analyzed through a One-Sample *t*-test (see section 4.2.1). That test showed that subjects seemed to feel more comfortable when dealing with the CPD system than when dealing with the VS system. That result is also supported by questions asked in questionnaires during Lab-4 of Experiments III and IV. There was a trend to consider the VS system more complex that the CPD system.

The "System" factor influence was also noticed when analyzing the "Time obtaining Domain Model" dependent variable. For the last three experiments we observed that subjects dealing with the VS system spent less time to draw their Domain Model than subjects dealing with the CPD system. This result could be explained by the fact that VS system had fewer elements or concepts to be modeled.

It is recommended, for future experiments where two software systems are used as experiment artifacts, to make them as equivalents as possible so as to remove any possible confounded effect of this factor.

In general, not significant interaction effects were found among the three factors.

4.4. Questionnaires analysis

Subjects were supposed to fill in a questionnaire at the end of each lab: see the description of the questionnaires in section 3.2.5, and the actual questionnaires in Appendix F. The results obtained after statistically analyzing those data have been used to support some of the conclusions reached in the previous sections when analyzing the Domain Models obtained by the subjects. In this section the significant results are discussed, but without referring again to conclusions drawn in previous sections.

The questionnaires were analyzed using One-Side Two-Sample *t*-test. Essentially, differences in responses were compared by considering the two confounding factors already mentioned, i.e., the subject's "Ability" and the "System" being used. The "Method" factor was used when possible, that is, when dealing with questions that were answered by subjects dealing and not dealing with the artifacts (i.e., not on questions that are specific to the use of the artifacts). We considered the "Ability" factor during the analysis of all the questionnaires, although our analysis of confounding variables in section 4.3 did not show in general a significant main effect of this factor. Detailed statistical results of questionnaires are organized in tables in Appendix E.

4.4.1 Experiment I (Summer 2005)

Table 162 (Page 113) shows the results of applying a Two-Sample *t*-test for questionnaire in Lab-1. Mean values related to the questions that inquire about the subjects' background are not that good in general, suggesting that, unfortunately, the subjects did not seem to be very familiar with SSD and SOC. Four significant differences were detected. It seems that high-ability subjects have a better background in UML and understanding of OCL than low-ability subjects ("p-value" of 0.004 and 0.05, respectively). The other two significant differences point out how much easier it was for the high-ability subjects to deal with the VS system ("p-value" of 0.0324 and 0.05) than low-ability subjects.

Table 163 (Page113) shows only one significant difference in Lab-2 ("p-value"=0.0033). Subjects defining SSD and SOC for the CPD system complained significantly much more than subjects working with VS about the lack of time to complete the task. Actually, the CPD system documentation contains more use cases than the VS system documentation. Once again, a high negative score is associated with how comfortable subjects feel when defining SSD and SOC.

Table 164 and Table 165 (Page 114) show the analysis of questionnaires for Lab-3. A couple of significant differences show that it was harder for low-ability subjects to understand the lab's tasks ("p-value"=0.0270). Similarly, low-ability subjects found the usage of OCL expressions less useful to detect possible attributes to be modeled in classes than high-ability subjects. Again, the subjects were not comfortable with defining constraints. Finally, subjects using the artifacts complained more than the others about the lack of time to complete the lab tasks.

Table 166 and Table 167 (Page 115) show the analysis of questionnaires for Lab-4. One of the two significant differences express that low-ability subjects feel much less confident when defining Domain Models than high-ability subjects ("p-value"=0.0077). The other significant difference shows that subjects using the artifacts complained much more than the other subjects about the lack of time to finish the lab's tasks.

4.4.2 Experiment II (Fall 2005)

Table 168 and Table 169 (page 116) show the results for Lab-1 and Lab-2 respectively. No significant differences were detected in any of them. Once more, we notice the high mean value computed in Lab-1 for questions related to the subjects' understanding of the artifacts being used ("Familiar with OCL

syntax" and "Confident using OCL to write contracts"). That could be a drawback to measure the impact of using the artifacts to obtain the Domain Model.

Table 170 and Table 171 (page 116) show the results for Lab-3. Only one significant difference was detected, which expresses that those subjects who used the artifacts felt much more comfortable than those not using them while defining the Domain Model ("p-value"=0.031).

Table 172 and Table 173 (Page 117) show the results for Lab-4. Many significant differences were identified. High-ability subjects expressed stronger than low-ability subjects that their Domain Model at Lab-3 would have been better if they would have been provided with the corresponding artifacts ("p-value"=0.0452). It was said "artifacts" (plural), even though the significant difference was only associated to SOC, because similar question associated to SSD was close (only) to revealing a significant difference ("p-value"=0.0646). In general those subjects using the artifacts were complaining much more than those not using them in different ways. They felt that they did not have enough time to complete the lab tasks ("p-value"<0.0001) and that the lab instructions were not clear enough ("p-value"=0.0042). Additionally, they did not feel comfortable defining the Domain Model ("p-value"=0.0068), maybe because they had to deal with more issues than in their previous lab. Perhaps those three previous concerns could have affected those subjects' performance.

4.4.3 Experiment III (Summer 2006)

Table 174 (Page 118) shows the results of applying Two-Sample *t*-test for Lab-1. Many significant differences were computed between high-ability subjects and low-ability subjects. All of them favored low-ability subjects. In general, low-ability subjects felt more comfortable while working with both the VS and the CPD systems than high-ability subjects. For this experiment, the subjects seemed to be more familiar with the artifacts than in previous experiments.

For Lab-2, no significant differences were detected when analyzing the questionnaire (Table 175, Page 118). Again, it is noticeable through the mean values for the last two questions that subjects seemed to be more confident regarding the use of SSD and SOC than in previous experiments.

Table 176 and Table 177 (Page 119) show the results for Lab-3. Similarly to previous Experiments, subjects who used the artifacts strongly complained about the lack of time to complete the lab tasks ("p-value"=0.048). This also happened for Lab-4 (Table 178 and Table 179, Page 120) ("p-

value"=0.0223). In addition, two other significant differences were detected for Lab-4. Subjects using the VS system complained more than the one using the CPD system about the clarity of lab instructions ("p-value"=0.0401). High-ability subjects were also complaining about the time constraint to finish the lab tasks. Notice than in general subjects considered the VS system more complex than the CPD.

4.4.4 Experiment IV (Fall 2006)

Table 180 shows the results for Lab-1. Only one significant difference was found ("p-value"=0.0171), suggesting that high-ability subjects felt more familiar with OCL expressions than low-ability subjects. Notice once again that students' background on those artifacts used for the experiment is better than the one computed for previous experiments. This parameter increased from lab to lab.

For Lab-2, only one significant difference was detected (Table 181, Page 121). Subjects using the CPD system complained more than subjects using the VS system about not having enough time to obtain the System Operation Contracts ("p-value"=0.0121). Recall that the CPD system had more use cases, and therefore more functionalities, than the VS system. Subjects also complained about the lack of time, but this was not significant, as opposed to previous experiments.

Table 182 (Page 121) shows the results for Lab-3, where only one significant difference was found ("p-value"=0.0003). As in previous experiments, subjects using the artifacts complained more that the other subjects about the lack of time to complete the lab tasks. However, they obtained a better Domain Model, which we already described, which could be explained because of a better understanding of how to deal with the artifacts.

5 Conclusions and recommendations

The Unified Process (UP) is an iterative and incremental software development process that is widely used by the software engineering community. It provides a set of software development principles and good practices whose goal is to transform a set of customer requirements into a high-quality software system, delivered on time and within budget. Within the Unified Process many artifacts have been defined. System Sequence Diagrams and System Operation Contracts are two of those artifacts. Both of them are extensions proposed by Larman [1] as means of improving the quality of the Domain Model — another artifact obtained during the analysis phase of a software system lifecycle.

Empirical studies are one important way to evaluate the efficiency of any method, technique or tool. This is particularly the case in the Software Engineering context, where there exists so many ways to perform the same tasks, for example object-oriented analysis and design, and no theoretical foundations that can help us decide what best practice to adopt in a given context. That is why, in this thesis we designed a controlled experiment which goal was to analyze the impact of obtaining System Sequence Diagrams and System Operations Contracts as a way to improve the quality of the Domain Model and the effort invested in obtaining this Domain Model during the analysis phase.

Four trials of the experiment were carried out (referred to as Experiment I, Experiment II, Experiment III and Experiment IV) with 4^{th} year students registered in a Software Engineering or Computer System Engineering bachelor program. The experiment can be referred to as a Repeated Measures study, where the same subjects are measured twice. One independent variable ("Method") and two dependent variables ("Domain Model Correctness" and "Time obtaining Domain Model") were considered. The experiment design also took into consideration possible confounding factors as a way to ensure that results are valid. Subjects' ability to deal with UML artifacts as well as the specific software systems being used for the study were the two factors we controlled.

According to the statistical analysis results, both artifacts under study, System Sequence Diagrams and System Operation Constraints, do make an important contribution to the quality of the Domain Model. For the last three experiments, subjects dealing with the artifact obtained in general a better Domain Model than those not dealing with the artifacts. That difference was, however, only significant for Experiment IV. This could be explained by the fact that subjects for this experiment were better trained

at the tasks (i.e., using those artifacts when building a Domain Model), and perhaps more motivated with the lab tasks as they better understood the benefits and objectives. The lack of time to complete the lab tasks could be another issue conspiring against detecting a significant difference. Indeed, subjects using the artifacts complained about not having enough time to finish the lab tasks. Software Engineering activities like producing a Domain Model are not straight forward, and among others things, require constant reviews as a way of revealing errors, misunderstanding and low quality solutions. Maybe three hour was not enough for the tasks at hand. It has to be pointed out that it is highly probable that the subjects' level of knowledge on how to use the artifacts plays a fundamental role in finding significant differences.

In addition, we think that the lack of significance observed for Experiment II and Experiment III does not necessarily mean that the result of the experiments is a negative one (i.e., that the artifacts do not help). Indeed, Larman mentions in [1] that those two artifacts are recommended for systems with a high level of complexity. Although the students' performance was not the expected one, and the students felt that the task was challenging (because of the system complexity or time pressure), the systems we used for these experiments were relatively easy to understand and not of high complexity.

One interesting result, consistent across the last three experiments, is that, although it took significantly more time to perform the tasks when using the artifacts (more documents to read and understand, more documents to produce), the artifacts helped the students building the Domain Model faster. Whether this time reduction has a practical impact would have to be studied in the future. We suggest to provide the subjects with additional training for future experiments in order to guarantee that they know how to properly deal with the artifacts under study.

Nowadays the approach of working in pairs is very popular, particularly in Agile Programming to develop small and middle-sized applications. Perhaps, a similar experiment could be implemented by using this tactic of working in pairs. In that case the only problematic issue could be the way of creating pairs in such a way that there are comparable in terms of ability for instance. This last concern could be solved by applying what is suggested by Extreme Programming practitioners for creating pairs composed of an experienced professional and a novice one. Transposing this strategy to this experiment, the solution would be pairs of high ability and low ability subjects.

Another suggestion that could provide much more realistic results is to run the same experiment, but in an industrial environment, where developers have the required skills and substantial practical experience. A couple of issues would have to be addressed though. First, it is usually harder to control such an experiment. Second, finding enough personnel to carry out the experiment could be also a difficult task, especially if subjects work in pairs. Remember that the more subjects, the more powerful the experiment.

6 References

[1] Larman, Craig: Applying UML and Patterns: An introduction to Object-Oriented Analysis and design and the Unified Process. Second edition. Prentice Hall. 2002.

[2] Bruegge , Bernd and Dutoit, Allen H.: Object-Oriented software engineering: Conquering complex and changing systems. Prentice Hall, 2000.

[3] Aksit, Mehmet; Van Den Berg, Klass and Van Den Broek, Pim: Modelling the object-oriented software process. Enschede: Telematica Instituut. 2000.

[4] Hruby, P: The Object-Oriented Model for a Development Process. ECOOP'97 Workshops Proceedings, Finland, 1998, p 303-6

[5] Heldal, Rogardt and Johannisson, Kristofer: Relating informal and formal Contracts using Domain Models. 4th KeY Symposium. Sweden. 2005.

[6] Briand, Lionel C.; Morasca, Sandro; Basili, Victor R.: An operational process for goal-driven definition of measures. IEEE Transactions on Software Engineering. Vol 28, No 12. 2002.

[7] Wohlin, Claes; Runeson, Per; Höst, Martin; Ohlsson, Magnus C.; Regnell, Björn and Wesslénn Anders: Experimentation in software engineering: An introduction. Kluwer Academic. 2000.

[8] Caulcutt, Roland: Statistic in research and development. Second edition. Chapman & Hall. 1991.

[9] Van Soligen, Rini and Berghout, Egon: Improvement by goal-oriented measurement. Proceedings of the European Software Engineering Process Group conference (E-SEPG), Amsterdam. 1997.

[10] Briand, Lionel C.; Labiche, Y.; Di Penta, M.; Yan-Bondoe, H. D.: An experimental investigation of formality in UML-based development. IEEE Transactions on Software Engineering. Vol 31, No 10. 2005.

[11] Perry, Dewayne E.; Porter, Adam A. and Votta, Lawrence G.: Empirical Studies of Software Engineering: A Roadmap. Proceedings of conference on The future of Software Engineering, Ireland. 2000.

[12] Mathiassen, Lars; Munk-Madsen, Andreas; Nielsen, Peter A. and Stage Jan: Object Oriented Analysis & Design. Marko Publishing. 2000.

[13] Devore, Jay L.: Probability and Statistic for Engineering and the science. Third Edition. Duxbury Press. 2003.

[14] Feise, Ronald J.: Do multiple outcome measures require p-value adjustment. BMC Med Res Methodol. Vol 2. 2002

[15] Tamhane, Ajit C. and Logan, Brent R.: Multiple endpoints: an overview and new developments. Electronic journal, http://www.biostat.mcw.edu/tech/tr043.pdf. Last update September 2003. Last access December 2005.

[16] Cody, Ronald P. and Smith, Jeffrey K.: Applied statistics and the SAS programming language. 4th edition. Prentice-Hall. 1997.

[17] Jones, Byron and Kenward, Michael G.: Design and analysis of cross-over trials. Second edition. Chapman & Hall/CRC. 2003.

[18] Hopkins, Will G.: A new view of statistic. Sport Science electronic journal, http://www.sportsci.org/resource/stats/contents.html. Last update June 2001. Last access May 2005.

[19] Wolfinger, Russ and Chang, Ming: Comparing the SAS GLM and MIXED procedures for Repeated Measures. SAS Users Group International (SUGI) Proceedings. Florida. 1995.

[20] Moser, E. Barry: Repeated Measures Modeling with PROC MIXED. SAS Users Group International (SUGI) Proceedings. Canada. 2004.

[21] Marciniak, John J.: Encyclopedia of Softaware Engineering. Volume I. Second edition. John Wiley & Sons, Inc. 2002.

[22] Oppenheim, A. N.: Questionnaire design, interviewing and attitude measurement. Pinter Publishers. 1996

[23] Dong, Bin: The impact of UML documentation on software maintenance. An experimental evaluation. Thesis document of Master in Applied Science (Carleton Univertsity). 2005.

[24] SAS Intitute Inc.: JMP 5.1 User documentation

[25] SAS Intitute Inc.: SAS 9.1.3 User documentation

[26] Cockburn, Alistair : Writing effective Use Cases. Addison Wesley. 2001.

[27] Devore, Jay; Farnum Nicholas: Applied statistic for engineers and scientists. Brooks/Cole Publishing Company. 1999.

[28] Cowan, Glen: Statistical data analysis. Clarendon Press. 1998.

[29] Dallal, Gerard E.: Repeated Measures Analysis of Variance: Before SAS's Mixed Procedure. http://www.tuftd.edu/~gdallal/Repeat.htm. Last update May 2002. Last access September 2005.

[30] Kowalchuk, Rhonda K; Keselman, H.J.; Algina, James and Wolfinger, Russell D.: The analysis of Repeated Measurements with mixed-model adjusted F Test. Sage Publications. 2004.

[31] Steven, Perdita and Pooley, Rob: Using UML software Engineering with object and components. Updated edition Addison-Wesley. 2000.

[32] Bennett, Simon; Skelton, John; Lunn, Kenn: UML. Second edition. McGraw-Hill. 2005.

[33] Kichenham, Barbara A.; PFleeger, Shari L.; Pckard, Lesly M.; Jones, Peter W.; Hoaglin, David C.; Emam, Khaled E.; Rosenberg, Jarrett: Preliminary guidelines for empirical research in Software Enginnering. IEEE Transactions on Software Engineering, VOL. 28, NO. 8, August 2002.

[34] Milliken, George A. and Johnson, Dallas E.: Analysis of messy data. Volume I: Designed experiments. Wadsworth Inc. 1984.

[35] Dallal Gerard E.: The Computer-Aided Analysis of Crossover Studies. http://www.tufts.edu/~gdallal/crossovr.htm. Last update August 2001. Last access May 2005.

[36] Arlow, Jim and Neustadt, Ila: UML and the Unified Process. Pearson Education Limited. 2002.

[37] Bushnell, William and Steiner, Martin: Use of Proc Mixed in the Analysis of Repeated Measures Data from a Clinical Trial in Obsessive Compulsive Disorder. SAS Users Group International (SUGI) Proceedings. California. 1997.

[38] Davis, Charles S: Statistical Methods for the analysis of repeated measurements. Springer-Verlag New York, Inc. 2002

[39] Giden, Ellen R.: ANOVA Repeated Measures. SAGE Publications. 1992.

[40] Khattree, Ravi: Repeated Measures Data and SAS: What SAS does and does not do. http://www.misug.org/mifolder/RKhattree_Repeated_Measures.pdf#search=%22Ravi%20Khattree%20REPEATED%20MEASURES%20DATA%20AND%20SAS%22. Last update November 2003. Last access December 2005.

[41] Sheskin, David J. Handbook of parametric and nonparametric statistical procedures. Third edition. Chapman & All/CRC . 2004.

Appendix A Statistic test results for the Summer-2005 experiment

			Lab 3			Lab 4		
Dependent variable		**Group**	**Mean**	**p-value (*t*-test)**	**p-value (Wilconxon)**	**Mean**	**p-value (*t*-test)**	**p-value (Wilconxon)**
Correctness	Missing classes	SSD/SC	0.37	0.64	0.98	0.39	0.67	0.65
		No_SSD/SC	0.35			0.37		
	Useless classes	SSD/SC	0.46	**0.01**	**0.01**	0.28	0.15	0.22
		No_SSD/SC	0.25			0.41		
	Missing relationships	SSD/SC	0.62	0.32	0.63	0.59	0.94	0.85
		No_SSD/SC	0.56			0.59		
	Wrong relationships	SSD/SC	0.11	0.32	0.28	0.08	0.21	0.20
		No_SSD/SC	0.08			0.11		
	Missing attributes	SSD/SC	0.62	0.92	0.85	0.58	0.55	0.42
		No_SSD/SC	0.61			0.55		
	Wrong attributes	SSD/SC	0.18	0.76	0.90	0.22	0.61	0.38
		No_SSD/SC	0.16			0.25		
	Averaged correctness	SSD/SC	0.37	0.23	0.26	0.36	0.48	0.31
		No_SSD/SC	0.34			0.38		
Time in lab		SSD/SC	176.9	0.43	0.42	173.3	**0.00**	**0.00**
		No_SSD/SC	174.5			153		

Table 62 Two-Sample *t*-test considering independent variable "Method" levels (Experiment I).

A.1 Simple Repeated Measures ANOVA

Effect	**DF**	**F-Ratio**	**Pros > F**
Method	1	0.22	0.6405
Method Order	1	0.04	0.8446
Attempt	1	0.39	0.5336
Covariance Structure	Toeplitz		

Table 63 Simple Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Experiment I)

Effect	DF	F-Ratio	Pros > F
Method	1	0.22	0.6443
Method Order	1	8.80	**0.0044**
Attempt	1	0.00	0.9503
Covariance Structure	Toeplitz		

Table 64 Simple Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Experiment I)

Effect	DF	F-Ratio	Pros > F
Method	1	0.29	0.5918
Method Order	1	0.48	0.4926
Attempt	1	0.04	0.8451
Covariance Structure	Toeplitz		

Table 65 Simple Repeated Measures ANOVA analysis for "Missing Relationships" sub-feature (Experiment I)

Effect	DF	F-Ratio	Pros > F
Method	1	0.11	0.7363
Method Order	1	2.51	0.1189
Attempt	1	0.28	0.5997
Covariance Structure	Toeplitz		

Table 66 Simple Repeated Measures ANOVA analysis for "Wrong Relationships" sub-feature (Experiment I)

Effect	DF	F-Ratio	Pros > F
Method	1	0.04	0.8505
Method Order	1	0.58	0.4484
Attempt	1	3.53	0.0656
Covariance Structure	Toeplitz		

Table 67 Simple Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Experiment I)

Effect	DF	F-Ratio	Pros > F
Method	1	0.04	0.8425
Method Order	1	0.21	0.6469
Attempt	1	0.94	0.3353
Covariance Structure	Toeplitz		

Table 68 Simple Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Experiment I)

A.2 Two-Way ANOVA / Mixed Repeated Measures ANOVA

Source	DF	F-Ratio	Prob > F
Method	1	0.67	0.4177
Ability	1	3.31	0.0741
Method * Ability	1	0.30	0.5882

Table 69 Mixed Model Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Method & Ability) (Experiment I)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.35	12	0.35	4	0.35	16
		SSD/SC	0.34	11	0.44	6	0.38	17
		All Methods	0.35	23	0.40	10		
	Lab 4	No_SSD/SC	0.34	10	0.43	5	0.37	15
		SSD/SC	0.37	12	0.45	4	0.39	16
		All Methods	0.36	22	0.44	9		

Table 70 Descriptive statistic for the test reported in Table 69

Source	DF	F-Ratio	Prob > F
Method	1	0.67	0.4150
Ability	1	0.41	0.5232
Method * Ability	1	0.39	0.5337

Table 71 Mixed Model Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Method & Ability) (Experiment I)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.28	12	0.18	4	0.26	16
		SSD/SC	0.42	11	0.54	6	0.46	17
		All Methods	0.34	23	0.40	10		
	Lab 4	No_SSD/SC	0.39	10	0.45	5	0.41	15
		SSD/SC	0.29	12	0.27	4	0.28	16
		All Methods	0.33	22	0.37	9		

Table 72 Descriptive statistic for the test reported in Table 71

Source	DF	F-Ratio	Prob > F
Method	1	0.35	0.5563
System	1	0.00	0.9819
Method * System	1	0.92	0.3408

Table 73 Mixed Model Repeated Measures ANOVA analysis for "Wrong Associations" sub-feature (Method & Ability) (Experiment I)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.08	12	0.10	4	0.09	16
		SSD/SC	0.11	11	0.12	6	0.11	17
		All Methods	0.09	23	0.11	10		
	Lab 4	No_SSD/SC	0.11	10	0.13	5	0.12	15
		SSD/SC	0.10	12	0.06	4	0.09	16
		All Methods	0.11	22	0.10	9		

Table 74 Descriptive statistic for the test reported in Table 73

Source	DF	F-Ratio	Prob > F
Method	1	0.40	0.5288
Ability	1	8.67	**0.0046**
Method * Ability	1	0.00	0.9586

Table 75 Mixed Model Repeated Measures ANOVA analysis for "Missing Associations" sub-feature (Method & Ability) (Experiment I)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.55	12	0.62	4	0.57	16
		SSD/SC	0.55	11	0.74	6	0.63	17
		All Methods	0.55	23	0.69	10		
	Lab 4	No_SSD/SC	0.53	10	0.72	5	0.59	15
		SSD/SC	0.59	12	0.64	4	0.60	16
		All Methods	0.56	22	0.68	9		

Table 76 Descriptive statistic for the test reported in Table 75

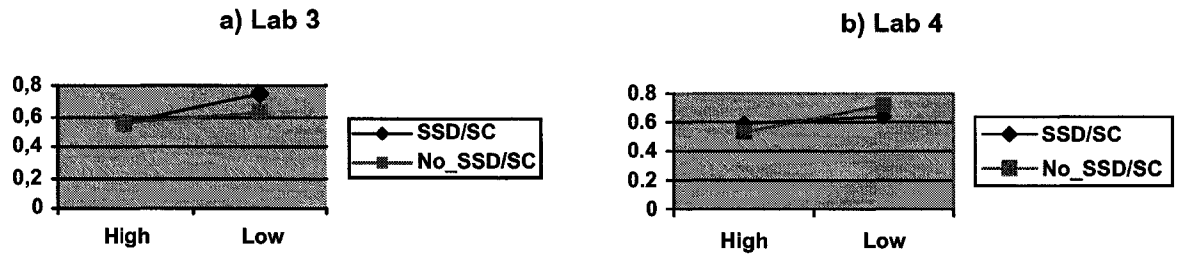

Figure 17 Graph of means for "Missing Associations" sub-feature (Method & Ability)

Source	DF	F-Ratio	Prob > F
Method	1	0.00	0.9686
Ability	1	1.42	0.2379
Method * Ability	1	0.01	0.9385

Table 77 Mixed Model Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Method & Ability) (Experiment I)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.18	12	0.11	4	0.16	16
		SSD/SC	0.19	11	0.16	6	0.18	17
		All Methods	0.19	23	0.14	10		
	Lab 4	No_SSD/SC	0.27	10	0.23	5	0.25	15
		SSD/SC	0.25	12	0.16	4	0.23	16
		All Methods	0.26	22	0.20	9		

Table 78 Descriptive statistic for test reported in Table 77

Source	DF	F-Ratio	Prob > F
Method	1	0.09	0.7594
Ability	1	1.03	0.3140
Method * Ability	1	0.03	0.8687

Table 79 Mixed Model Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Method & Ability) (Experiment I)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.62	12	0.60	4	0.62	16
		SSD/SC	0.57	11	0.71	6	0.62	17
		All Methods	0.60	23	0.67	10		
	Lab 4	No_SSD/SC	0.52	10	0.62	5	0.55	15
		SSD/SC	0.59	12	0.58	4	0.59	16
		All Methods	0.56	22	0.60	9		

Table 80 Descriptive statistic for test reported in Table 79

Source	DF	F-Ratio	Prob > F
Method	1	0.43	0.5136
System	1	0.63	0.4320
Method * System	1	0.06	0.8129

Table 81 Mixed Model Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Method & System) (Experiment I)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.37	8	0.34	8	0.35	16
		SSD/SC	0.31	8	0.45	9	0.38	17
		All Methods	0.34	16	0.39	17		
	Lab 4	No_SSD/SC	0.39	8	0.35	7	0.37	15
		SSD/SC	0.48	8	0.30	8	0.39	16
		All Methods	0.44	16	0.32	15		

Table 82 Descriptive statistic for test reported in Table 81

Source	DF	F-Ratio	Prob > F
Method	1	0.32	0.5751
System	1	9.70	**0.0029**
Method * System	1	0.90	0.3479

Table 83 Mixed Model Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Method & System) (Experiment I)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.21	8	0.30	8	0.26	16
		SSD/SC	0.36	8	0.57	9	0.46	17
		All Methods	0.28	16	0.44	17		
	Lab 4	No_SSD/SC	0.33	8	0.51	7	0.41	15
		SSD/SC	0.15	8	0.41	8	0.28	16
		All Methods	0.24	16	0.46	15		

Table 84 Descriptive statistic for test reported in Table 83

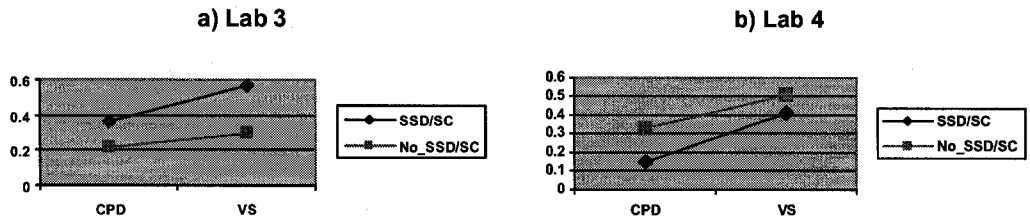

Figure 18 Graph of means for "Useless Classes" sub-feature (Method & System)

Source	DF	F-Ratio	Prob > F
Method	1	0.61	0.4392
System	1	5.02	**0.0288**
Method * System	1	0.90	0.3476

Table 85 Mixed Model Repeated Measures ANOVA analysis for "Missing Associations" sub-feature (Method & System) (Experiment I)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.63	8	0.50	8	0.57	16
		SSD/SC	0.63	8	0.63	9	0.63	17
		All Methods	0.63	16	0.56	17		
	Lab 4	No_SSD/SC	0.66	8	0.52	7	0.59	15
		SSD/SC	0.66	8	0.54	8	0.60	16
		All Methods	0.66	16	0.53	15		

Table 86 Descriptive statistic for test reported in Table 85

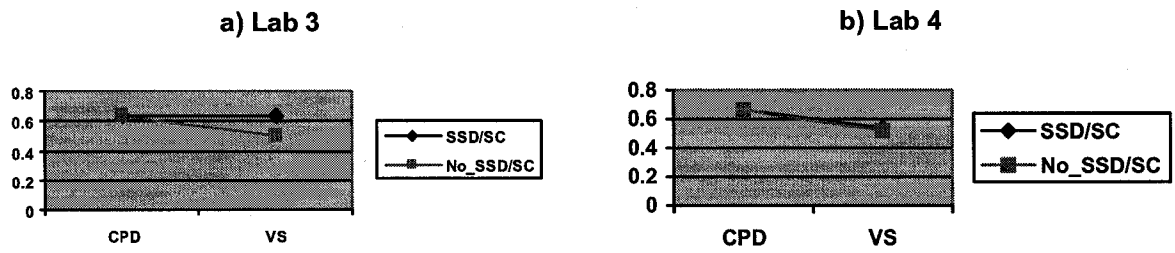

Figure 19 Graph of means (Method & System & Ability) for "Missing Associations" sub-feature

Source	DF	F-Ratio	Prob > F
Method	1	0.03	0.8588
System	1	1.97	0.1658
Method * System	1	0.71	0.4032

Table 87 Mixed Model Repeated Measures ANOVA analysis for "Wrong Associations" sub-feature (Method& System) (Experiment I)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.10	8	0.07	8	0.09	16
		SSD/SC	0.16	8	0.06	9	0.11	17
		All Methods	0.13	16	0.07	17		
	Lab 4	No_SSD/SC	0.11	8	0.13	7	0.12	15
		SSD/SC	0.09	8	0.09	8	0.09	16
		All Methods	0.10	16	0.11	15		

Table 88 Descriptive statistic for test reported in Table 87

Source	DF	F-Ratio	Prob > F
Method	1	0.00	0.9695
System	1	6.35	**0.0146**
Method * System	1	9.75	**0.0028**

Table 89 Mixed Model Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Method & System) (Experiment I)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.25	8	0.08	8	0.16	16
		SSD/SC	0.07	8	0.31	9	0.18	17
		All Methods	0.16	16	0.19	17		
	Lab 4	No_SSD/SC	0.19	8	0.33	7	0.25	15
		SSD/SC	0.15	8	0.31	8	0.23	16
		All Methods	0.17	16	0.32	15		

Table 90 Descriptive statistic for test reported in Table 89

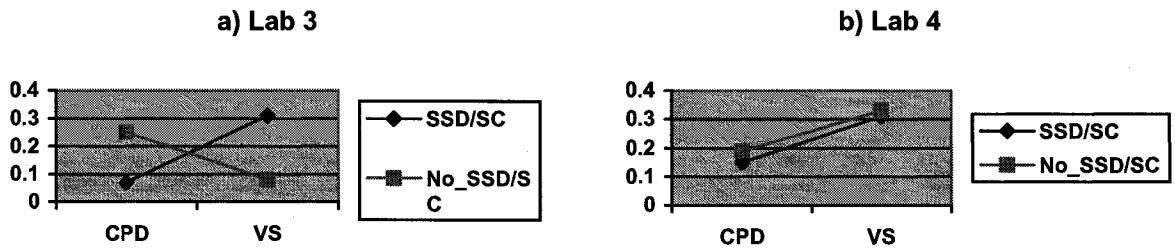

Figure 20 Graph of means for "Wrong Attributes" sub-feature (Method & System)

Source	DF	F-Ratio	Prob > F
Method	1	0.15	0.7000
System	1	10.40	**0.0021**
Method * System	1	0.48	0.4891

Table 91 Mixed Model Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Method & System) (Experiment I)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.71	8	0.53	8	0.62	16
		SSD/SC	0.66	8	0.59	9	0.62	17
		All Methods	0.68	16	0.56	17		
	Lab 4	No_SSD/SC	0.61	8	0.49	7	0.55	15
		SSD/SC	0.66	8	0.51	8	0.59	16
		All Methods	0.64	16	0.50	15		

Table 92 Descriptive statistic for test reported in Table 91

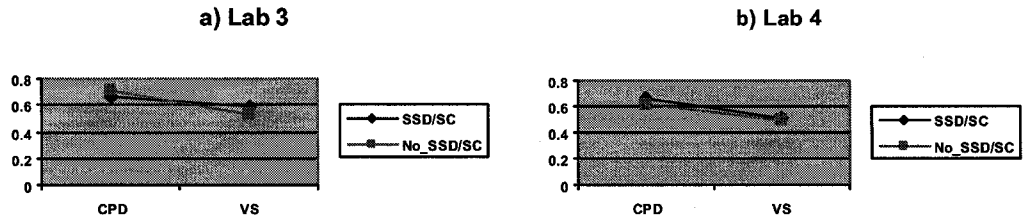

Figure 21 Graph of means for "Missing Attributes" sub-feature (Method & System)

Appendix B Statistic test results for the Fall-2005 experiment

			Lab 3			Lab 4		
Dependent variable		**Group**	**Mean**	**p-value (*t*-test)**	**p-value (Wilcoxon)**	**Mean**	**p-value (*t*-test)**	**p-value (Wilcoxon)**
Correctness	Missing classes	SSD/SC	0.4519	0.9633	0.8590	0.4612	0.8832	0.8658
		No_SSD/SC	0.4538			0.4554		
	Useless classes	SSD/SC	0.2067	0.0913	0.0264	0.2263	0.9800	0.4632
		No_SSD/SC	0.2996			0.2249		
	Missing relationships	SSD/SC	0.7113	0.6438	0.6814	0.7637	0.2524	0.3283
		No_SSD/SC	0.7279			0.716		
	Wrong relationships	SSD/SC	0.1711	0.3895	0.1697	0.19.2	0.4839	0.7297
		No_SSD/SC	0.1988			0.2078		
	Missing attributes	SSD/SC	0.7015	0.4732	0.3891	0.6982	0.5591	0.3358
		No_SSD/SC	0.725			0.6803		
	Wrong attributes	SSD/SC	0.1725	0.9538	0.3587	0.1784	0.6046	0.9785
		No_SSD/SC	0.1743			0.196		
	Averaged correctness	SSD/SC	0.4025	0.1208	0.1045	0.4197	0.7374	0.6248
		No_SSD/SC	0.4299			0.4134		
Time obtaining Domain Model		SSD/SC	108.95	0.1400	0.1696	104.12	0.7915	0.8358
		No_SSD/SC	120.79			106.51		
Time in lab		SSD/SC	159.35	0.3759	0.2097	154.56	**0.0117**	**0.0198**
		No_SSD/SC	153.94			136.63		

Table 93 Two-Sample *t*-test experiment considering the independent variable "Method" levels (Experiment II).

B.1 Simple Repeated Measures ANOVA

Effect	**DF**	**F-Ratio**	**Pros > F**
Method	57	0.00	0.9457
Method Order	57	0.02	0.8927
Attempt	57	0.04	0.8513
Covariance Structure	Toeplitz		

Table 94 Simple Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Experiment II)

Effect	**DF**	**F-Ratio**	**Pros > F**
Method	57	1.41	0.2367
Method Order	57	1.50	0.2230
Attempt	57	0.51	0.4752
Covariance Structure	Toeplitz		

Table 95 Simple Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Experiment II)

Effect	DF	F-Ratio	Pros > F
Method	57	0.32	0.5711
Method Order	57	1.38	0.2424
Attempt	57	0.55	0.4611
Covariance Structure	Toeplitz		

Table 96 Simple Repeated Measures ANOVA analysis for "Missing Relationships" sub-feature (Experiment II)

Effect	DF	F-Ratio	Pros > F
Method	57	1.26	0.2646
Method Order	57	0.06	0.8032
Attempt	57	0.49	0.4860
Covariance Structure	Toeplitz		

Table 97 Simple Repeated Measures ANOVA analysis for "Wrong Relationships" sub-feature (Experiment II)

Effect	DF	F-Ratio	Pros > F
Method	57	0.18	0.6723
Method Order	57	0.12	0.7299
Attempt	57	0.36	0.5478
Covariance Structure	Toeplitz		

Table 98 Simple Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Experiment II)

Effect	DF	F-Ratio	Pros > F
Method	57	0.02	0.9002
Method Order	57	0.86	0.3547
Attempt	57	1.16	0.2844
Covariance Structure	Toeplitz		

Table 99 Simple Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Experiment II)

B.2 Three-Way ANOVA / Mixed Repeated Measures ANOVA

Source	DF	F-Ratio	Prob > F
Method	1	0.01	0.9063
System	1	0.50	0.4827
Method * System	1	1.65	0.2020
Ability	1	4.76	**0.0312**
Method * Ability	1	0.01	0.9102
System * Ability	1	0.13	0.7186
Method * System * Ability	1	1.68	0.1979

Table 100 Mixed Model Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Method & System & Ability) (Experiment II)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.44	26	0.50	7	0.45	33
		SSD/SC	0.42	27	0.57	8	0.45	35
		All Methods	0.43	53	0.53	15		
	Lab 4	No_SSD/SC	0.44	27	0.51	8	0.46	35
		SSD/SC	0.46	26	0.48	7	0.46	33
		All Methods	0.45	53	0.49	15		

Table 101 Descriptive statistic for test reported in Table 100

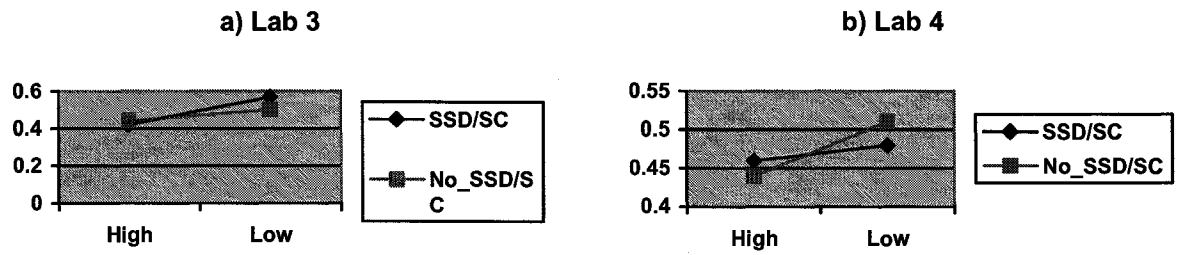

Figure 22 Graph of means for *"Missing classes"* sub-feature (Method & System & Ability)

Source	DF	F-Ratio	Prob > F
Method	1	2.09	0.1509
System	1	17.64	**<.0001**
Method * System	1	0.35	0.5555
Ability	1	0.28	0.5999
Method * Ability	1	0.88	0.3493
System * Ability	1	2.70	0.1030
Method * System * Ability	1	0.33	0.5671

Table 102 Mixed Model Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Method & System & Ability) (Experiment II)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.24	18	0.37	15	0.30	33
		SSD/SC	0.15	19	0.29	16	0.21	35
		All Methods	0.19	37	0.33	31		
	Lab 4	No_SSD/SC	0.12	16	0.32	19	0.22	35
		SSD/SC	0.16	15	0.29	18	0.23	33
		All Methods	0.14	31	0.30	37		

Table 103 Descriptive statistic for test reported in Table 102

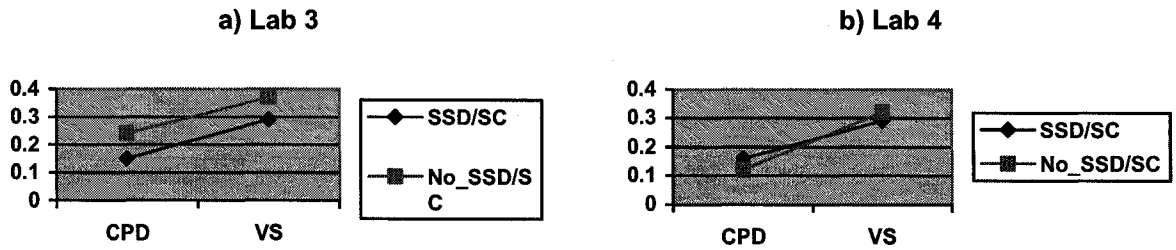

Figure 23 Graph of means for "*Missing classes*" sub-feature(Method & System & Ability)

Source	DF	F-Ratio	Prob > F
Method	1	0.32	0.5717
System	1	0.38	0.5393
Method * System	1	0.06	0.8008
Ability	1	1.63	0.2047
Method * Ability	1	0.01	0.9311
System * Ability	1	0.00	0.9739
Method * System * Ability	1	0.43	0.5152

Table 104 Mixed Model Repeated Measures ANOVA analysis for "Missing Associations" sub-feature (Method & System & Ability) (Experiment II)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.70	26	0.81	7	0.73	33
		SSD/SC	0.68	27	0.83	8	0.71	35
		All Methods	0.69	53	0.82	15		
	Lab 4	No_SSD/SC	0.72	27	0.70	8	0.72	35
		SSD/SC	0.78	26	0.72	7	0.76	33
		All Methods	0.75	53	0.71	15		

Table 105 Descriptive statistic for test reported in Table 104

Source	DF	F-Ratio	Prob > F
Method	1	0.38	0.5377
System	1	3.39	0.0683
Method * System	1	1.63	0.2049
Ability	1	0.46	0.4986
Method * Ability	1	0.18	0.6714
System * Ability	1	4.99	**0.0273**
Method * System * Ability	1	0.72	0.3989

Table 106 Mixed Model Repeated Measures ANOVA analysis for "Wrong Associations" sub-feature (Method, System & Ability) (Experiment II)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Ability	Lab 3	High	0.21	28	0.19	25	0.20	53
		Low	0.11	9	0.14	6	0.12	15
		All Abilities	0.19	37	0.18	31		
	Lab 4	High	0.19	25	0.19	28	0.19	53
		Low	0.16	6	0.29	9	0.24	15
		All Abilities	0.18	31	0.21	37		

Table 107 Descriptive statistic for test reported in Table 106

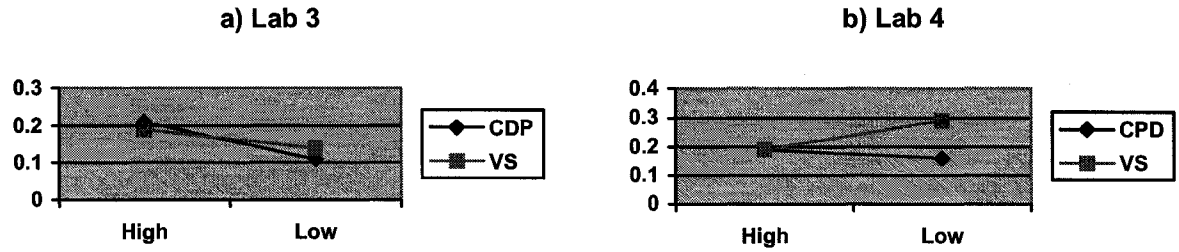

Figure 24 Graph of means for "Wrong Associations" sub-feature (Method & System & Ability)

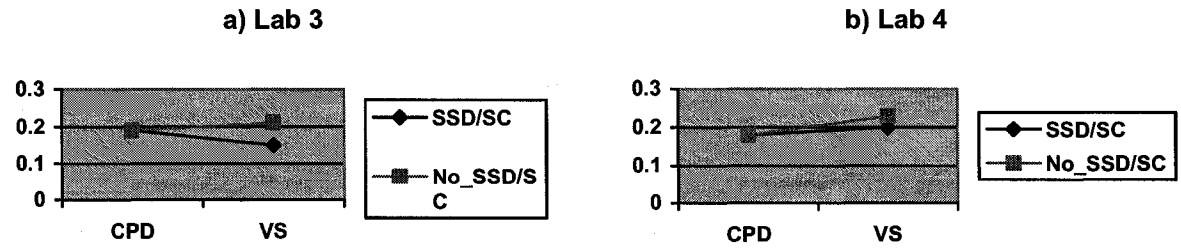

Figure 25 Graph of means for "Wrong Associations" sub-feature (Method & System & Ability) (Test was close to reveal a significant main effect for "System" factor)

Source	DF	F-Ratio	Prob > F
Method	1	0.28	0.5977
System	1	2.25	0.1361
Method * System	1	1.26	0.2639
Ability	1	1.65	0.2017
Method * Ability	1	0.13	0.7151
System * Ability	1	0.88	0.3504
Method * System * Ability	1	0.11	0.7367

Table 108 Mixed Model Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Method, System & Ability) (Experiment II)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.19	18	0.15	15	0.17	33
		SSD/SC	0.16	19	0.19	16	0.17	35
		All Methods	0.17	37	0.17	31		
	Lab 4	No_SSD/SC	0.18	16	0.21	19	0.20	35
		SSD/SC	0.14	15	0.21	18	0.18	33
		All Methods	0.16	31	0.21	37		

Table 109 Descriptive statistic for test reported in Table 108

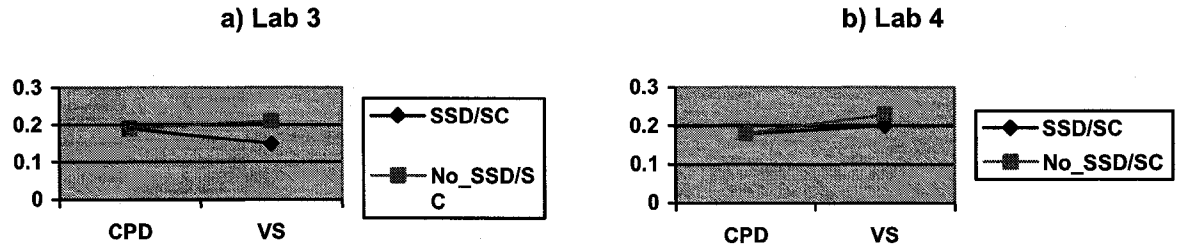

Figure 26 Graph of means for "Wrong Attributes" sub-feature (Method & System & Ability)

Source	DF	F-Ratio	Prob > F
Method	1	0.18	0.6695
System	1	3.52	0.0630
Method * System	1	0.09	0.7693
Ability	1	1.89	0.1717
Method * Ability	1	0.83	0.3628
System * Ability	1	0.02	0.8795
Method * System * Ability	1	0.35	0.5529

Table 110 Mixed Model Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Method, System & Ability) (Experiment II)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.75	18	0.70	15	0.73	33
		SSD/SC	0.72	19	0.67	16	0.70	35
		All Methods	0.74	37	0.68	31		
	Lab 4	No_SSD/SC	0.69	16	0.67	19	0.68	35
		SSD/SC	0.75	15	0.65	18	0.70	33
		All Methods	0.72	31	0.66	37		

Table 111 Descriptive statistic for test reported in Table 110

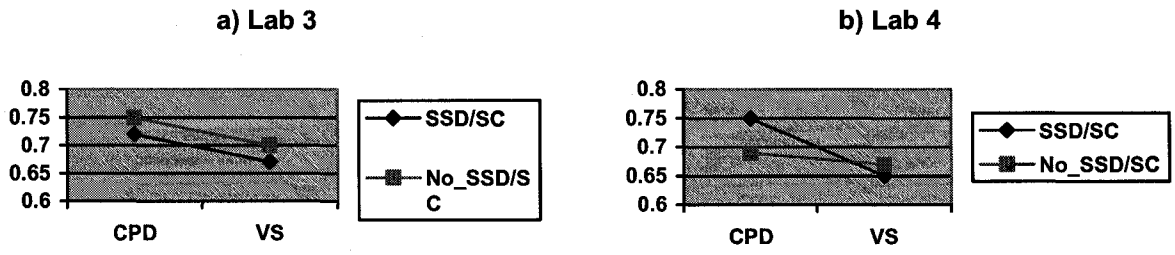

Figure 27 Graph of means for "Missing Attributes" sub-feature (Method & System & Ability)

(Test was close to reveal a significant main effect for "System" factor)

Appendix C Statistic tests results for the summer-2006 experiment

			Lab 3			Lab 4		
Dependent variable		**Group**	**Mean**	**p-value (*t*-test)**	**p-value (Wilcoxon)**	**Mean**	**p-value (*t*-test)**	**p-value (Wilcoxon)**
Correctness	Missing classes	SSD/SC	0.3236	0.2095	0.2727	0.4038	0.8041	0.8043
		No_SSD/SC	0.3866			0.3896		
	Useless classes	SSD/SC	0.2112	0.4226	0.4987	0.3946	0.6858	0.6237
		No_SSD/SC	0.3097			0.3475		
	Missing relationships	SSD/SC	0.5889	0.7980	0.6778	0.6307	0.8723	1.0000
		No_SSD/SC	0.5727			0.6219		
	Wrong relationships	SSD/SC	0.1875	0.1063	0.1430	0.1977	0.5743	0.4864
		No_SSD/SC	0.2852			0.1656		
	Missing attributes	SSD/SC	0.6181	0.4523	0.5739	0.6174	**0.0328**	0.0556
		No_SSD/SC	0.571			0.7057		
	Wrong attributes	SSD/SC	0.1875	0.8715	0.7903	0.1657	0.4345	0.4889
		No_SSD/SC	0.178			0.2031		
	Averaged correctness	SSD/SC	0.3528	0.2525	0.2684	0.3993	0.7719	0.7759
		No_SSD/SC	0.3839			0.408		
Time obtaining Domain Model		SSD/SC	108.95	0.7865	0.6804	101.5	0.3225	0.550
		No_SSD/SC	120.79			89.033		
Time in lab		SSD/SC	171.44	**0.0340**	**0.0501**	165.27	**0.0006**	**0.0079**
		No_SSD/SC	149.36			121.44		

Table 112 Two-Sample *t*-test experiment considering the independent variable "Method" levels (Experiment III).

C.1 Simple Repeated Measures ANOVA

Effect	**DF**	**F-Ratio**	**Pros > F**
Method	1	1.18	0.2920
Method Order	1	0.36	0.5535
Attempt	1	1.37	0.2576
Covariance Structure	Toeplitz		

Table 113 Simple Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Experiment III)

Effect	DF	F-Ratio	Pros > F
Method	1	0.11	0.7493
Method Order	1	0.85	0.3680
Attempt	1	1.63	0.2166
Covariance Structure	Toeplitz		

Table 114 Simple Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Experiment III)

Effect	DF	F-Ratio	Pros > F
Method	1	0.12	0.7372
Method Order	1	0.01	0.9266
Attempt	1	1.26	0.2759
Covariance Structure	Toeplitz		

Table 115 Simple Repeated Measures ANOVA analysis for "Missing Relationships" sub-feature (Experiment III)

Effect	DF	F-Ratio	Prob > F
Method	1	0.83	0.3728
Method Order	1	2.17	0.1566
Attempt	1	2.20	0.1543
Covariance Structure	Toeplitz		

Table 116 Simple Repeated Measures ANOVA analysis for "Wrong Relationships" sub-feature (Experiment III)

Effect	DF	F-Ratio	Prob > F
Method	1	0.40	0.5400
Method Order	1	3.31	0.0890
Attempt	1	2.90	0.1154
Covariance Structure	Unstructured		

Table 117 Simple Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Experiment III)

Effect	DF	F-Ratio	Prob > F
Method	1	0.13	0.7236
Method Order	1	0.27	0.6088
Attempt	1	0.00	0.9928
Covariance Structure	Toeplitz		

Table 118 Simple Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Experiment III)

C.2 Two-WAY ANOVA / Mixed Repeated Measures ANOVA

Source	DF	F-Ratio	Prob > F
Method	1	1.17	0.2878
Ability	1	0.54	0.4681
Method * Ability	1	0.36	0.5520

Table 119 Mixed Model Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Method & Ability) (Experiment III)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.38	10	0.46	1	0.39	11
		SSD/SC	0.34	8	0.26	2	0.32	10
		All Methods	0.36	18	0.33	3		
	Lab 4	No_SSD/SC	0.40	8	0.46	2	0.40	10
		SSD/SC	0.37	10	0.57	1	0.39	11
		All Methods	0.38	18	0.52	3		

Table 120 Descriptive statistic for test reported in Table 119

Source	DF	F-Ratio	Prob > F
Method	1	0.26	0.6130
Ability	1	3.52	0.0689
Method * Ability	1	0.52	0.4766

Table 121 Mixed Model Repeated Measures ANOVA analysis for "Useless classes" sub-feature (Method & Ability) (Experiment III)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.34	10	0	1	0.31	11
		SSD/SC	0.23	8	0.14	2	0.21	10
		All Methods	0.30	18	0.10	3		
	Lab 4	No_SSD/SC	0.39	8	0.08	2	0.35	10
		SSD/SC	0.41	10	0.29	1	0.39	11
		All Methods	0.40	18	0.18	3		

Table 122 Descriptive statistic for test reported in Table 121

Source	DF	F-Ratio	Prob > F
Method	1	0.34	0.5613
Ability	1	0.29	0.5958
Method * Ability	1	1.33	0.2566

Table 123 Mixed Model Repeated Measures ANOVA analysis for "Missing Associations" sub-feature (Method & Ability) (Experiment III)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.56	10	0.69	1	0.57	11
		SSD/SC	0.64	8	0.42	2	0.59	10
		All Methods	0.59	18	0.51	3		
	Lab 4	No_SSD/SC	0.61	8	0.69	2	0.62	10
		SSD/SC	0.60	10	0.90	1	0.63	11
		All Methods	0.61	18	0.79	3		

Table 124 Descriptive statistic for test reported in Table 123

Source	DF	F-Ratio	Prob > F
Method	1	0.00	0.9691
Ability	1	1.41	0.2431
Method * Ability	1	0.59	0.4468

Table 125 Mixed Model Repeated Measures ANOVA analysis for "Wrong Associations" sub-feature (Method & Ability) (Experiment III)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.30	10	0.19	1	0.29	11
		SSD/SC	0.18	8	0.21	2	0.19	10
		All Methods	0.25	18	0.20	3		
	Lab 4	No_SSD/SC	0.18	8	0.06	2	0.17	10
		SSD/SC	0.21	10	0.10	1	0.20	11
		All Methods	0.20	0.08	0.18	3		

Table 126 Descriptive statistic for test reported in Table 125

Source	DF	F-Ratio	Prob > F
Method	1	0.03	0.8715
Ability	1	2.50	0.1227
Method * Ability	1	0.41	0.5277

Table 127 Mixed Model Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Method & Ability) (Experiment III)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.18	10	0.13	1	0.18	11
		SSD/SC	0.16	8	0.27	2	0.19	10
		All Methods	0.18	18	0.22	3		
	Lab 4	No_SSD/SC	0.18	8	0.34	2	0.20	10
		SSD/SC	0.15	10	0.29	1	0.17	11
		All Methods	0.17	18	0.32	3		

Table 128 Descriptive statistic for test reported in Table 127

Source	DF	F-Ratio	Prob > F
Method	1	0.45	0.5071
Ability	1	3.04	0.0899
Method * Ability	1	0.59	0.4480

Table 129 Mixed Model Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Method & Ability) (Experiment III)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.58	10	0.47	1	0.57	11
		SSD/SC	0.67	8	0.43	2	0.62	10
		All Methods	0.62	18	0.44	3		
	Lab 4	No_SSD/SC	0.71	8	0.69	2	0.71	10
		SSD/SC	0.62	10	0.63	1	0.62	11
		All Methods	0.65	18	0.66	3		

Table 130 Descriptive statistic for test reported in Table 129

Source	DF	F-Ratio	Prob > F
Method	1	0.54	0.4661
System	1	3.18	0.0834
Method * System	1	0.70	0.4088

Table 131 Mixed Model Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Method & System) (Experiment III)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.42	6	0.34	6	0.39	11
		SSD/SC	0.35	5	0.29	5	0.32	10
		All Methods	0.39	18	0.32	3		
	Lab 4	No_SSD/SC	0.45	5	0.33	5	0.40	10
		SSD/SC	0.40	5	0.38	6	0.39	11
		All Methods	0.42	10	0.37	11		

Table 132 Descriptive statistic for test reported in Table 131

Source	DF	F-Ratio	Prob > F
Method	1	0.48	0.4915
System	1	17.84	**0.0002**
Method * System	1	4.20	**0.0481**

Table 133 Mixed Model Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Method & System) (Experiment III)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.12	6	0.54	5	0.31	11
		SSD/SC	0.12	5	0.32	5	0.21	10
		All Methods	0.12	11	0.44	10		
	Lab 4	No_SSD/SC	0.18	5	0.62	5	0.35	10
		SSD/SC	0.35	5	0.43	6	0.39	11
		All Methods	0.27	10	0.49	11		

Table 134 Descriptive statistic for test reported in Table 133

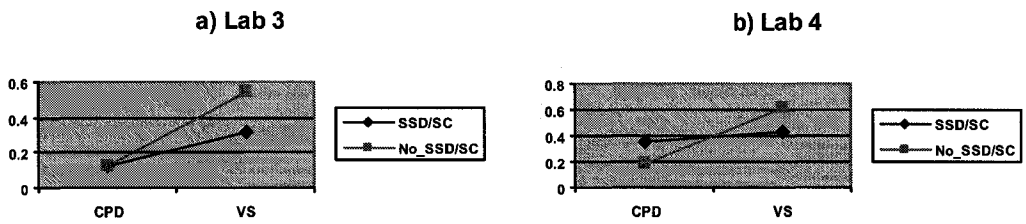

Figure 28 Graph of means for "Useless Classes" sub-feature (Method & System)

Source	DF	F-Ratio	Prob > F
Method	1	0.36	0.5545
System	1	0.22	0.6382
Method * System	1	1.85	0.1820

Table 135 Mixed Model Repeated Measures ANOVA analysis for "Missing Associations" sub-feature (Method & System) (Experiment III)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.58	6	0.56	5	0.57	11
		SSD/SC	0.60	5	0.58	5	0.59	10
		All Methods	0.59	11	0.57	10		
	Lab 4	No_SSD/SC	0.68	5	0.53	5	0.62	10
		SSD/SC	0.59	5	0.67	6	0.63	11
		All Methods	0.63	10	0.62	11		

Table 136 Descriptive statistic for test reported in Table 135

Source	DF	F-Ratio	Prob > F
Method	1	1.58	0.2172
System	1	1.15	0.2918
Method * System	1	2.97	0.0939

Table 137 Mixed Model Repeated Measures ANOVA analysis for "Wrong Associations" sub-feature (Method & System) (Experiment III)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.24	6	0.34	5	0.29	11
		SSD/SC	0.24	5	0.13	5	0.19	10
		All Methods	0.24	11	0.24	10		
	Lab 4	No_SSD/SC	0.13	5	0.23	5	0.17	10
		SSD/SC	0.18	5	0.22	6	0.20	11
		All Methods	0.15	10	0.22	11		

Table 138 Descriptive statistic for test reported in Table 137

Source	DF	F-Ratio	Prob > F
Method	1	0.28	0.6005
System	1	4.68	**0.0374**
Method * System	1	0.03	0.8633

Table 139 Mixed Model Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Method & System) (Experiment III)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.15	6	0.22	5	0.18	11
		SSD/SC	0.16	5	0.22	5	0.19	10
		All Methods	0.15	11	0.22	10		
	Lab 4	No_SSD/SC	0.18	5	0.25	5	0.20	10
		SSD/SC	0.11	5	0.22	6	0.17	11
		All Methods	0.14	10	0.23	11		

Table 140 Descriptive statistic for test reported in Table 139

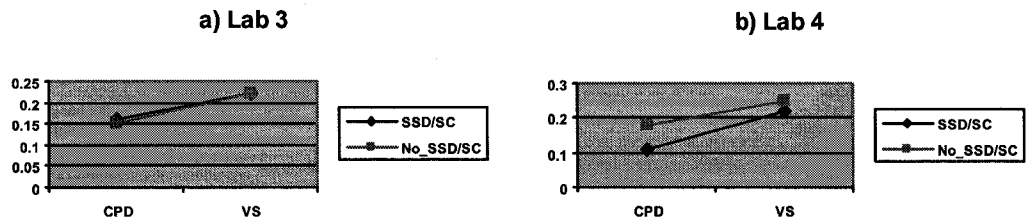

Figure 29 Graph of means for "Wrong Attributes" sub-feature (Method & System)

Source	DF	F-Ratio	Prob > F
Method	1	0.01	0.9420
System	1	10.11	**0.0031**
Method * System	1	0.22	0.64

Table 141 Mixed Model Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Method & System) (Experiment III)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.61	6	0.53	5	0.57	11
		SSD/SC	0.71	5	0.50	5	0.62	10
		All Methods	0.66	11	0.51	10		
	Lab 4	No_SSD/SC	0.74	5	0.65	5	0.71	10
		SSD/SC	0.65	5	0.59	6	0.62	11
		All Methods	0.69	10	0.61	11		

Table 142 Descriptive statistic for test reported in Table 141

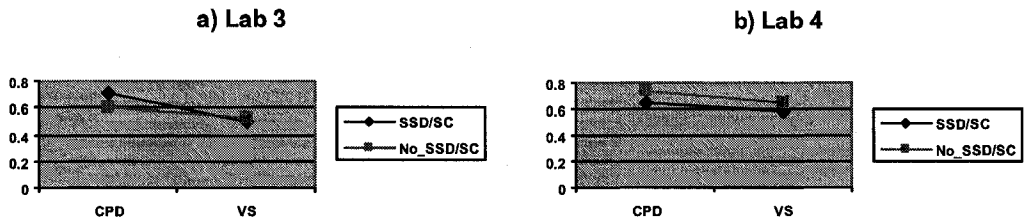

Figure 30 Graph of means for "Missing Attributes" sub-feature (Method & System)

Appendix D Statistic tests results for the Fall-2006 experiment

			Lab 3			Lab 4		
Dependent variable		**Group**	**Mean**	**p-value (*t*-test)**	**p-value (Wilcoxon)**	**Mean**	**p-value (*t*-test)**	**p-value (Wilcoxon)**
Correctness	Missing classes	SSD/SC	0.3545	*0.0689*	0.1092	0.3889	0.6424	0.9049
		No_SSD/SC	0.4293			0.3732		
	Useless classes	SSD/SC	0.2574	0.5400	0.5872	0.3162	**0.0034**	**0.0037**
		No_SSD/SC	0.2978			0.4946		
	Missing relationships	SSD/SC	0.5944	**0.0074**	**0.0124**	0.629	0.5565	0.2121
		No_SSD/SC	0.7206			0.6526		
	Wrong relationships	SSD/SC	0.1024	**0.0008**	**0.0029**	0.1371	0.2955	0.5424
		No_SSD/SC	0.1895			0.1655		
	Missing attributes	SSD/SC	0.6401	0.4840	0.7520	0.6243	0.5896	0.4708
		No_SSD/SC	0.6694			0.6437		
	Wrong attributes	SSD/SC	0.171	0.9542	0.5599	0.1462	0.5532	0.7337
		No_SSD/SC	0.1727			0.1588		
	Averaged correctness	SSD/SC	0.3533	**0.0175**	**0.0455**	0.3614	*0.0541*	**0.0386**
		No_SSD/SC	0.4132			0.4006		
Time obtaining Domain Model		SSD/SC	120.84	0.2504	0.2339	98.847	0.4327	0.5410
		No_SSD/SC	128.9			106.51		
Time in lab		SSD/SC	170.52	**0.0087**	**0.0095**	157.39	0.4948	0.2241
		No_SSD/SC	156.87			151.86		

Table 143 Two-Sample *t*-test experiment considering the independent variable "Method" levels (Experiment IV).

D.1 Simple Repeated Measures ANOVA

Effect	**DF**	**F-Ratio**	**Pros > F**
Method	1	1.62	0.2086
Method Order	1	2.45	0.1226
Time	1	0.29	0.5946
Covariance Structure	Unstructured		

Table 144 Simple Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Experiment IV)

Effect	**DF**	**F-Ratio**	**Pros > F**
Method	1	5.18	**0.0264**
Method Order	1	2.14	0.1485
Time	1	1.37	0.2466
Covariance Structure	Unstructured		

Table 145 Simple Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Experiment IV)

Effect	DF	F-Ratio	Pros > F
Method	1	8.21	**0.0058**
Method Order	1	2.09	0.1535
Time	1	0.55	0.4613
Covariance Structure	Toeplitz		

Table 146 Simple Repeated Measures ANOVA analysis for "Missing Relationships" sub-feature (Experiment IV)

Effect	DF	F-Ratio	Prob > F
Method	1	11.79	**0.0011**
Method Order	1	2.13	0.1499
Time	1	0.08	0.7721
Covariance Structure	Unstructured		

Table 147 Simple Repeated Measures ANOVA analysis for "Wrong Relationships" sub-feature (Experiment IV)

Effect	DF	F-Ratio	Prob > F
Method	1	1.08	0.3025
Method Order	1	0.01	0.9200
Time	1	0.69	0.4098
Covariance Structure	Toeplitz		

Table 148 Simple Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Experiment IV)

Effect	DF	F-Ratio	Prob > F
Method	1	0.08	0.7723
Method Order	1	0.04	0.8487
Time	1	0.97	0.3287
Covariance Structure	Unstructured		

Table 149 Simple Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Experiment IV)

D.2 Three-Way ANOVA / Mixed Repeated Measures ANOVA

Source	DF	F-Ratio	Prob > F
Method	1	1.65	0.2021
System	1	2.31	0.1314
Method * System	1	2.28	0.1336
Ability	1	4.12	**0.0446**
Method * Ability	1	0.95	0.3321
System * Ability	1	0.02	0.8978
Method * System * Ability	1	1.65	0.2011

Table 150 Mixed Model Repeated Measures ANOVA analysis for "Missing Classes" sub-feature (Method & System & Ability) (Experiment IV)

			Ability					
			High		Low		All Abilities	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.41	22	0.49	8	0.43	30
		SSD/SC	0.35	22	0.37	9	0.35	31
		All Methods	0.38	44	0.43	17		
	Lab 4	No_SSD/SC	0.34	20	0.44	9	0.37	29
		SSD/SC	0.37	22	0.43	9	0.39	31
		All Methods	0.36	42	0.44	18		

Table 151 Descriptive statistic for test reported in Table 150

Source	DF	F-Ratio	Prob > F
Method	1	5.43	**0.0215**
System	1	5.88	**0.0169**
Method * System	1	1.86	0.1750
Ability	1	0.33	0.5642
Method * Ability	1	0.10	0.7490
System * Ability	1	0.20	0.6554
Method * System * Ability	1	2.91	0.0908

Table 152 Mixed Model Repeated Measures ANOVA analysis for "Useless Classes" sub-feature (Method & System & Ability) (Experiment IV)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.16	14	0.42	16	0.30	30
		SSD/SC	0.18	15	0.33	16	0.26	31
		All Methods	0.17	29	0.38	32		
	Lab 4	No_SSD/SC	0.49	15	0.33	14	0.41	29
		SSD/SC	0.16	16	0.33	15	0.24	31
		All Methods	0.32	31	0.33	29		

Table 153 Descriptive statistic for test reported in Table 152

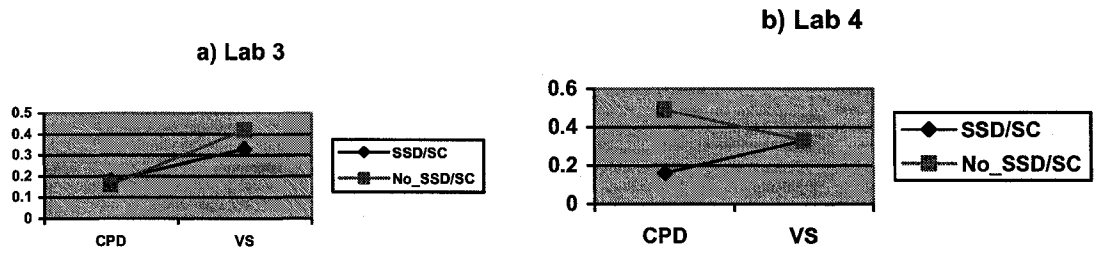

Figure 31 Graph of means for "Wrong Attributes" sub-feature (Method & System & Ability)

Source	DF	F-Ratio	Prob > F
Method	1	7.16	**0.0086**
System	1	1.73	0.1908
Method * System	1	5.33	**0.0228**
Ability	1	4.43	**0.0375**
Method * Ability	1	0.06	0.7999
System * Ability	1	0.06	0.7993
Method * System * Ability	1	1.57	0.2129

Table 154 Mixed Model Repeated Measures ANOVA analysis for "Missing Associations" sub-feature (Method & System & Ability) (Experiment IV)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.75	14	0.71	16	0.72	30
		SSD/SC	0.54	15	0.64	16	0.59	31
		All Methods	0.64	29	0.68	32		
	Lab 4	No_SSD/SC	0.74	15	0.56	14	0.65	29
		SSD/SC	0.66	16	0.60	15	0.63	31
		All Methods	0.70	31	0.58	29		

Table 155 Descriptive statistic for test reported in Table 154

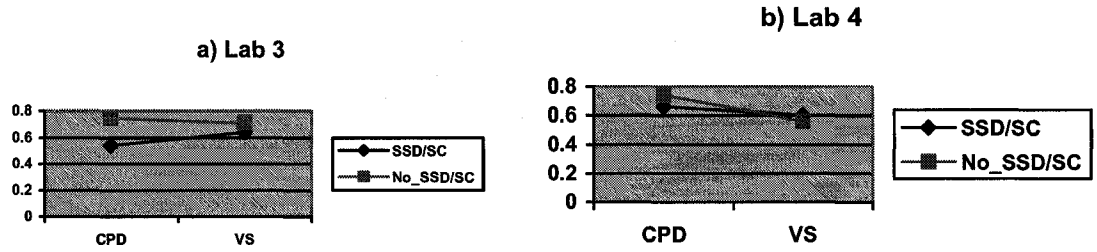

Figure 32 Graph of means (Method & System & Ability) for "Wrong Attributes" sub-feature

Source	DF	F-Ratio	Prob > F
Method	1	12.16	**0.0007**
System	1	7.49	**0.0072**
Method * System	1	17.91	**<0.0001**
Ability	1	0.10	0.7479
Method * Ability	1	0.78	0.3799
System * Ability	1	0.03	0.8556
Method * System * Ability	1	0.04	0.8327

Table 156 Mixed Model Repeated Measures ANOVA analysis for "Wrong Associations" sub-feature (Method & System & Ability) (Experiment IV)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.13	14	0.23	16	0.19	30
		SSD/SC	0.16	15	0.05	16	0.10	31
		All Methods	0.15	29	0.14	32		
	Lab 4	No_SSD/SC	0.10	15	0.24	14	0.17	29
		SSD/SC	0.11	16	0.17	15	0.14	31
		All Methods	0.10	31	0.20	29		

Table 157 Descriptive statistic for test reported in Table 156

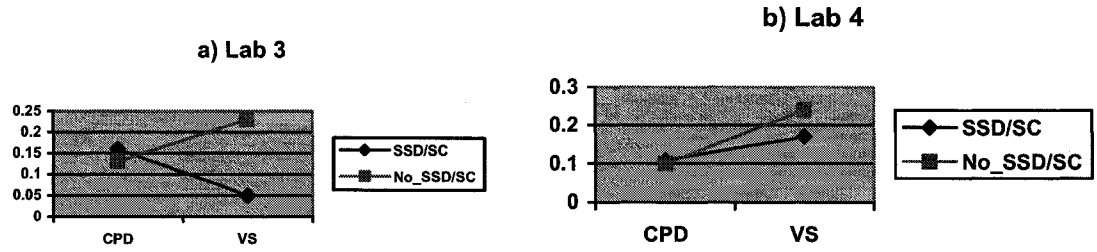

Figure 33 Graph of means for "Wrong Attributes" sub-feature (Method & System & Ability)

Source	DF	F-Ratio	Prob > F
Method	1	0.09	0.7648
System	1	9.78	**0.0022**
Method * System	1	0.97	0.3270
Ability	1	0.05	0.8218
Method * Ability	1	0.08	0.7825
System * Ability	1	1.02	0.3155
Method * System * Ability	1	0.21	0.6476

Table 158 Mixed Model Repeated Measures ANOVA analysis for "Wrong Attributes" sub-feature (Method & System & Ability) (Experiment IV)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.16	14	0.18	16	0.17	30
		SSD/SC	0.12	15	0.22	16	0.17	31
		All Methods	0.14	29	0.20	32		
	Lab 4	No_SSD/SC	0.14	15	0.18	14	0.16	29
		SSD/SC	0.12	16	0.18	15	0.15	31
		All Methods	0.13	31	0.18	29		

Table 159 Descriptive statistic for test reported in Table 158

Source	DF	F-Ratio	Prob > F
Method	1	0.86	0.3555
System	1	3.13	0.0795
Method * System	1	0.05	0.8276
Ability	1	0.36	0.5517
Method * Ability	1	0.06	0.8096
System * Ability	1	0.00	0.9628
Method * System * Ability	1	1.24	0.2682

Table 160 Mixed Model Repeated Measures ANOVA analysis for "Missing Attributes" sub-feature (Method & System & Ability) (Experiment IV)

			System					
			CPD		VS		All Systems	
			Mean	Size	Mean	Size	Mean	Size
Method	Lab 3	No_SSD/SC	0.64	14	0.69	16	0.67	30
		SSD/SC	0.63	15	0.65	16	0.64	31
		All Methods	0.64	29	0.67	32		
	Lab 4	No_SSD/SC	0.73	15	0.56	14	0.64	29
		SSD/SC	0.68	16	0.57	15	0.62	31
		All Methods	0.70	31	0.56	29		

Table 161 Descriptive statistic for test reported in Table 160

Appendix E Statistic tests results for the Questionnaires

E.1 EXPERIMENT I

Question	Ability mean		p-value
	High	Low	
Good background in UML	2.087	2.9	**0.004**
Familiar with OCL syntax	2.6957	3.3	**0.05**
Previous experience with SSD/SOC	3.8261	3.3	0.2599
Confident using OCL to write contracts	3.1739	3.6	0.3156
VS questions easy to understand	1.7391	2.5	**0.0324**
VS questions easy to answer	1.8261	2.5	0.0697
Confident dealing with VS	1.6957	2.3	**0.05**
CPD questions easy to understand	1.913	2.3	0.1641
CPD questions easy to answer	1.8696	2.1	0.49
Confident dealing with CPD	1.8261	2.2	0.1645
VS more complex that CPD	3.4783	3.2	0.4227

Table 162 Two-Sample *t*-test considering dependent variable "Ability" during Summer/2005 experiment (questionnaire at Lab-1)

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	4.5	3.375	**0.0033**	3.8636	4.1	0.5931
Clear lab instructions	2.5	2.6875	0.5530	2.5455	2.7	0.6507
Comfortable defining SSD	2.125	2.3125	0.4889	2.1364	2.4	0.3657
Comfortable defining SC	2.5625	2.75	0.5304	2.5909	2.8	0.5165

Table 163 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Summer/2005 experiment (questionnaire at Lab-2)

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	3.25	2.75	0.2521	3.0455	2.9	0.7594
Clear lab instructions	2.0	2.25	0.4047	1.9091	2.6	**0.0270**
Comfortable defining Domain Model	2.375	2.375	1.0000	2.1818	2.8	0.0614
Comfortable defining constraints	3.625	3.556	0.8848	3.8182	3.167	0.1795
OCL useful detecting classes	3.125	2.888	0.5560	2.8182	3.333	0.2090
OCL useful detecting attributes	2.625	2.778	0.7443	2.2727	3.5	**0.0041**
OCL useful detecting relationships	3.0	3.0	1.0000	2.8182	3.333	0.2537

Table 164 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Summer/2005 experiment (questionnaire at Lab-3)[1]

Question	Method mean		p-value
	No_SSD/SC	SSD/SC	
Enough time to perform lab tasks	2.3333	3.5882	**0.0022**
Clear lab instructions	1.9333	2.2941	0.2271
Comfortable defining Domain Model	2.4000	2.3529	0.8817
Comfortable defining constraints			
OCL useful detecting classes			
OCL useful detecting attributes			
OCL useful detecting relationships			

Table 165 Two-Sample *t*-test considering independent variable "Method" during Summer/2005 experiment (questionnaire at Lab-3)[1]

[1] For this test the last four questions were not considered during the test because they were answered exclusively by those subjects that used the artifacts.

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	2.5294	2.3333	0.6566	2.2174	3.000	0.1026
Clear lab instructions	1.8824	2.1333	0.4650	1.8261	2.444	0.0985
Comfortable defining Domain Model	2.3529	2.3333	0.9523	2.087	3.000	**0.0077**
SSD would have been improved Domain Model	2.5556	2.5714	0.9672	2.3636	3.000	0.1064
SOC would have been improved Domain Model	3.0000	2.8333	0.7075	2.700	3.400	0.1119
Comfortable defining constraints	3.6250	3.0000	0.2303	3.2308	3.500	0.6674
OCL useful detecting classes	2.7500	2.7500	1.0000	2.9167	2.250	0.1410
OCL useful detecting attributes	2.6250	2.7500	0.7856	2.7500	2.500	0.6365
OCL useful detecting relationships	2.7500	2.6250	0.8018	2.7500	2.500	0.6631
Better previous Domain Model with SSD and SC	3.0000	2.8750	0.8119	3.0833	2.500	0.3281

Table 166 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Summer/2005 experiment (questionnaire at Lab-4)

Question	Method mean		p-value
	No_SSD/SC	SSD/SC	
Enough time to perform lab tasks	1.8125	3.0625	**0.0022**
Clear lab instructions	1.2636	2.125	0.4659
Comfortable defining Domain Model	2.3125	2.375	0.8483
SSD would have been improved Domain Model			
SOC would have been improved Domain Model			
Comfortable defining constraints			
OCL useful detecting classes			
OCL useful detecting attributes			
OCL useful detecting relationships			
Better previous Domain Model with SSD and SC			

Table 167 Two-Sample *t*-test considering independent variable "Method" during Summer/2005 experiment (questionnaire at Lab-4)[2]

[2] For this test the last seven questions were not considered during the test because they were answered exclusively by those subjects that used the artifacts.

E.2 EXPERIMENT II

Question	Ability mean		p-value
	High	Low	
Good background in UML	2.5741	2.4	0.5334
Familiar with OCL syntax	3.2593	3.0667	0.4855
Previous experience with SSD/SOC	3.0	3.2667	0.3907
Confident using OCL to write contracts	3.5741	3.4667	0.7139
VS questions easy to understand	2.0185	2.0	0.9454
VS questions easy to answer	2.1667	1.7333	0.1042
Confident dealing with VS	1.9815	1.6667	0.1365
CPD questions easy to understand	2.4815	2.1333	0.2825
CPD questions easy to answer	2.4074	2.1333	0.3573
Confident dealing with CPD	2.3519	1.9333	0.1190
VS more complex that CPD	3.8148	3.4667	0.2902

Table 168 Two-Sample *t*-test considering dependent variable "Ability" during Fall/2005 experiment (questionnaire at Lab-1)

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to obtain SSD	2.9706	3.0938	0.7158	2.9423	3.3571	0.3142
Enough time to obtain SC	3.0588	3.2813	0.4820	3.0385	3.6429	0.1152
Clear lab instructions	1.9706	2.4688	0.0662	2.2308	2.1429	0.7935
Comfortable defining SSD	2.2647	2.5313	0.2092	2.3846	2.4286	0.8663
Comfortable defining SC	2.9118	2.8438	0.7816	2.9038	2.7857	0.6935

Table 169 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Fall/2005 experiment (questionnaire at Lab-2)

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	2.8056	2.6552	0.6409	2.7451	2.714	0.9371
Clear lab instructions	2.3056	2.3793	0.7866	2.2745	2.571	0.3660
Comfortable defining Domain Model	2.5556	2.4643	0.6876	2.4706	2.692	0.4274
Comfortable defining constraints	3.5294	3.5333	0.9906	3.4231	4.000	0.1677
OCL useful detecting classes	2.9412	2.500	0.1934	2.72	2.833	0.7937
OCL useful detecting attributes	2.8235	2.500	0.2889	2.72	2.5	0.5697
OCL useful detecting relationships	2.8235	2.500	0.2889	2.76	2.333	0.2664

Table 170 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Fall/2005 experiment (questionnaire at Lab-3)

Question	Method mean		p-value
	No_SSD/SC	SSD/SC	
Enough time to perform lab tasks	2.4516	3.0000	0.0840
Clear lab instructions	2.3548	2.3235	0.9081
Comfortable defining Domain Model	2.7667	2.2941	**0.0331**
Comfortable defining constraints			
OCL useful detecting classes			
OCL useful detecting attributes			
OCL useful detecting relationships			

Table 171 Two-Sample *t*-test considering independent variable "Method" during Fall/2005 experiment (questionnaire at Lab-3)[1]

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	2.4138	1.8571	0.0756	2.000	2.538	0.1670
Clear lab instructions	2.0345	1.6857	0.1913	1.902	1.615	0.3873
Comfortable defining Domain Model	2.6207	1.9714	0.0095	2.333	2.000	0.2926
SSD would have been improved Domain Model	2.6000	2.8824	0.4806	2.923	2.000	0.0646
SOC would have been improved Domain Model	2.7333	2.8235	0.8163	2.961	2.000	**0.0452**
Comfortable defining constraints	3.4167	3.4211	0.9900	3.360	3.667	0.4741
OCL useful detecting classes	2.7500	3.1579	0.2015	3.080	2.667	0.2961
OCL useful detecting attributes	2.9167	2.6667	0.4829	2.917	2.167	0.0785
OCL useful detecting relationships	3.0000	3.0556	0.8494	3.000	3.167	0.6413
Better previous Domain Model with SSD and SC	2.9231	2.5625	0.3415	2.818	2.429	0.3771

Table 172 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Fall/2005 experiment (questionnaire at Lab-4)

Question	Method mean		p-value
	No_SSD/SC	SSD/SC	
Enough time to perform lab tasks	1.5152	2.7419	**<0.0001**
Clear lab instructions	1.4848	2.2258	**0.0042**
Comfortable defining Domain Model	1.9394	2.6129	0.0068
SSD would have been improved Domain Model			
SOC would have been improved Domãin Model			
Comfortable definiîg constraints			
OCL useful deþecting classes			
OCL useful dçtecting attributes			
OCL useflßl detecting relationships			
Bítter previous Domain Model with[a]SSD and SC			

Table 173 Two-Sample *t*-test considering independent variable "Method" during Fall/2005 experiment (questionnaire at Lab-4)[2]

E.3 EXPERIMENT III

Question	Ability mean		p-value
	High	Low	
Good background in UML	2.2941	2.3333	0.9112
Familiar with OCL syntax	2.4706	2.3333	0.79.87
Previous experience with SSD/SOC	3.1765	3.0000	0.8037
Confident using OCL to write contracts	2.7059	2.3333	0.4791
VS questions easy to understand	1.7059	1.0000	**0.0203**
VS questions easy to answer	1.8235	1.0000	**0.0167**
Confident dealing with VS	1.9412	1.0000	**0.0016**
CPD questions easy to understand	2.0588	1.6667	0.2766
CPD questions easy to answer	2.3529	1.6667	0.1292
Confident dealing with CPD	2.4706	1.6667	**0.0242**
VS more complex that CPD	3.9412	4.6667	0.2567

Table 174 Two-Sample *t*-test considering dependent variable "Ability" during Summer/2006 experiment (questionnaire at Lab-1)

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	2.3333	2.4545	0.8716	2.3529	2.6667	0.7638
Clear lab instructions	3.6667	3.0909	0.3951	3.2353	4.0000	0.4180
Comfortable defining SSD	1.8889	2.3636	0.2372	2.0588	2.6667	0.2790
Comfortable defining SC	2.3333	2.2727	0.8809	2.2353	2.6667	0.4405

Table 175 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Summer/2006 experiment (questionnaire at Lab-2)

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	2.4545	2.2222	0.7040	2.1765	3.3333	0.1637
Clear lab instructions	2.1818	2.0000	0.6210	2.0000	2.6667	0.1836
Comfortable defining Domain Model	1.9091	1.7778	0.7483	1.8235	2.000	0.7569
Comfortable defining constraints	3.0000	4.0000	0.1514	3.5714	3.0000	0.5190
OCL useful detecting classes	3.2000	3.0000	0.7980	2.7143	4.5000	0.0207
OCL useful detecting attributes	3.2000	2.2500	0.2149	2.5714	3.5000	0.3205
OCL useful detecting relationships	3.2000	2.7500	0.5389	2.7143	4.0000	0.1114

Table 176 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Summer/2006 experiment (questionnaire at Lab-3)

Question	Method mean		p-value
	No_SSD/SC	SSD/SC	
Enough time to perform lab tasks	1.8182	3.0000	**0.0408**
Clear lab instructions	1.9091	2.3333	0.2410
Comfortable defining Domain Model	1.9091	1.7778	0.7483
Comfortable defining constraints			
OCL useful detecting classes			
OCL useful detecting attributes			
OCL useful detecting relationships			

Table 177 Two-Sample *t*-test considering independent variable "Method" during Summer/2006 experiment (questionnaire at Lab-3) [1]

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	2.5000	2.4000	0.8808	2.6111	1.0000	**0.0002**
Clear lab instructions	1.6000	2.5000	**0.0401**	2.0556	2.0000	0.8260
Comfortable defining Domain Model	1.8000	2.0000	0.6132	1.9444	1.5000	0.4990
SSD would have been improved Domain Model	3.0000	2.0000	0.0892	2.6250	2.0000	0.5406
SOC would have been improved Domain Model	3.0000	3.000	1.0000	3.1250	2.0000	0.1419
Comfortable defining constraints	3.0000	3.4286	0.2574	3.1818	4.0000	0.2231
OCL useful detecting classes	3.0000	3.3333	0.5571	3.1000	4.0000	0.3527
OCL useful detecting attributes	2.6000	3.0000	0.4790	2.8000	3.0000	0.8402
OCL useful detecting relationships	2.8000	2.8333	0.9536	2.9000	2.0000	0.3527
Better previous Domain Model with SSD and SC	2.4000	2.8333	0.3129	2.6000	3.0000	0.5987
VS system more complex than CPD system	3.7000	3.4000	0.5375	3.5000	4.0000	0.5375
CPD system more complex than VS system	2.3000	2.6000	0.5375	2.5000	2.0000	0.5375
SSD and SC are useful artifacts to use when doing software systems analysis	2.1000	2.5000	0.1800	2.3333	2.0000	0.0549

Table 178 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Summer/2006 experiment (questionnaire at Lab-4)

Question	Method mean		p-value
	No_SSD/SC	SSD/SC	
Enough time to perform lab tasks	1.6667	3.0909	**0.0223**
Clear lab instructions	1.6667	2.3636	0.1233
Comfortable defining Domain Model	1.5556	2.1818	0.1032
SSD would have been improved Domain Model			
SOC would have been improved Domain Model			
Comfortable defining constraints			
OCL useful detecting classes			
OCL useful detecting attributes			
OCL useful detecting relationships			
Better previous Domain Model with SSD and SC			
VS system more complex than CPD system	3.2222	3.8182	0.2154
CPD system more complex than VS system	2.7778	2.1818	0.2154
SSD and SC are useful artifacts to use when doing software systems analysis	2.0000	2.5455	0.0626

Table 179 Two-Sample *t*-test considering independent variable "Method" during Summer/2006 experiment (questionnaire at Lab-4)

E.4 EXPERIMENT IV

Question	Ability mean		p-value
	High	Low	
Good background in UML	2.3953	2.1765	0.2984
Familiar with OCL syntax	2.3256	2.9412	**0.0171**
Previous experience with SSD/SOC	3.186	2.8235	0.2507
Confident using OCL to write contracts	2.907	3.2353	0.2427
VS questions easy to understand	1.9535	1.8824	0.7758
VS questions easy to answer	2.0465	2.0588	0.9560
Confident dealing with VS	2.0000	1.9412	0.7723
CPD questions easy to understand	2.1163	1.9412	0.4721
CPD questions easy to answer	2.2093	2.1765	0.8808
Confident dealing with CPD	2.186	1.9412	0.2191
VS more complex that CPD	3.6279	3.1176	0.1041

Table 180 Two-Sample *t*-test considering dependent variable "Ability" during Fall/2006 experiment (questionnaire at Lab-1)

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to obtain SSD	3.4231	2.8571	0.1318	3.0256	3.4000	0.3749
Enough time to obtain SOC	4.0000	3.1071	**0.0121**	3.5128	3.6000	0.8313
Clear lab instructions	2.3077	2.2857	0.9244	2.2821	2.3333	0.8427
Comfortable defining SSD	2.2692	2.2500	0.9348	2.1795	2.4667	0.2707
Comfortable defining SC	3.1154	2.9286	0.4808	2.9487	3.2000	0.3947

Table 181 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Fall/2006 experiment (questionnaire at Lab-2)

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	2.7586	2.3636	0.2408	2.5909	2.4706	0.7531
Clear lab instructions	1.9310	2.2424	0.2257	2.0682	2.1765	0.7111
Comfortable defining Domain Model	2.1724	2.5313	0.1492	2.3953	2.2353	0.5696
Comfortable defining constraints	3.2500	3.1176	0.7295	3.0417	3.5556	0.2264
OCL useful detecting classes	2.7500	2.5000	0.3813	2.7391	2.3333	0.1979
OCL useful detecting attributes	2.7500	2.5000	0.4446	2.7391	2.3333	0.2618
OCL useful detecting relationships	2.6875	2.5625	0.6783	2.6087	2.6667	0.8628

Table 182 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Fall/2006 experiment (questionnaire at Lab-3)

Question	Method mean		p-value
	No_SSD/SC	SSD/SC	
Enough time to perform lab tasks	1.9677	3.1290	**0.0003**
Clear lab instructions	1.9355	2.2581	0.2083
Comfortable defining Domain Model	2.4000	2.3226	0.7574
Comfortable defining constraints			
OCL useful detecting classes			
OCL useful detecting attributes			
OCL useful detecting relationships			

Table 183 Two-Sample *t*-test considering independent variable "Method" during Fall/2006 experiment (questionnaire at Lab-3)[1]

Question	System mean		p-value	Ability mean		p-value
	CPD	VS		High	Low	
Enough time to perform lab tasks	2.400	1.8621	0.0687	1.9762	2.5294	0.0905
Clear lab instructions	2.0667	1.5862	**0.0218**	1.7857	1.9412	0.5105
Comfortable defining Domain Model	2.0667	1.8276	0.2656	1.8810	2.1176	0.3186
SSD would have been improved Domain Model	2.7857	2.2667	0.0923	2.5238	2.5000	0.9464
SOC would have been improved Domain Model	2.9286	2.4667	0.2217	2.5714	3.0000	0.3127
Comfortable defining constraints	3.5625	3.6429	0.8109	3.4762	3.8889	0.2537
OCL useful detecting classes	3.1875	3.2857	0.7489	3.1429	3.4444	0.3634
OCL useful detecting attributes	2.8750	2.4286	0.1717	2.4762	3.1111	0.0706
OCL useful detecting relationships	3.0625	3.0000	0.8630	2.9048	3.3333	0.2721
Better previous Domain Model with SSD and SC	2.4118	2.2857	0.7199	2.4762	2.1000	0.3109
VS system more complex than CPD system	3.2667	3.5862	0.2550	3.5000	3.2353	0.3941
CPD system more complex than VS system	2.6667	2.1034	**0.0237**	2.3810	2.4118	0.9127
SSD and SC are useful artifacts to use when doing software systems analysis	2.3103	2.1379	0.4141	2.1429	2.4375	0.2103

Table 184 Two-Sample *t*-test considering dependent variables "Ability" and "System" during Fall/2006 experiment (questionnaire at Lab-4)

Question	Method mean		p-value
	No_SSD/SC	SSD/SC	
Enough time to perform lab tasks	1.6786	2.5484	0.0026
Clear lab instructions	1.5357	2.0962	0.0070
Comfortable defining Domain Model	1.6429	2.2258	0.0053
SSD would have been improved Domain Model			
SOC would have been improved Domain Model			
Comfortable defining constraints			
OCL useful detecting classes			
OCL useful detecting attributes			
OCL useful detecting relationships			
Better previous Domain Model with SSD and SC			
VS system more complex than CPD system	3.5357	3.3226	0.4497
CPD system more complex than VS system	2.3929	2.3871	0.9820
SSD and SC are useful artifacts to use when doing software systems analysis	2.1111	2.3226	0.3169

Table 185 Two-Sample *t*-test considering independent variable "Method" during Fall/2006 experiment (questionnaire at Lab-3)

Question	**Method mean**		**p-value**
	No_SSD/SC	**SSD/SC**	
Enough time to perform lab tasks	1.6786	2.5484	**0.0026**
Clear lab instructions	1.5357	2.0968	**0.0070**
Comfortable defining Domain Model	1.6429	2.2258	**0.0053**
SSD would have been improved Domain Model			
SOC would have been improved Domain Model			
Comfortable defining constraints			
OCL useful detecting classes			
OCL useful detecting attributes			
OCL useful detecting relationships			
Better previous Domain Model with SSD and SC			

Table 186 Two-Sample *t*-test considering independent variable "Method" during Fall/2006 experiment (questionnaire at Lab-4)[3]

[3] For this test the last seven questions were not considered during the test because they were answered exclusively by those subjects that used the artifacts.

Appendix F Questionnaires and Template documents

Questionnaire for Lab-1

Levels of agreement:

1 – Strongly agree 2 – Agree 3 – Not certain 4 – Disagree 5 – Strongly disagree

Question:

	1	2	3	4	5
1. I have a good background dealing with UML models (Use Cases, Class Diagrams, StateChart Diagrams) for real software systems	☐	☐	☐	☐	☐
2. I am familiar with the OCL syntax	☐	☐	☐	☐	☐
3. I have a previous experience dealing with System Sequence Diagrams and System Operations Contracts (without considering what it was learned in class)	☐	☐	☐	☐	☐
4. I feel confident that I can use the OCL language to write contracts for a real system	☐	☐	☐	☐	☐
5. The questions for the Video Store System were easy to understand	☐	☐	☐	☐	☐
6. The questions for the Video Store System were easy to answer	☐	☐	☐	☐	☐
7. I feel confident that I understand the Video Store System	☐	☐	☐	☐	☐
8. The questions for the Car Parts Dealer System were easy to understand	☐	☐	☐	☐	☐
9. The questions for the Car Parts Dealer System were easy to answer	☐	☐	☐	☐	☐
10. I feel confident that I understand the Car Parts Dealer System	☐	☐	☐	☐	☐
11. The Video Store System is more complex than the Car Parts Dealer System	☐	☐	☐	☐	☐

Figure 34 Questionnaire to answer at the end of Lab-1

Questionnaire for Lab-1

Levels of agreement:

1 – Strongly agree 2 – Agree 3 – Not certain 4 – Disagree 5 – Strongly disagree

Question:

	1	2	3	4	5
1. I had enough time to obtain the System Sequence Diagram	☐	☐	☐	☐	☐
2. I had enough time to obtain the System Operation Contract	☐	☐	☐	☐	☐
3. The instructions and objectives of the lab were perfectly clear to me	☐	☐	☐	☐	☐
4. I felt comfortable defining System Sequence Diagrams	☐	☐	☐	☐	☐
5. I felt comfortable defining Operation Contracts	☐	☐	☐	☐	☐

Figure 35 Questionnaire to answer at the end of Lab-2

Questionnaire for Lab-3

Levels of agreement:

1 – Strongly agree 2 – Agree 3 – Not certain 4 – Disagree 5 – Strongly disagree

Question:

	1	2	3	4	5
1. I had enough time to perform the lab tasks	☐	☐	☐	☐	☐
2. The instructions and objectives of the lab were perfectly clear to me	☐	☐	☐	☐	☐
3. I felt comfortable defining the Domain Model Class Diagram	☐	☐	☐	☐	☐
4. I felt comfortable defining OCL contracts	☐	☐	☐	☐	☐
5. OCL contracts and System Sequence Diagram were useful to refine the class diagram by					
5.1 Detecting new classes	☐	☐	☐	☐	☐
5.2 Detecting new attributes	☐	☐	☐	☐	☐
5.3 Detecting new relationships between classes	☐	☐	☐	☐	☐

6. How much time (in terms of percentage) did you spend in writing OCL contracts during this lab? ☐
7. How much time (in terms of percentage) did you spend in building the Domain Model during this lab? ☐
8. How much time (in terms of percentage) did you spend in writing the Glossary? ☐

The answer for the last three questions will be in the range from A to D, where the meaning of each letter is as follow:

A. <25% B. >=25% ~ <50% C. >=50% ~ <75% D: >=75%

Figure 36 Questionnaire to answer at the end of Lab-3 by subjects using the artifacts

Questionnaire for Lab-3

Levels of agreement:

1 – Strongly agree 2 – Agree 3 – Not certain 4 – Disagree 5 – Strongly disagree

Question:

	1	2	3	4	5
1. I had enough time to perform the lab tasks	☐	☐	☐	☐	☐
2. The instructions and objectives of the lab were perfectly clear to me	☐	☐	☐	☐	☐
3. I felt comfortable defining the Domain Model Class Diagram	☐	☐	☐	☐	☐

4. How much time (in terms of percentage) did you spend in building the Domain Model during this lab? ☐
5. How much time (in terms of percentage) did you spend in writing the Glossary? ☐

The answer for the last three questions will be in the range from A to D, where the meaning of each letter is as follow:

A. <25% B. >=25% ~ <50% C. >=50% ~ <75% D: >=75%

Figure 37 Questionnaire to answer at the end of Lab-3 by subjects not using the artifacts

Questionnaire for Lab-4

Levels of agreement:

1 – Strongly agree 2 – Agree 3 – Not certain 4 – Disagree 5 – Strongly disagree

Question:

	1	2	3	4	5
1. I had enough time to perform the lab tasks	☐	☐	☐	☐	☐
2. The instructions and objectives of the lab were perfectly clear to me	☐	☐	☐	☐	☐
3. I felt comfortable defining the Domain Model Class Diagram	☐	☐	☐	☐	☐
4. I felt comfortable defining OCL contracts	☐	☐	☐	☐	☐
5. OCL contracts and System Sequence Diagram were useful to refine the class diagram by					
5.1 Detecting new classes	☐	☐	☐	☐	☐
5.2 Detecting new attributes	☐	☐	☐	☐	☐
5.3 Detecting new relationships between classes	☐	☐	☐	☐	☐
6. How much time (in terms of percentage) did you spend in writing OCL contracts during this lab? ☐					
7. How much time (in terms of percentage) did you spend in building the Domain Model during this lab? ☐					
8. How much time (in terms of percentage) did you spend in writing the Glossary? ☐					
9. My Domain Model would have been better in the previous lab when working in the other system if I had been given the corresponding System Sequence Diagram and OCL contracts.	☐	☐	☐	☐	☐

The answer for question six, seven and eight will be in the range from A to D, where the meaning of each letter is as follow:

A. <25% B. >=25% ~ <50% C. >=50% ~ <75% D: >=75%

Figure 38 Questionnaire to answer at the end of lab 4 by subjects using the artifacts

Questionnaire for Lab-4

Levels of agreement:

1 – Strongly agree 2 – Agree 3 – Not certain 4 – Disagree 5 – Strongly disagree

Question:

	1	2	3	4	5
1. I had enough time to perform the lab tasks	☐	☐	☐	☐	☐
2. The instructions and objectives of the lab were perfectly clear to me.	☐	☐	☐	☐	☐
3. I felt comfortable defining Domain Model Class Diagram	☐	☐	☐	☐	☐
4. System Sequence Diagrams would have been useful to improve the Domain Model	☐	☐	☐	☐	☐
5. System Operation Contracts would have been useful to improve the Domain Model	☐	☐	☐	☐	☐
7. How much time (in terms of percentage) did you spend in building the Domain Model during this lab? ☐					
8. How much time (in terms of percentage) did you spend in writing the Glossary? ☐					

The answer for question six, seven and eight will be in the range from A to D, where the meaning of each letter is as follow:

A. <25% B. >=25% ~ <50% C. >=50% ~ <75% D: >=75%

Figure 39 Questionnaire to answer at the end of Lab-4 by subjects not using the artifacts

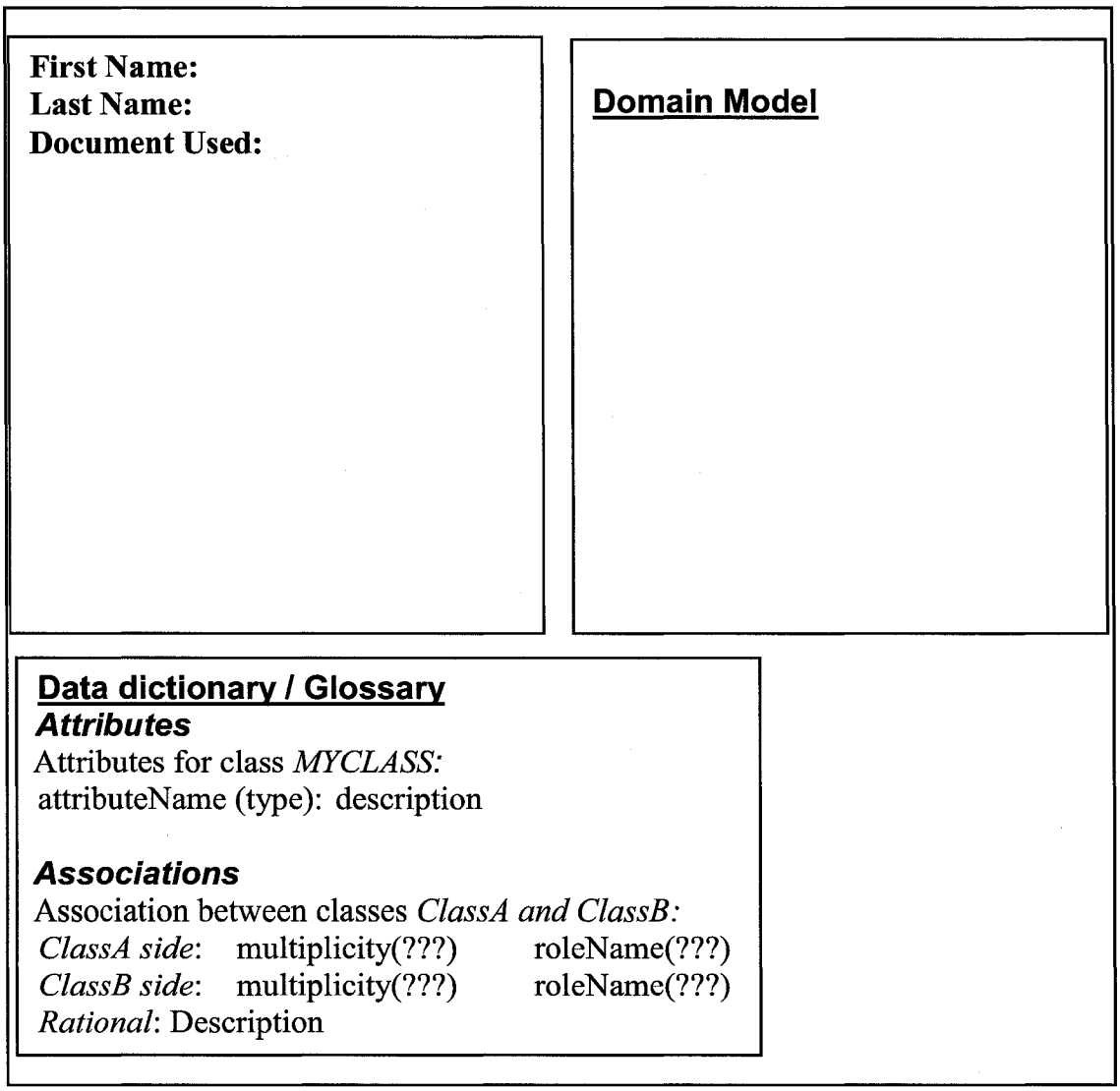

Figure 40 Template for subjects providing Domain Model without using the artifacts

Students copied and pasted the solution they modeled in Visio-2000. Additionally, they provided an explanation about each attribute and relationship between classes they devised.

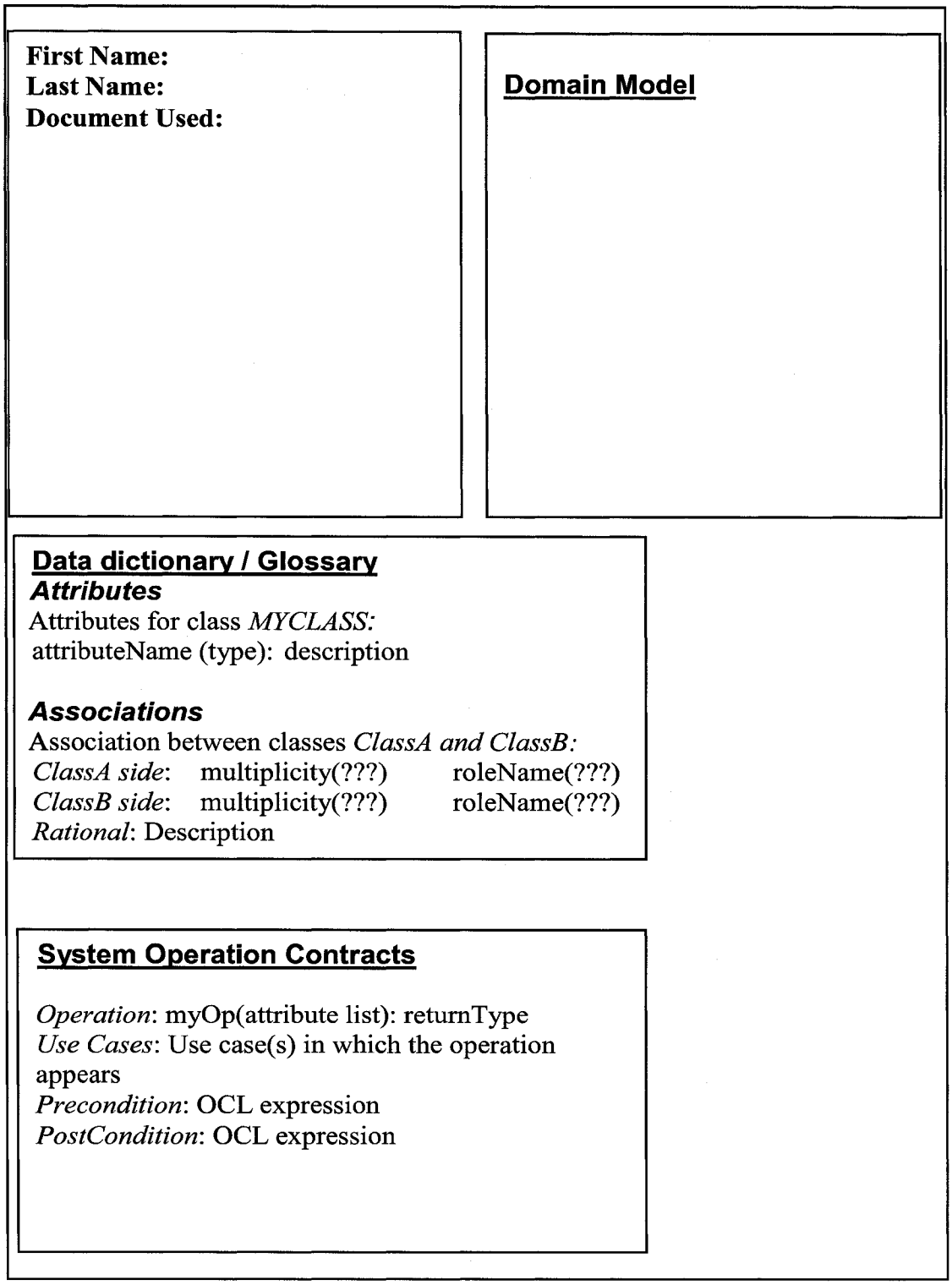

Figure 41 Template for subjects providing Domain Model using the artifacts

Students did similar as it was explained in Figure 40. Additionally, they provided System Operation Contracts using OCL expressions.