M. Staron, L. Kuzniarz and C. Wohlin, "Empirical Assessment of Using Stereotypes to Improve Comprehension of UML Models: A Set of Experiments", Journal of Systems and Software, Vol. 79, No. 5, pp. 727-742, 2006.

Empirical Assessment of Using Stereotypes to Improve Comprehension of UML Models: A Set of Experiments

Miroslaw Staron, Ludwik Kuzniarz, Claes Wohlin

Department of Systems and Software Engineering

School of Engineering

Blekinge Institute of Technology

Box 520, Soft Center

SE-372 25 Ronneby, Sweden

(Miroslaw.Staron, Ludwik.Kuzniarz, Claes.Wohlin)@bth.se

Corresponding author:

Miroslaw Staron (affiliation as above)

Tel: +46 457 38 58 12

Fax: +46 457 271 25

Abstract

Stereotypes were introduced into the Unified Modeling Language to provide means of customizing this general purpose modeling language for its usage in specific application domains. The primary role of stereotypes is to brand an existing model element with specific semantics, but stereotypes can also be used to provide means of a secondary classification of modeling elements. This paper elaborates on the influence of stereotypes on the comprehension of models. The paper describes a set of controlled experiments performed in academia and industry which were aimed at evaluating the role of stereotypes in improving comprehension of UML models. The results of the experiments show that stereotypes play a significant role in the comprehension of models and the improvement achieved both by students and industry professionals.

1 Introduction

The Unified Modeling Language (UML, [18]) is a general purpose modeling language, which has been gaining popularity in the last decade and has become the de facto standard for modeling artifacts produced during object-oriented software development. The language is composed of general purpose model elements and mechanisms, called extension mechanisms, which allow for customization of the language according to local needs and requirements such as a specific domain, specific software development process or specific problem. The notion of a stereotype is one of these mechanisms. The principle for using stereotypes in UML is that they allow branding an existing element with specific properties. The intention of using the specific properties was that the express some specific semantics associated with the branded model element. It seems that the notion of stereotype can be used not only to express properties of model elements that are beyond the core semantics, but also to introduce new virtual modeling elements which could improve quality properties of the models. This role of stereotypes in UML was indicated by Atkinson et al. [1] and is still not well investigated. This role, however, reflects the original intent of introducing stereotypes into object-oriented software development [28]. This paper is a contribution to the evaluation of the role of UML stereotypes in understanding of the UML encoded development models in both academia and industry. It presents a set of experiments designed to evaluate

the influence of stereotypes on the understanding of UML models. The set of experiments consists of four experiments preceded by a pilot study. Our intention was to evaluate to what degree stereotypes help in the comprehension of UML models. In order to perform a thorough evaluation considering several aspects, we conducted a series of experiments. The main evaluation is conducted in the first and the last experiment respectively while the potentially confounding aspects are evaluated in two auxiliary experiments. The results from a single experiment conducted in academic environment [15] cannot be generalized to industry professionals. A replication of the experiment in industry [24], if considered in separation to the other experiments, might not be generalizable to other contexts. Therefore, we performed a series of experiments in order to obtain the industrial validity of the results of the whole series of experiments and to use the series of experiments as stepping stones in a technology transfer process from academia to industry. Using industry professionals as subjects in experimentation is expensive and sometimes even impossible. In order to minimize the threat that the results are not significant due to errors in the design of the experiment, the three experiments in academia was performed in order to verify that the experiment performed in industry used the available resources to the best possible extent. Performing a set of experiments in academia before replicating the experiment in industry allowed us also to choose the optimal configuration of the experiment to replicate. The expectations of the results from industry in consequence allowed us to draw conclusions from the results of the industrial experiment despite a smaller number of subjects in the industrial experiment.

The results presented in this paper show the degree to which the stereotypes could help in improving the comprehension of UML models both in academia and in industry.

The structure of the paper is as follows. Firstly, the related work in the field of UML models and source code comprehension is presented in Section 2. The role of the specific kind of stereotypes used in the experiment is described in more detail in Section 3. The experiment design is described in Section 4. Section 5 presents the first experiment with student subjects and stereotypes. Section 6 presents two auxiliary experiments aimed at evaluating whether certain elements in the experiment design do not influence the results of the

experiments. Section 7 presents the industrial experiment. Finally, Section 8 discusses the results of all four experiments and Section 9 presents conclusions.

2 Related work

Stereotypes have been analyzed in different contexts in a theoretical way by Atkinson et al. [1-4], where the authors discuss and analyze various scenarios of using stereotypes in UML designs. The experiment design presented in this paper builds on the analyses in order to use appropriate stereotypes in an appropriate context for empirical evaluation. The analysis of the notion of stereotype from the perspective of the expressiveness of various stereotypes has been presented by Berner et al. [6], which also influenced the design of this study.

Each of the experiments presented in this paper follows an approach similar to that of other empirical studies in software engineering, object-orientation and UML, for example an experiment aimed at comparison of the understanding of sequence and collaboration diagrams in UML presented by Otero and Dolado [20]. The experiment conducted by Otero and Dolado provides results for evaluation of the diagrams for different purposes and domains conducted with student subjects. Empirical investigations, such as our experiments and the experiment presented by Otero and Dolado, provide tangible figures on how good or bad various alternatives (that a software developer can choose) can be in practice.

Briand et al. [7] evaluate whether object-oriented documents are easier to maintain. Similar to our experiments, it is an evaluation of two different ways of presenting equivalent information. The design of our experiments uses the experiences reported in [7]. Additionally, the experiment presented in this paper is similar to an experiment, presented by Hendrix et al. [10], on the influence of additional graphical information for source code comprehension. As stereotypes are intended to abstract certain properties of designs, the results of the experiments presented in this paper could be seen as evaluation of the way in which the abstractions are used in UML designs – in a similar way providing abstractions of source code for the purpose of improving comprehension.

Basili et al. [5] advocate organizing experiments into families by modifying a base design of experiment or subjects in each experiment in the family. In this manner, the results of separate experiments can be grouped, thus facilitating knowledge building. Our set of experiments can be seen as a small family of experiments since the experiments are closely related to each other and are based on a common design.

The UML 2.0 specification [19] introduces several changes to the definition of the notion of stereotype. Although our experiments are designed in the context of UML 1.x family of languages, they evaluate the role of stereotypes from the perspective of software developers (who are not language engineers), for whom the changes in the definition of the notion of stereotype are not significant. Therefore, the results of our experiments are valid for UML 2.0 as well.

3 Roles of stereotypes in UML

As defined in the UML specification documents [18], the main purpose for using stereotypes is to introduce new semantics to the existing model elements. The UML definition of stereotypes involves the definitions of other extension mechanisms – tagged values and constraints. The relationship between tagged values, constraints and stereotypes is analyzed by Gogolla and Henderson-Sellers [8] or Kuzniarz and Staron [13]. Stereotypes allow extending the language in a way, which is consistent with the definition of the language and they are useful in automatic model transformations, like for example code generation for a specific purpose [12, 21, 26].

Stereotypes and the new semantics expressed by them become even more important if they form profiles, which are closed sets of stereotypes definitions. Profiles provide a way of grouping stereotypes according to their purpose, allowing using UML for more specific needs. The most recognized profiles are the UML Profile for Business Modeling – which is a part of the UML specification, the UML Profile for Schedulability, Performance, and Time [17], the UML Profile for CORBA [16] and the Data Modeling Profile [9].

There is also another way of perceiving stereotypes. They provide a secondary classification of model elements. This concept was initially introduced by Wirfs-Brock et al [28], and discussed in detail by Atkinson et

al. [1]. Such stereotypes provide a means of expressing some classification of the stereotyped model elements, adding properties, which cannot be defined for all model elements of the same kind, but only for some. This kind of stereotypes can be called *model simplification stereotypes* [13], since they are intended to make models less complicated, not always involving the definition of a new semantics. Such usage of stereotypes can help readers of the stereotyped model to understand the model better. These stereotypes can also be classified as *transitive stereotypes* according to the classification presented by Atkinson et al. [1], because they are added to classifiers on the model level, but should also be recognized on the instance level. They are useful as a secondary classification mechanism [27] since they brand both the classifier and its instances with additional meaning. An example of such a stereotype taken from the empirical study presented in forthcoming sections is shown in Figure 1. The stereotype name is sender and it (in brief) means that instances of classes stereotyped as a sender is able only to send telecommunication signals (defined in Section 4.1) to instances of other classes, but cannot receive signals from other instances. In this sense, the stereotype is attached to a classifier (a class), but its meaning and restrictions apply to the instances of this class. This explains the reason why the graphical representation is attached to both the class and its instance. Further details on such applications of stereotypes can be found in the discussion provided by Atkinson et al. [1]. The application of stereotypes to both classes and objects is visible also in the diagrams presented in Appendix A.

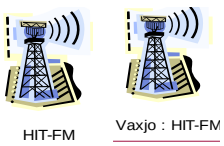

Figure 1. Example of a transitive stereotype. The sender stereotype is applied to a class (left-hand side), while its restrictions apply to instances.

Transitive stereotypes could help to distinguish between instances of standard model elements and instances of stereotyped model elements. The distinction seems to be useful in understanding the model and finding inconsistencies in models, or logical errors. An evaluation of this role of stereotypes could reveal the extent to which they influence the understanding.

4 Experiment design

The evaluation of the role of stereotypes in software development was done using a set of controlled experiments. The set of experiments had the same basic design. Each experiment presented in this paper could be summarized in the following way: “The goal of the experiment is to analyze *the comprehension of UML models* for the purpose of *evaluation of UML stereotypes* with respect to *their role in understandability of UML models* from the point of view of *the software developer* in the context of *UML domain modeling*”.

The type of the experiment design was a paired comparison design [29]. The treatments were model types, with two possible values - stereotyped model and non-stereotyped model. All subjects were assigned randomly to two groups: group 1 and group 2. There were two rounds conducted in each experiment. Two different sets of artifacts were presented to each subject in each group in each round. Table 2 in Section 4.6 summarizes the artifacts and their presentation to the groups. To avoid a learning effect, the artifacts came from two different application domains (A and B). The above are discussed in more detail in the following sections.

4.1 Experiment objects

The set of experimental objects consisted of four artifacts, summarized below. They are fully presented in [23]:

- Set A-S: stereotyped model A and description of stereotypes used in this model
- Set B-N: non-stereotyped model B,
- Set A-N: non-stereotyped model A,
- Set B-S: stereotyped model B and description of stereotypes used in this model.

Artifact set A-x describes a domain of radio and TV transmissions. It consists of a class diagram describing different types of existing objects: for example radio station, retransmission station, different types of antennas, etc., and a corresponding collaboration diagram describing one of possible situations such as sending a news program across a country. Excerpts of these diagrams can be found in Appendix A and are described later in this section.

Artifact set B-x describes a domain of GSM telephony. It consists of a class diagram describing different types of existing objects: for example mobile phone, transmission station, connection to conventional telephone network, etc., and a corresponding object diagram describing one possible situation of using the network, like making phone calls at a given time.

The best solution would be to have the same models (e.g. A) in both rounds as described in Section 5.3, but because of the learning effect in the second round it could not be done. If the subject saw the same model in both rounds the observed improvement could be caused by the fact that they see the model for the second time. Thus, another set of artifacts (i.e. B) was introduced, to avoid this situation. All the materials contained the same information and are of approximately equal complexity.

The sets of experiment objects which use stereotypes are based on a telecommunication profile. This profile was used as an example because the domain of telecommunication is intuitive with respect to the basic concepts gathered in the profile, although other profiles can be used in the experiment. The telecommunication profile contains stereotypes which should be seen as model simplification stereotypes according to the classification of Kuzniarz and Staron [13]. Considering the classification of stereotypes by Atkinson et al. [1], the stereotypes are classified as transitive stereotypes. In this paper, the definition of the profile is summarized in Table 1.

The profile introduces the following modeling elements from the telecommunication domain:

- active elements: sender, receiver and transmitter;
- passive links between the elements: transmission lines;
- signals sent between the active elements: transmissions;
- format of the content of transmission: transmission content.

Stereotype	Base class	Description
Sender	Class	It represents a class, which instances send telecommunication signals to instances of other classes, stereotyped Receiver or Transmitter
Receiver	Class	This stereotype represents a class, which instances receive telecommunication signals from instances of other classes, stereotyped Sender or Transmitter.
Transmitter	Class	It represents a class, which instances are used to relay telecommunication signals.

Table 1. Simplified stereotype definition table for the telecommunication profile

The stereotypes were presented to the subjects in a simplified form, which contained the representation and the description of the stereotype.

Excerpts of two models, A-S and A-N, can be found in Appendix A. Each model contains two diagrams, collaboration diagram and class diagram, which are presented in figures 10-13. The non-stereotyped models use inheritance and notes describing the intent and restrictions of the element to substitute stereotypes. The notes are attached to classes, since classes are regarded as definitions of objects. The description in the notes, however, applied to objects. The reason for not attaching the notes to objects is that it would explicitly distinguish objects of different types from each other and in that sense it would not be different than stereotyping and such a situation was identified as one of the threats to validity of the study. It can be observed that the stereotyped collaboration diagrams provide more information about the intent of the objects and therefore are more readable. The designs contain 30 objects in collaboration diagrams each and 11 (models with stereotypes) or 14 (models without stereotypes) classes in class diagrams.

4.2 Subjects

The study reported here was carried out using software engineering (and related) students and software engineering industry professionals. It was a desired sample of the population, as is explained later in this section. There are three main potential cases (types of subjects), summarized as follows:

1. A subject has worse knowledge of UML than knowledge of the telecommunication domain,
2. A subject has an equal knowledge on UML and the telecommunication domain, and
3. A subject has better knowledge of UML than knowledge of the telecommunication domain.

The third type of subject is the worst-case situation. Given the knowledge of UML the subjects should be able to better understand the model in standard UML than the model with stereotypes since the stereotypes are new concepts, which the subjects are not used to. This is the situation in which the introduction of stereotypes could deteriorate the comprehension of UML models. The introduction of stereotypes in this case requires an additional effort to learn them, whereas this effort is not required when standard UML is used. This could have a negative effect on the introduction of stereotypes. Other kinds of subjects could be either expected to perform better for stereotyped models (type 1) or at least equally well (type 2).

Students learnt UML during the object-oriented software development course, but they were not introduced to the notion of stereotype nor the telecommunication domain which makes them subjects of the third type. On the other hand, the industry professionals in the study work on the daily basis with UML and are familiar with stereotypes and the domain (which makes them subjects of the second type). This type of subjects is expected to achieve equal or larger improvement than the student subjects. Some indications on the differences between student subjects and professionals are also given by Höst et al. [11] or Tichy [25].

4.3 Independent and dependent variables

There was one independent variable in the experiment, the diagram type, with values: S (stereotyped) and N (non-stereotyped). Understandability of the designs was measured by three dependent variables:

I. Total score: NRESP – the number of correct answers for each subject when asked questions about the design.
II. Time: TSEC – the time, measured in seconds, which was required to fill in the questionnaire.
III. Relative time for a correct answer: TSEC/NRESP – the time, measured in seconds per answer, which is required to produce one correct answer.

The type of system could be considered as a second independent variable, but it was introduced only to minimize the learning effect in the second round of the experiment, and therefore it is not an independent variable.

4.4 Hypotheses and data analysis methods

Each experiment tested the null hypothesis. If it was falsified, it would mean that the introduction of stereotypes influences the understanding of UML models. The hypotheses were formulated as follows:

- Null hypothesis - H_0: Introduction of stereotypes does not influence understandability of UML models.
- Alternative hypothesis - H_1: Introduction of stereotypes improves understandability of UML models;

The hypotheses were evaluated separately for each variable. Each variable (TSEC, NRESP and TSEC/NRESP) was tested for the normal distribution using the Shapiro-Wilk test. If the test indicated that a distribution could not be classified as normal, a non-parametric Wilcoxon test was used to analyze the data. If the distribution could be classified as normal, the parametric paired t-test was used. The tests are recommended for this type of design by, for example Wohlin et al. [29], and their detailed description is presented by Siegel and Castellan [22].

4.5 Instrumentation

The main instruments used in the experiment were two comprehension questionnaires that measure the level of understanding of the presented UML models (one for each round). There were 12 questions in each questionnaire. The questions in the questionnaire concerning the same system, A or B, are identical regardless of whether the model is stereotyped or non-stereotyped. There were three types of questions asked in each questionnaire:

1. asking for the number of instances of classes of a certain type: sender, receiver or transmitter;
2. asking for the number of different types of elements in the diagram; and
3. checking whether some elements were placed correctly according to their definition.

The questions allowed measuring the level of understanding of UML models in terms of correct answers. An example of the first type of question is “How many receivers are shown on the object diagram A-S?” which requires the subject to count objects that are either stereotyped “receiver” or inherit from the receiver class (in case of non-stereotyped model). The second type of question concerned the class diagrams, to attract the attention of subjects to the definitions of objects and to enable them to get accustomed with the class diagram. A sample question is “How many types (kinds) of transmitters are shown in class diagram A-S?” The original diagrams contained more than one type of each element depicted by inheritance. An example of the third kind of question is “A signal cannot be transmitted via more than 2 transmitters; otherwise it is too weak to be received. How many too weak signals are shown in object diagram A-S?” It was aimed at checking the correctness of the model, i.e. an inspection-like question.

Subjects were asked to write down the time before starting answering the questions and after completing the questionnaire. The current time was displayed on the wall using a beamer. The measured total time for answering the questionnaire allowed for measuring the understanding of the model in terms of the required time to answer the questions concerning the model.

After the experiment, there is a post-experiment phase, where the subjects were asked to fill in the third, additional questionnaire about their background, prior knowledge of UML, prior knowledge of stereotypes and experience in the fields of software development as well as object-orientation. This was required to check that the subjects belong to the desired population. Since the participation in the experiments was not obligatory, the questionnaires about the background of the subjects could not be distributed beforehand. Nevertheless, in our experiments we knew the subjects and we knew that their level of knowledge about object-orientation, UML, stereotypes and telecommunication domain was sufficient for the experiment. It allowed us to classify them as the third type of subjects. In the case of industry professionals we had the possibility of having dialogues with each of them before the experiment and based on that we could classify them into the second type of subjects.

4.6 Experiment operation

In the course of the experiment, there were two rounds. In each round, each of the two groups is given a different treatment. Table 2 presents the outline of the experiment operation. The artifacts sets presented in the table are described in detail in Section 4.1. In each round each subject got a different type of artifact set as presented in Table 2.

	Round 1	Round 2
Group 1	Set A-S	Set B-N
Group 2	Set A-N	Set B-S

Table 2. Experiment rounds

The design of the experiment includes a lecture, given directly before the experiment. The intention of the lecture was to explain the notion of stereotypes and show some basic examples of the usage of stereotypes, but it did not introduce the set of stereotypes used in the experiment.

4.7 Design validation – pilot study

In order to initially verify and validate the design of the experiment as well as identify potential flaws in the design we conducted a pilot study. The pilot study was done with a group of two subjects, who were chosen based on their knowledge of UML and telecommunication domain. Each of them was acting as one group as described in Table 2. The results from the pilot study showed that there existed a confounding factor in the study stemming from the design of experiment objects. For a subject who was given the stereotyped model in the first round, the time for solving the assignment in the second round was shorter than in the first round. It was because the subject used some of his knowledge about stereotypes from round 1 to introduce the stereotypes to the non-stereotyped model in the second round. This introduction helped the subject to improve the time for solving the assignment.

From the pilot study it was also found that one of the models (objects of the experiment) was slightly less complex. There was also an ordering effect in the questionnaire, which made the introduction of stereotypes in the second round easier and intuitional. A dialogue with the subjects proved that they perceived stereotypes as helpful in understanding. The subjects indicated that one of the models was less complex.

As a result of the pilot study, the order of questions in the questionnaires was changed and the complexity of all models was balanced. The pilot study also indicated the extent to which the introduction of stereotypes could be useful. By introducing the stereotypes in the second round, one of the subjects showed that the set of stereotypes was indeed helpful in understanding the model.

4.8 Threats to validity of the study related to the design

As any empirical study, this study has threats to its validity. The threats that stem from the design of the study are grouped as suggested by Wohlin et al. [29] and presented below. Threats to the validity of an experiment are described in its section.

One of the most important threats to construct validity is the effect of interaction of testing and treatment. In group 2, each subject was given the stereotyped model prior to the non-stereotyped model. This could result in the introduction of stereotypes into non-stereotyped model in the second round. Since it was indicated by the pilot study, the ordering of questions was such that it minimized the effect and the models were prepared in a way, that introduction of stereotypes required some effort, which could be seen in the analysis of times for solving the assignment.

There is also a conclusion validity threat. Since the experimenters prepared the objects, there is a danger that the complexity of the design documents is not the same as the complexity of real-world design documents. On the other hand, the prior preparation of the objects resulted in an equal complexity of the models while the unbalanced complexity was seen as a larger threat.

An external validity threat is that the results of the study should be used with caution if they are to be considered for the purpose of evaluating how the enriched graphical notation improves the understanding of models

by different stakeholders. The evaluation of how the stereotypes improve the communication between different stakeholders would be possible if there were subjects of the first type in the study (knowing the domain better than the UML notation would indicate that subjects of this type are different stakeholders than the subjects of the third type – subjects that know UML better than the domain). It would require also a different empirical approach – i.e. observation rather than a controlled experiment.

The design of the study was done in a way to minimize the threats to the internal validity of the study (although sometimes introducing the other kinds of threats as discussed above), and thus there are no major threats in this category.

4.9 Summary of the set of experiments

We have conducted three experiments in academia in order to ensure that the fourth experiment in industry could be conducted as efficiently as possible given the potential problems with a small size of the sample in the industrial experiment. The first experiment was performed with students as subjects. Its intention was to evaluate the influence of stereotypes on comprehension of UML models. The experiment was intended to be replicated in an industrial setting. Nevertheless, before the replication we intended to verify that there are no potential confounding factors. One of these factors could be the order of presentation of experiment objects to subjects. In order to verify whether this factor has an influence on the results, we have conducted another (i.e. second) experiment in the academic setting. The results show that the order of presentation of objects does not influence the results. The other factor that could potentially influence the results was the representation of stereotypes in the experiment objects. In the first and second experiments the stereotypes were represented graphically as icons which are an alternative to the textual representation of stereotypes. We have conducted the third experiment to verify whether this factor could influence the results. The results of that experiment showed that icons might have some influence on the results. After these two auxiliary experiments, we have eventually conducted the fourth experiment in the industrial setting, which was an exact replication of the first experiment

except for having industrial professionals as subjects. The summary of the set of experiments is presented in Figure 2. The auxiliary experiments are drawn with dark filling.

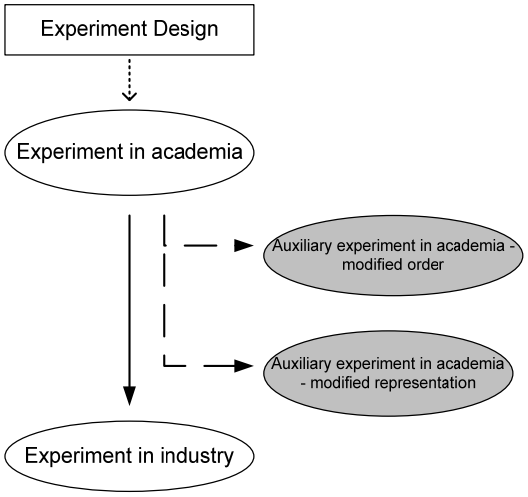

Figure 2. Summary of the set of experiments

Figure 2 summarizes the experiments conducted in our study and presents the experiments according to the order they were performed in. The first experiment in academia was conducted exactly according to the experiment design and is described in detail in Section 5.

The second experiment introduced changes to the order of presentation of objects in the study. Since two systems could be used as experiment objects (systems A and B), the results obtained could have been influenced by the fact that system A was used in the first round of the first experiment as presented in Table 2 for both groups. In the first experiment the results for the non-stereotyped model for group 2 in the second round could be influenced by the fact that even though the subjects were presented with different models, knowledge of stereotypes could help subjects in understanding the model better. In this experiment we changed the ordering for group 2 – both groups were presented with the stereotyped model first. The experiment is presented in Section 6.

The third experiment was the auxiliary experiment with modified representation of stereotypes in the experiment objects. The icons representing stereotypes were replaced with text in the so-called guillements to

represent stereotypes. An example of a stereotyped element using this representation of one of the stereotypes used in the experiment is presented in Figure 3.

Figure 3. The transmitter stereotype represented in the textual form

The main reason for introducing the change to the experiment design in this experiment was to investigate how much the comprehension was influenced by icons as such and not stereotypes that these icons represented. Since the icons are an appealing graphical representation, the sole presence of icons and not for example restrictions on using the stereotypes represented by the icons, could be the factor that cause improvement in terms of comprehension. Our intention was to verify whether it is the case in our experiments and hence the modification of the objects of the study. The sequence of presenting objects to groups is according to the sequence presented in the experiment design in Table 2. The experiment is presented in Section 6.

After testing various alternatives of the experiment we decided to replicate the first experiment. The replication was done as the fourth experiment – the experiment in industry. It was conducted exactly according to the experiment design and no changes were introduced.

5 Academic experiment

The first experiment was conducted during approximately two hours on a sample of 44 students of Software Engineering at Blekinge Institute of Technology in Ronneby, Sweden in March 2003. At the start, the subjects were given a 45 minutes lecture introducing the notion of stereotypes, explaining the usage of stereotypes and its graphical representation. The telecommunication profile was not explained during the lecture. Then, the subjects were divided into two equal groups using blocking. The blocking was done based on the study program of the students and the laboratory group. Both groups were in the same room as the lecture. Then, the subjects were given a short introduction to their task. The time was displayed with a beamer during the whole

time of experiment. The subjects were given the first comprehension questionnaire with stereotyped or non-stereotyped model according to the group they belonged to. After completing the first comprehension questionnaire the subjects were given the second comprehension questionnaire and after completing the second one, they were given the background questionnaire to fill in. The experiment was conducted in a classroom, where the students were supervised and no communication among them took place.

5.1 Analysis of experiment results

Although the experiment was performed on a group of 44 students it was found that answers from only 39 subjects could be taken into consideration: 20 in group 1 and 19 in group 2. The results of five subjects were removed due to some errors in choosing questionnaires, since two subjects solved two tests for the same system, and personal issues as one subject did not hand in one of the questionnaires. The results of the experiment indicate that introduction of stereotypes improves the understandability of UML models. It is interesting to examine the results from different perspectives: each variable independently and relative times for a correct answer - TSEC/NRESP. For each analyzed variable a bar plot is used as a presentation of the acquired data before the statistical significance tests. The influence of stereotypes on the number of correct responses for each subject is summarized in Figure 4.

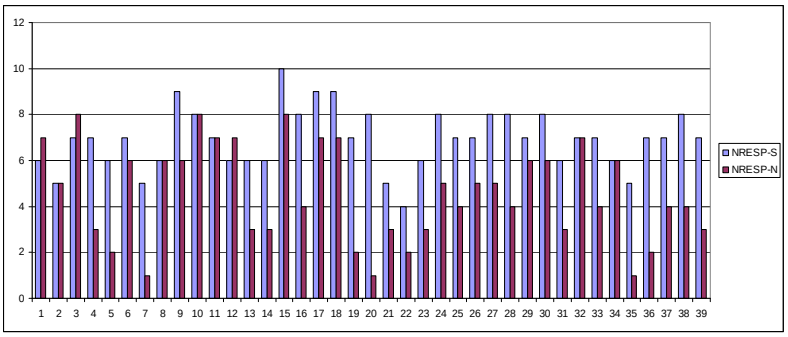

Figure 4. Number of correct answers for each subject. NRESP-S is the number of correct answers for the stereotyped model and NRESP-N is the number of correct answers in the non-stereotyped model.

The figure shows that there is a lot of variability in the data set although for the majority of subjects the correctness was larger for the stereotyped model. The times acquired from the subjects are summarized in Figure 5.

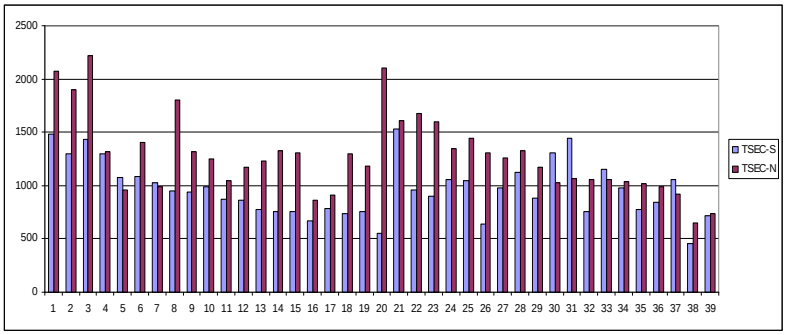

Figure 5. Time spent (in seconds) for answering the questionnaire for each subject. TSEC-S is the time spent for the stereotyped model and TSEC-N is the time spent for the non-stereotyped model.

The figure shows that for the majority of subjects the time required for answering the questionnaire is larger for the non-stereotyped model.

The comparison between the relative times required for a correct answer in each round for each subject gives an overview of the overall performance of a subject in each round. Figure 6 presents the results for each subject.

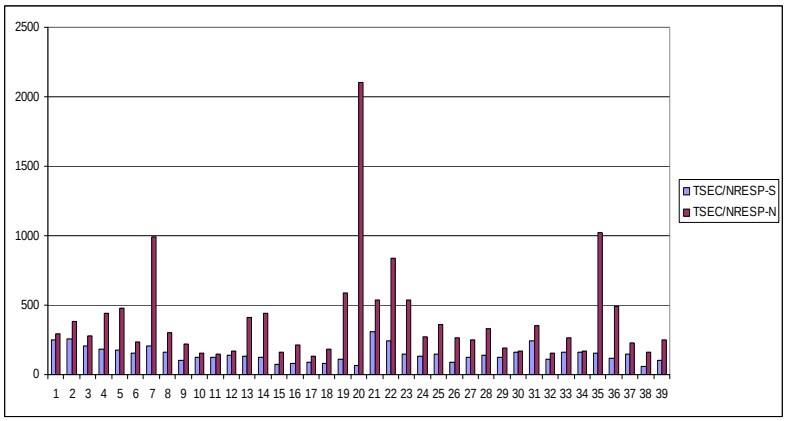

Figure 6. Relative time (in seconds) required for a correct answer. TSEC/NRESP-S is the relative time for the stereotyped model and TSEC/NRESP-N is the relative time for the non-stereotyped model.

Subjects 7, 20 and 35 can be identified as outliers in the analysis for this variable, since their results are more than two standard deviations away from the mean. They were not identified as outliers in the analyses of the previous variables, because the data was more unevenly distributed as indicated by their larger standard deviation.

The basic descriptive statistics for each variable for both treatments are presented in Table 3.

	Variable	Mean	Standard Deviation
Non-stereotyped model	NRESP	4.56	2.01
	TSEC	1281	367
	TSEC/NRESP	307	157
Stereotyped model	NRESP	6.92	1.29
	TSEC	967	260
	TSEC/NRESP	147	56
Differences: stereotyped – non-stereotyped	NRESP	2.36	1.87
	TSEC	-315	356
	TSEC/NRESP	-159	137

Table 3. Descriptive statistics for the first experiment.

The descriptive statistics indicate that the null hypotheses regarding each variable could be rejected, which is supported by the statistical significance testing which is summarized in Table 4. The column with the significance level of the hypothesis testing contains the name of the test used. The usage of the test for each variable is determined by the result of the normality test using Shapiro-Wilk test for normality for which the borderline significance level is chosen to be 0.05.

Variable	Improvement [value]	Improvement [%]	Significance level – Shapiro-Wilk	Significance level	H_0-accepted
NRESP	2.36	52	0.039	p<0.0001 (Wilcoxon)	No
TSEC	-315	25	0.074	p<0.0001 (paired t-test)	No
TSEC/NRESP	-159	52	p<0.0001	p<0.0001 (Wilcoxon)	No

Table 4. Results from statistical significance testing

The results from statistical significance testing show that the null hypotheses for all variables can be rejected. This means that the introduction of stereotypes improved the understanding of the models as measured by: number of correct answers, time spent for answering the questionnaire, and the relative time for one correct answer.

5.2 Threats to validity of the study

The main external validity threat is the sample, which may be considered as too homogenous, since it consists of students of the same year. This in a sense is a desired effect since the students represent a group of subjects that has the most undesired background for the evaluation. The results of the background questionnaire showed that the subjects know UML in practical applications, at the same time they are not familiar with the domain at the same level.

6 Auxiliary experiments

As mentioned previously, we conducted two additional experiments in order to verify the credibility of the design of the experiment before replicating it in industry. In this section we describe the experiments, briefly summarize their results and present implications of the auxiliary experiment on the whole study.

6.1 Experiment with modified order of object presentation

Subjects of this experiment were students of the Information Systems Program at Blekinge Institute of Technology. Their knowledge of object orientation, UML, stereotypes and the telecommunication domain was equal to the knowledge of the subjects in the first experiment. As the knowledge of all subjects was similar seven subjects were randomly assigned to group 1 and eight subjects to group 2. The experiment took place in a lecture room to which the students were accustomed to and no communication between them took place. The students attended the same course as the students in the first experiment.

6.1.1 Summary of the results

The results of the experiment are summarized in Table 5. The differences in variables denote the difference between the value of the variable for the stereotyped model and the value of the variable for the non-stereotyped model, e.g. NRESP-S – NRESP-N. The difference is presented as a value of the difference and as a percentage of the improvement/deterioration. As a basis for the calculation of the percentages, the value of the variable for the non-stereotyped model was taken (although it is not presented in the table). The calculations are the same as in Section 5. As in the case of the first experiment, the Shapiro-Wilk test was used with respect to the normal distribution; the significance level of the test is presented in the table. The significance levels of the Shapiro-Wilk test allowed using the paired t-test for statistical significance testing. The significance level of the statistical significance testing is gathered in the last column in the table.

Variable	Difference [value]	Difference [%]	Significance level – Shapiro-Wilk	Significance level – paired t-test
NRESP	2.80	69	0.073	0.001
TSEC	1.47	0.15	0.655	0.987
TSEC/NRESP	-155	52	0.102	0.003

Table 5. Summary of the results of the auxiliary experiment with modified presentation order

Statistical significance tests used show that the observed difference in means for NRESP and TSEC/NRESP variables are significant. It means that the stereotypes improved understanding of UML models as measured by the number of correct answers and by relative time for a correct answer. Average number of correct answers in the test was higher in the case of the stereotyped model by 2.80 and the average relative time for a correct answer was shorter by 155 seconds per answer in case of the stereotyped model. The values of the TSEC variable indicate deterioration in the time required for answering the questionnaire. The average time for answering the questionnaire was 1.47 seconds longer for the stereotyped model, but it was not found to be statistically significant.

As two out of three variables are significant and show improvement, we can conclude that the effect of the order of presentation of objects to subjects does not influence the results. The non significant results for one of the variables are a result of the small number of subjects in the study.

6.1.2 Threats to validity

In addition to the threats of validity of the experiment design there is a threat to conclusion validity in the case of this experiment - a small number of subjects participating in the study. Only fifteen subjects participating, reducing the power of the statistical significance tests used. Another threat is the lack of significance of the paired t-test for the time spent for answering the questionnaire, but the comparison using the relative time for a correct answer is significant.

6.2 Experiment with textual representation of stereotypes

The sample in this experiment consisted of students of the Information Systems Program at Blekinge Institute of Technology. The students were in the same point in the course as students in the previous experiments and their knowledge was similar to the knowledge of the students in the previous experiments. Nine subjects participated in the experiment, four of them were randomly assigned to group 1 as defined in Table 2, and five subjects were assigned to group 2. The experiment was conducted in a lecture room to which the subjects were used and there was no communication between the subjects during the course of the experiment.

6.3 Summary of the results

The results of the experiment are summarized in Table 6. The Shapiro-Wilk test used for testing whether the distributions can be regarded as normal allowed using the parametric test (paired t-test) in the case of NRESP and TSEC variables. The TSEC/NRESP needed to be analyzed with the non-parametric test (Wilcoxon).

Variable	Difference [value]	Difference [%]	Significance level – Shapiro-Wilk	Significance level
NRESP	2.33	45	0.905	0.065 (paired t-test)
TSEC	-320	20	0.211	0.222 (paired t-test)
TSEC/NRESP	-170	48	0.032	0.028 (Wilcoxon)

Table 6. Summary of the results of the auxiliary experiment with textual representation of stereotypes

The significance levels of the tests did not allow rejecting the null hypothesis on the equality of means for NRESP and TSEC, which means that the observed difference is not significant. In the case of the TSEC/NRESP variable, the observed difference is significant, which in turn means that stereotypes have influenced the relative time for a correct answer – the average relative time for a correct answer was shorter in case of the stereotyped model by 170 seconds per answer.

Based on the results, it seems that the presence of icons is an additional help in comprehension of UML models as the sole presence of stereotypes resulted in a smaller improvements which are only statistically significant for TSEC/NRESP variable. However, it should be noted that the result for NRESP (0.065) is close to significant and hence the findings are in line with the second experiment. Stereotypes help in particular when it comes to having correct answers and then also in the relative time for a correct answer.

6.4 Threats to validity

This experiment is burdened with a threat to conclusion validity – a small size of the sample in the experiment. Only nine subjects participated in the experiment making the statistical tests less powerful. Although the improvements were not small, the variability of the improvements did not allow statistically rejecting the null hypothesis in the case of separate variables NRESP and TSEC. Nevertheless, the significance test allows rejecting the null hypothesis of the equality of means in the case of the TSEC/NRESP variable.

7 Industrial experiment

The industrial experiment was conducted at Volvo Information Technology, at their site in Gothenburg, Sweden. The experiment was conducted during approximately two hours on a sample of four professionals chosen randomly from the personnel involved in UML modeling. At the start, the subjects were given a short introduction the experiment and their tasks. Then, the subjects were divided into two equal groups using blocking. The blocking was done based on short interviews with the subjects before the experiment. During the interviews with subjects we were able to find out enough information to classify them into the second type – subjects who have an equal knowledge in UML and the telecommunication domain. Both groups were in the same room. The time was displayed with the beamer during the whole time of the experiment. The subjects were given the first comprehension questionnaire with stereotyped or non-stereotyped model according to the group they belonged to. After completing the first comprehension questionnaire the subjects were given the second comprehension questionnaire and after completing the second one, they were given the background questionnaire to

fill in. The experiment was conducted in a meeting-room, where the subjects used to have regular working meetings and no communication among them took place. Nevertheless, there was some discussion at the beginning which led to clarification of certain concepts. Despite its presence, the discussion is not regarded as a confounding factor because it was supposed to clarify certain concepts which the subjects understood differently from the experimenters.

No changes were introduced into the design of the experiment; nevertheless, the statistical significance was not tested due to the small sample size.

7.1 Experiment results

The results of the experiment indicate that the alternative hypothesis H_1 can be supported based on that the null hypothesis could be rejected. The evaluation is based on the mean values without running statistical tests due to the small sample size. Due to the small number of subjects in the industry study, the discussion of the results is presented in Section 9, embedded in the context of the original experiment and the expected improvements from the industry professionals.

The results of the number of correct responses for each subject are summarized in Figure 7.

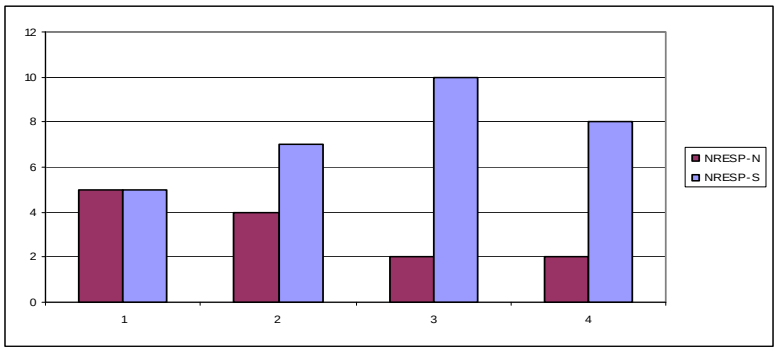

Figure 7. Number of correct answers for each subject. NRESP-S is the number of correct answers for the stereotyped model and NRESP-N is the number of correct answers in the non-stereotyped model.

Although it is not shown in the figure, the first subject did not provide answers to 7 out of 12 questions while examining the stereotyped model. The subject answered all of the remaining questions correctly. It could be

caused by accidental omission of one page of the second comprehension questionnaire, since all questions that were not answered were on the same page. Nevertheless, this page was not the last page and was included in the comprehension questionnaire that was returned by the subject. Despite this situation, the subject was not removed from the sample afterwards.

The time spent for answering the comprehension questionnaire is summarized in Figure 8.

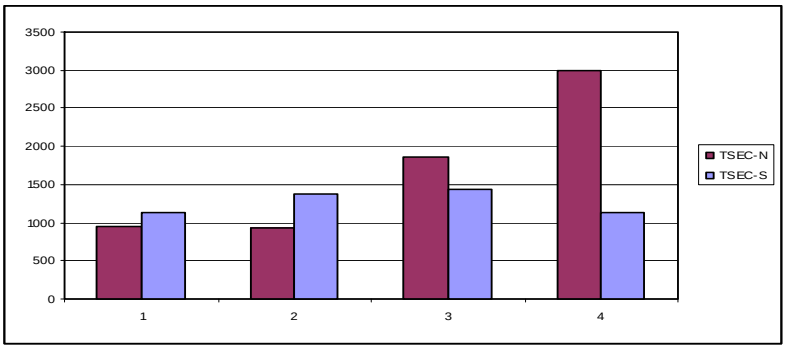

Figure 8. Time spent (in seconds) for answering the questionnaire for each subject.

The results show that two subjects actually showed deterioration in the time for answering the questionnaire and two subjects showed improvement. Despite that the magnitude of the improvements seems larger than the magnitude of the deterioration.

Figure 9 presents the summary of the results from the industrial experiment with respect to the relative time for a correct answer – TSEC/NRESP.

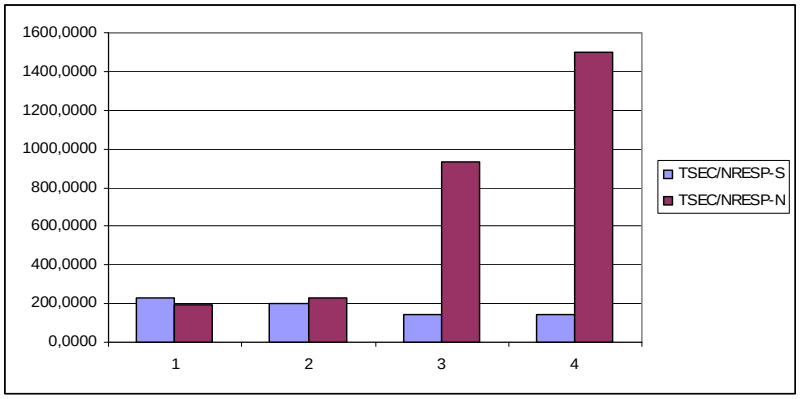

Figure 9 Relative time (in seconds) required for a correct answer.

The figure shows that the fourth subjects achieved a large improvement for the relative time for one correct answer. The descriptive statistics for all variables in the industrial experiment are presented in Table 7.

	Variable	Mean	Standard Deviation
Non-stereotyped model	NRESP	3.25	1.50
	TSEC	1688	976
	TSEC/NRESP	714	624
Stereotyped model	NRESP	7.50	2.08
	TSEC	1276	159
	TSEC/NRESP	178	42
Differences: stereotyped – non-stereotyped	NRESP	4.25	3.50
	TSEC	-411	1033
	TSEC/NRESP	-536	662

Table 7. Descriptive statistics for industrial experiment

The descriptive statistics indicate that the null hypotheses can be rejected. The introduction of stereotypes increases the number of correct answers by 131%, decreases the time required for answering the questionnaire by 24%, and decreases the relative time for one correct answer by 75%.

7.2 Threats to validity of the study

A threat to the construct validity in this experiment is that the lecture before the experiment was shorter than in the case of the other experiments. However there were fewer subjects and the experimenters were able to freely discuss issues with them – in particular the level of understanding of the stereotypes – so this threat has been minimized.

A threat to conclusion validity is the small sample size and thus the lack of statistical significance testing which is minimized by drawing conclusions in the context of the first experiment, in which the size of the sample was larger.

8 Discussion

The results of all four experiments differ slightly from each other. The descriptive statistics and improvements of the experiments are gathered in Table 8. The values of TSEC, TSEC/NRESP, and percentages are rounded to the nearest integer value except for the TSEC improvement in the second experiment, which is left more precise due to its very small value.

Variable	First experiment		Second experiment		Third experiment		Industrial experiment	
	Value	Percentage	Value	Percentage	Value	Percentage	Value	Percentage
NRESP	**2.36**	**52%**	**2.80**	**69%**	2.33	45%	4.25	131%
TSEC	**-345**	**25%**	*1.47*	*0.15%*	-320	20%	-411	24%
TSEC/NRESP	**-159**	**52%**	**-154**	**52%**	**-170**	**48%**	-536	75%

Table 8. Summary of all experiments' results. Significant results are in boldface.

The presented summary of the results for all the experiments shows that a statistically significant improvement was achieved in half of the analyzed cases. There is only one case of a small deterioration – the deterioration in terms of time for answering the questionnaire in the second experiment presented in italics in Table 8, which was not found to be statistically significant. The lack of significance in the statistical hypothesis testing was caused by the small number of data points in the second and third experiments or by a variability of the

results as in the case of the TSEC variable in the second experiment. In the case of the industrial experiment, the hypotheses were not statistically tested since there was too few data points.

The results from the industrial experiment allow for making statements based on only four subjects, which is not a big sample. Nevertheless, since it is an exact replication of the first experiment performed in academia except for subjects, the results can be discussed together for both experiments. The results show that the industry professionals achieved larger improvements than students in the first experiment in the correctness (NRESP – 131% vs. 52%) and relative time for a correct answer (TSEC/NRESP – 75% vs. 52%) analysis. The difference between the results in case of students and professionals can be caused by the difference in their background. According to the discussion in Section 4.2 regarding the types of subjects, the students in all experiments can be classified to the third type – subjects who have a better knowledge in UML than in stereotypes and the telecommunication domain. It was checked with the background questionnaire filled in by all subjects in all experiments. On the other hand, the industrial sample based on the interviews was classified as subjects of the second type, i.e. subjects with equal knowledge of both factor levels. As it was indicated in Section 4.2, the subjects of the second type should perform at least equally well as the subjects of the third type. This was the outcome in the case of this study. As the industry professionals are expected to perform better than students based on their background (and did so), it seems that the results of the industrial experiment are valid and trustworthy despite the lack of statistical significance testing. A detailed discussion on the expected differences between the types of subjects has been presented by Kuzniarz et al. [14].

As far as the TSEC variable is concerned the industry professional in the fourth experiment achieved the improvement which was smaller than the students' improvement (24% vs. 25%) in the first experiment. However, the difference between improvements is rather small compared to the magnitude of the improvements hence the results can be regarded as basically the same. As we indicate in Section 4.2, the performance of the subjects of the second type is expected to be at least the same as the performance of the subjects of the third type. The results of the industrial experiment for the TSEC variable seem to be trustworthy.

In general the stereotypes improve the correctness of understanding of UML models. Improvements were achieved in all experiments – and half of the results were statistically significant. In all configurations of the experiment design used in the presented set of experiments the stereotypes were found to improve the correctness of understanding of UML models and improved the ability of subjects to find logical errors in the design. The size of the improvement varies from 45% for the third experiment (stereotypes represented as text) to 131% in the industrial experiment (stereotypes represented by graphical icons). Supported by the fact that the deterioration was achieved only in one case and it was a very small one, we could conclude that the stereotypes improve the comprehension of UML designs.

The same is true with respect to the relative time for a correct answer – an improvement was achieved in all experiments and the results were statistically significant in all experiments for which a statistical significance analysis was performed.

Furthermore, the lowest improvements in terms of NRESP and TSEC/NRESP were achieved in the third experiment, in which the stereotypes were not represented by graphical icons. One of the intentions of the auxiliary experiments was to check whether the results are biased towards stereotypes since icons were used to represent stereotypes. We wanted to check whether there are improvements if we use text as a representation for stereotypes instead of icons. The results show that the icons help, but they are not the only cause of the improvement. Any representation of stereotypes improves the understanding. Although the results are not significant, they can be regarded in the course of qualitative analysis, since the lack of significance was caused by the small sample size. The fact that the size of the improvement is smaller than in other experiments also increases the external validity of the results – the smaller improvement could be expected since icons are more visible than text in models.

On the other hand, during the second experiment there was a result that was in contradiction to other results – deterioration in the time for answering the questionnaire (TSEC – deterioration of 0.15%, in italics in Table 8). Nevertheless, the correctness of answering the questionnaire was improved (NRESP - improved by 69%, statistically significant). Such a result is an example of a situation for which we needed to provide two depen-

dent variables (TSEC and NRESP) in the study – the subjects were answering better when they were examining models with stereotypes, but it took them longer time to do so. In order to investigate whether it was an actual deterioration or improvement, we calculate and analyze the indirect measure – the relative time for a correct answer TSEC/NRESP – which in the case of the second experiment indicates a statistically significant improvement of 52%.

9 Conclusions and further work

The paper presents a set of four controlled experiments conducted both in academia and in industry, totaling of 72 subjects. The results of all four experiments, aimed at evaluating the role of stereotypes in understanding of UML models, confirm that the use of stereotypes improve UML model comprehension and show the magnitude of the improvements. Improvements were achieved in the following three aspects:

- an increase in the number of correct answers in the tests checking the level of understanding,
- a decrease in the time required for answering the questionnaire, and
- a decrease in the relative time for a correct answer.

The first experiment was conducted in academia with the largest number of subjects. It was also replicated with industrial subjects in the fourth experiment to obtain industrial validity of the results. The order of presenting of stereotyped and non-stereotyped models was chosen to be different for each experimental group. Since the order of presenting the models in experiment rounds could influence the results, we have conducted another experiment – the second experiment presented here – which changed the order of presentation of models. The results showed that the order could influence the results, but there was still improvement after introduction of stereotypes. Furthermore, the graphical icons used for representation of stereotypes in experiment objects could be the main factor that caused the improvement. To address this threat we conducted the third experiment, in which the icons were replaced by a text. Even in this case, the introduction of stereotypes showed improvements. Finally, the fact that the study subjects were students could be one of the factors that

influence the results, so the fourth experiment was conducted in the same way as the first experiment with subjects that were industry professionals.

Considerable improvements were achieved in all four experiments for stereotypes that were represented both as text and with icons. The stereotypes with their graphical representation as icons improve the understanding of UML models more than stereotypes represented as textual adornments of model elements. The claim is based on the evaluation of the hypothesis with statistical significance testing for three experiments and a qualitative analysis for the industrial experiment. The largest improvement was achieved in the case of the industrial experiment. The results of the industrial experiment are in accordance with the expectations based on the division into different types of subjects. Thus, although the number of professionals participating in this study was small, the results are fully aligned with those of the original study and therefore can be regarded as valid.

The presented results contribute to the evaluation of the role of stereotypes in software development. Further research in this field should include an industrial case study on usage of stereotypes in more specific industrial applications, for example stereotypes defining a specific domain or a specific software development process. Furthermore, the evaluation of the scale of improvement in understanding the models in the industrial applications can be done by a case study. The results of the study can be a basis for evaluation of the cost of introduction of stereotypes of different kinds. Such analysis is planned to be done in the next step of our research.

Acknowledgements

The authors would like to thank the subjects, without whom the study would not have been possible. The authors would also like to thank the anonymous reviewers for their valuable comments on the paper and Lawrence E. Henesey for proof-reading the paper.

References

1. Atkinson, C., Kuhne, T. and Henderson-Sellers, B., Stereotypical Encounters of the Third Kind. in *UML 2002*, (Dresden, 2002), Springer-Verlag, 100-114.
2. Atkinson, C. and Kühne, T. Profiles in a Strict Metamodeling Framework. *Science of Computer Programming*, *44* (1). 5-22.
3. Atkinson, C. and Kühne, T. Rearchitecting the UML Infrastructure. *ACM Transactions on Modeling and Computer Simulation*, *12* (4). 290-321.
4. Atkinson, C., Kühne, T. and Henderson-Sellers, B. Systematic Stereotype Usage. *Software and Systems Modeling*, *2* (3). 153-163.
5. Basili, V.R., Shull, F. and Lanubile, F. Building Knowledge through Families of Experiments. *IEEE Transactions on Software Engineering*, *25* (4). 456-473.
6. Berner, S., Glinz, M. and Joos, S., A Classification of Stereotypes for Object-Oriented Modeling Languages. in *The Second International Conference on the Unified Modeling Language*, (Fort Collins, CO, USA, 1999), Springer-Verlag, 249-264.
7. Briand, L.C., Bunse, C., Daly, J.W. and Differding, C. An Experimental Comparison of the Maintainability of Object-Oriented and Structured Design Documents. *Empirical Software Engineering*, *2* (3). 291-312.
8. Gogolla, M. and Henderson-Sellers, B., Analysis of UML Stereotypes in the UML Metamodel. in *UML 2002*, (Dresden, 2002), Springer-Verlag, 84-99.
9. Gornik, D. UML Data Modeling Profile, Rational Corp., 2002.
10. Hendrix, D., II, J.H.C. and Maghsoodloo, S. The Effectiveness of Control Structure Diagrams in Source Code Comprehension Activities. *IEEE Transactions on Software Engineering*, *28* (5). 463-477.
11. Höst, M., Regnell, B. and Wohlin, C. Using Students as Subjects - a Comparative Study of Students and Professionals in Lead-Time Impact Assessment. *Empirical Software Engineering*, *5* (3). 201-214.
12. Kuzniarz, L. and Ratajski, J., Code Generation Based on a Specific Stereotype. in *Information Systems Modeling*, (Roznov, Czech Republic, 2002), MARQ, 119-128.

13. Kuzniarz, L. and Staron, M., On Practical Usage of Stereotypes in UML-Based Software Development. in *Forum on Design and Specification Languages*, (Marseille, 2002), FDL, 262-270.
14. Kuzniarz , L., Staron, M. and Wohlin, C., Students as Subjects in Software Engineering Experimentation. in *Third Conference on Software Engineering Research and Practise in Sweden*, (Lund, Sweden, 2003), Lund Institute of Technology, 19-24.
15. Kuzniarz, L., Staron, M. and Wohlin, C., An Empirical Study on Using Stereotypes to Improve Understanding of UML Models. in *The 12th International Workshop on Program Comprehension*, (Bari, Italy, 2004), IEEE Computer Society, 14-23.
16. Object Management Group. UML Profile for CORBA, Object Management Group, www.omg.org, 2002, last accessed 2004.
17. Object Management Group. UML Profile for Schedulability, Performance and Time, Object Management Group, www.omg.org, 2002, last accessed 2005.
18. Object Management Group. Unified Modeling Language Specification V. 1.5, Object Management Group, www.omg.org, 2003, last accessed 2003.
19. Object Management Group. Unified Modeling Language Specification: Infrastructure Version 2.0, Object Management Group, www.omg.org, 2004, last accessed 2005.
20. Otero, M.C. and Dolado, J.J. An Initial Experimental Assessment of Dynamic Modeling in UML. *Empirical Software Engineering*, *7*. 27-37.
21. Rational. Rose Data Modeling Profile, Rational, www.rational.com, 1999, last accessed 2005.
22. Siegel, S. and Castellan, N.J. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, 1988.
23. Staron, M. Experiment on the Role of Stereotypes in UML Based Software Development - Materials, http://www.ipd.bth.se/mst/Experiment/index.html, 2003, last accessed 2005.

24. Staron, M., Kuzniarz, L. and Wohlin, C., An Industrial Replication of an Empirical Study on Using Stereotypes to Improve Understanding of UML Models. in *Software Engineering Research and Practice in Sweden*, (Linköping, Sweden, 2004), Department of Computer and Information Science, 53-62.
25. Tichy, W. Hints for Reviewing Empirical Work in Software Engineering. *Empirical Software Engineering*, *5* (4). 309-312.
26. Uhl, A. and Lichter, H., A UML Variant for Modeling System Searchability. in *Object Oriented Information Systems*, (Monpellier, 2002), Springer-Verlag, 199-211.
27. Wirfs-Brock, R. Stereotyping: A Technique for Characterizing Objects and Their Interactions. *Object Magazine*, *3* (4). 50-53.
28. Wirfs-Brock, R., Wilkerson, B. and Wiener, L. Responsibility-Driven Design: Adding to Your Conceptual Toolkit. *ROAD*, *2* (1). 27-34.
29. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B. and Wesslèn, A. *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publisher, Boston MA, 2000.

Appendix A – artefacts excerpts

Sample parts of models shown to subjects in the study are presented herein to give an overview of the diagrams. For the sake of simplicity, only parts of the models are included and only from one model (A), the other model (B) was similar in the content, although it defined different elements (for instance telephones instead of antennas).

Artefact set A-S

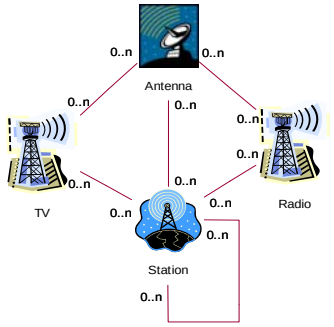

Figure 10 Excerpt from class diagram in artefacts set A-S (about 50% of the diagram)

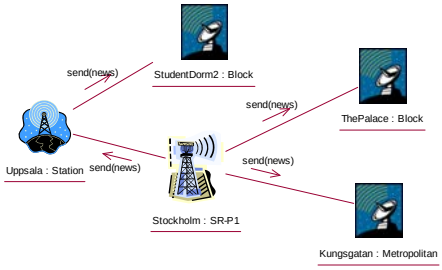

Figure 11 Excerpt from the collaboration diagrams from artefacts set A-S (around 20% of the diagram).

Artefact set A-N

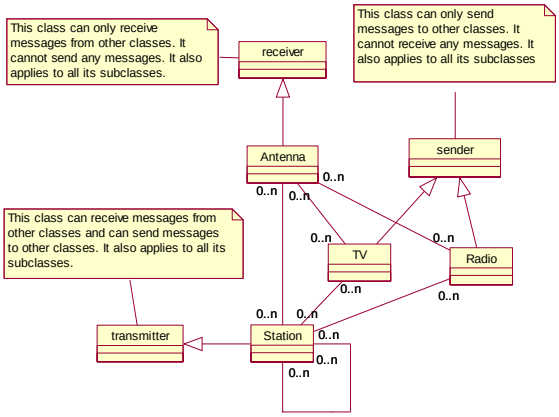

Figure 12 Excerpt from the class diagram in artefacts set A-N (about 50% of the diagram).

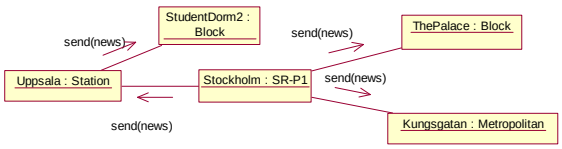

Figure 13 Excerpt from the collaboration diagram from set A-N (about 20% of the diagram).