An Empirical Study on Using Stereotypes to Improve Understanding of UML Models

Ludwik Kuzniarz, Miroslaw Staron, Claes Wohlin
Department of Software Engineering and Computer Science
Blekinge Institute of Technology
Box 520, Soft Center
SE-372 25 Ronneby, Sweden
(lku, mst, cwo)@bth.se

Abstract

Stereotypes were introduced into the Unified Modeling Language (UML) to provide means of customizing this visual, general purpose, object-oriented modeling language, for its usage in specific application domains. The primary purpose of stereotypes is to brand an existing model element with a specific semantics. In addition, stereotypes can also be used as notational shorthand. The paper elaborates on this role of stereotypes from the perspective of UML, clarifies the role and describes a controlled experiment aimed at evaluation of the role – in the context of model understanding. The results of the experiment support the claim that stereotypes with graphical icons for their representation play a significant role in comprehension of models and show the size of the improvement.

1. Introduction

The Unified Modeling Language (UML, [1]) is a general purpose visual object-oriented modeling language, which has been gaining popularity in the last decade and became de facto standard for modeling artifacts produced during object oriented software development. The language was designed in such a way that there is a well defined set of general purpose model elements and there are mechanisms, called extension mechanisms, which allow for its customization according to local needs and requirements such as a specific domain, specific software development process or specific problem. One such mechanism is the notion of a stereotype. The idea behind a stereotype in UML is that it allows branding an existing element with specific properties. The intention was that the specific properties should express some specific semantics associated with the branded model element. And this is how the stereotypes have been commonly used. But it seems that the notion can not only be used 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 in [2] and is still not well investigated, although it reflects the original intent of introducing stereotypes into object-oriented software development in [3]. This paper is a contribution to the evaluation of the role of UML stereotypes, which are dedicated for simplification of models and improving understanding of the UML encoded development models. It presents an empirical experiment designed to evaluate the influence of such stereotypes on the understanding of UML models. The results presented show the extent to which the stereotypes represented with graphical icons could help in improving the comprehension of UML models. The empirical study follows a similar approach to other empirical studies in software engineering, object-orientation and UML [4, 5]. The study was done on student subjects but its results can be generalized to a broader population, since the students represent a sample, which is expected to have the smallest improvement.

The structure of the paper is as follows. Firstly, the role of the specific kind of stereotypes is described in more detail in section 2. Then the design of an experiment aimed at evaluation of the role of stereotypes is presented in sections 3 and 4. Finally, the results of the experiment are presented in chapter 5, followed by a discussion of results and scale of improvement, conclusions and indications for further works in the last section.

2. Roles of stereotypes in UML

As defined in the UML specification documents [1], the main idea behind 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 (how they are involved is analyzed in [6, 7]). Such a definition of stereotypes is useful for their automatic processing in UML tools, because they separate the definition of syntactical (tagged values) and semantic information (constraints). It also allows extending the language in a way, which is consistent with the definition of the language. Stereotypes are useful in automatic model transformations, like for instance code generation for a specific purpose (i.e. [8-10]).

Stereotypes (and the new semantics expressed by them) are very important if they form profiles, which are closed sets of stereotypes definitions (along with constraints and tagged values). Profiles provide a way of grouping stereotypes according to their purpose, allowing using UML for other, more specific needs (i.e. changing the language so that it is as some other notation, for instance Entity Relationship Diagrams [7]), whereas the separate stereotypes which do not constitute profiles change the separate model elements only. The most recognized profiles are the UML profile for business modeling (part of the UML specification [1]), the UML profile for scheduling and performance [11], the UML profile for CORBA [12] and the data modeling profile [13].

There is also another way of perceiving stereotypes. They provide a secondary classification of model elements. This concept was initially introduced in [14], and discussed in detail in [2]. 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* [7], 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 it better. These stereotypes can also be classified as *transitive stereotypes* (according to the classification presented in [2]), 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 ([15]) since they both brand the classifier and its instances with additional meaning. An example of such a stereotype (taken also from the empirical study presented in forthcoming sections) is shown on Figure 1. The stereotype name is sender and it (in brief) means that instances of classes stereotyped as sender are only able to send signals (telecommunication signals – see section 3.1) 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 (see [2] for further discussions on such application of stereotypes).

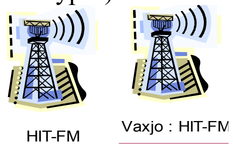

Figure 1 Example of a transitive stereotype. The sender stereotype is applied to a class (left-hand side), while its restrictions apply to instances

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. This paper presents a controlled experiment where the effect of stereotypes as a means for increased understanding is evaluated.

3. Experiment design

The evaluation of the role of stereotypes in software development is done using a controlled experiment. The study 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 represented graphically with icons* 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 is a paired comparison design [16]. The treatments are model types, with two possible values - stereotyped model and non-stereotyped model. All subjects are assigned randomly to two groups (group 1 and 2). There are two rounds in the experiment. Two different sets of artifacts are presented to every subject in each group in each round. Table 1 (see section 3.6) summarizes the artifacts and their presentation to the groups. To avoid a learning effect, the artifacts come from two different application domains (A and B). The artifacts are similar in complexity and the domains are similarly common to all subjects. The above are discussed in more detail in the following sections.

3.1. Experiment objects

The set of experimental objects consists of four artifacts, summarized below (they are fully presented in [17]):

- 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 instance radio station, retransmission station, different types of antennas, etc.) and a corresponding collaboration diagram describing one of possible situations (like 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 instance 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 (A) in both rounds (c.f. section 5.3), but because of the learning effect in the second (subjects could understand the model better in the second round simply because they see the model for the second time) round it cannot be done. So another set of artifacts (B) is introduced, to avoid it. All the materials were ensured to contain the same information and to be of equal complexity.

The sets of experiment objects, which use stereotypes, are based on a telecommunication profile, introduced briefly in this section. This profile is 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 in [7]). Considering the classification of stereotypes in [2], the stereotypes are added to the elements on the model level, but their semantics concerns also elements at the model instance level (transitive stereotypes). In this paper, the definition of the profile is only informal, while the details are omitted for the sake of simplicity of the description.

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

- active elements like: 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.

Table 1 summarizes the elements in the profile along with their description.

The constraints are expressed in the Object Constraint Language (OCL, [18]) where possible. In situations where the constraints should be defined on the meta-model level, but refer to the model instance level (for instance restriction that the operation stereotyped Transmission should be transmitted on links which are instances of classes stereotyped Transmission Line), the constraints are expressed only in natural language. Such constraints cannot be checked automatically and therefore it is important to investigate how many of violations of these constraints can be found by manually checking the models. In addition to OCL, every constraint is specified in natural language to make it more easily readable. An example constraint is presented below.

SC_1: Sender cannot receive any transmission - there are no operations stereotyped Transmission defined for Sender

```
context sender

inv: extendedElement->allOperations-> exists
(op:Operation| op.stereotype.name="transmission")
```

The details of the definitions are presented in [19].

Stereotype	Base class	Con-straints	Description
Sender	Class	SC_x	It represents a class, which instances send telecommunication signals to instances of other classes, stereotyped Receiver or Transmitter
Receiver	Class	RC_x	This stereotype represents a class, which instances receive telecommunication signals from instances of other classes, stereotyped Sender or Transmitter.
Transmitter	Class	TRC_x	It represents a class, which instances are used to relay telecommunication signals.
Trans-mission Line	Asso-ciation	TLC_x	This stereotype represents a transmission line, which allows communication between instances of stereotyped classes. It allows the transmission of telecommunication signals
Trans-mission	Operation	TOC_x	It depicts a telecommunication signal transmitted between instances of stereotyped classes.
Trans-mission Content	Classifier	TCC_x	It represents the format of information being sent as a communication signal.

Table 1 Simplified stereotype definition table for the telecommunication profile

Although the stereotypes are specified according to the UML specification, their intention is primarily to improve the understanding the UML models in which they are used, therefore they were presented to subjects in a simplified form, which contained the graphical representation and the description of the stereotype as presented in table 1.

Excerpts of two models (A-S and A-N) can be found in Appendix A. Each model contains two diagrams (collaboration and class diagrams) and they are presented in figures 6-9. Inheritance hierarchies and notes describing the intent and restrictions of the element substituted stereotypes used in the stereotyped models. The notes were attached to classes, since classes are usually regarded as definitions of objects. The description in the notes, however, applied to objects. The reason for not attaching the notes to objects was 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. It should reflect the results of the study. The designs contain 30 objects in collaboration diagrams each and 11 (models with stereotypes) or 14 (models without stereotypes) classes in class diagrams. The designs are as complex as the subjects are used to working with.

3.2 Subjects

The study was carried out using software engineering (and related programs) students. It was a desired sample of the population, as is explained later in this section. The students were taking part in an object-oriented software development course, which consisted of theoretical lectures, practical exercises and individual projects. Stereotypes were not introduced during the course. There are four main potential cases (types of subjects), summarized as follows:

- A subject has worse knowledge of UML than knowledge of the telecommunication domain,
- A subject has an equal knowledge on UML and the telecommunication domain,
- A subject has better knowledge of UML than knowledge of the telecommunication domain, and
- A subject has no knowledge of either UML or the telecommunication domain.

The third type of subject is the worst-case situation. Given the knowledge of UML the subject should be able to better understand the model in standard UML (since this is the notation the subject is used to) than the model with stereotypes (since this is a new concept, which the subject is 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 and 4).

In the study students were chosen as subjects. Since they were taught UML on a course, the knowledge of UML was sufficient to understand the given non-stereotyped model. In addition, they were not very familiar with the telecommunication domain. Therefore additional effort was required from them for understanding the stereotypes for this domain. This means that the subjects in this study primarily come from the third type above. The students were not graded for their performance in the experiment, but they were rewarded for the participation in the experiment.

This group of subjects is expected to achieve the smallest improvement from all of the other groups. Therefore, the experiment can also be generalized to professionals, for who the scale of the improvement should be larger than for the sample in this experiment. Some indications on the differences between student subjects and professionals are also given in [20, 21].

3.3. Independent and dependent variables

There is one independent variable in the experiment, the diagram type, with values: S (stereotyped) and N (non-stereotyped)

Understandability of the designs is measured by two dependent variables. The variables are:

I. Total score (NRESP) – the number of correct answers for each subject when asked questions about the design

II. Time (TSEC) – the time (in seconds) which was required to fill in the questionnaire.

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.

3.4. Hypotheses

The experiment tests the null hypothesis. If it is falsified, it would mean that the introduction of stereotypes represented graphically with icons influences the understanding of UML models. Moreover, if the alternative hypothesis is supported, then it would mean that the introduction of stereotypes using icons improves the understanding of UML models. The hypotheses are 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;

3.5. Instrumentation

The main instruments used in the experiment are two comprehension questionnaires that measure the level of understanding of the presented UML models (one for each round). There are 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 allow 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 contain 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 is aimed at checking the correctness of the model, i.e. an inspection-like question.

Subjects are asked to write down the time before starting answering the questions and after completing the questionnaire. The current time is displayed on the wall using a beamer. The measured total time for answering the questionnaire allows 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 are 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 is required to check that the subjects belong to the desired population.

3.6. Experiment operation

Before the main experiment phase, there is a pilot study, which should be operated in a similar way as the main experiment. However, its intention is to validate the experiment objects and identify potential confounding factors of the study.

In the course of the main experiment, there are 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 3.1. In each round each subject gets 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 lecture explains the notion of stereotypes and showing some basic examples of the usage of stereotypes, but it is not meant to introduce the set of stereotypes used in the experiment.

4. Conducting the study

The study was performed in two steps. The first step was a pilot study to examine the context of the study and to determine some of possible confounding factors that could influence the results of the study. The second step was the experiment, which was aimed at hypothesis testing. The results of the pilot study identified a confounding factor, which caused a change of instrumentation in the experiment (as presented in section 5.1).

The experiment was conducted in approximately 2 hours on a sample of 44 students. 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 on the projector during the whole time of experiment. The subjects were given the first comprehension questionnaire (with stereotyped or non-stereotyped model with respect 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. Experiment results

The results of the experiment indicate that the alternative hypothesis H_1 can be supported and that the null hypothesis can be rejected. The results are presented in section 5.2, but some interesting findings, which influenced the main experiment, are taken from the pilot study as described in section 5.1.

5.1 Pilot study results

The pilot study was done on a group of two subjects, who were chosen based on their knowledge of UML and telecommunication domain (see section 3.2). 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. For a subject who was given the stereotyped model in the first round, the time for solving the test 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 – questionnaire (since the questions in both comprehension questionnaires were of the same kind) and the number of correct answers.

From the pilot study it was also found that one of the models was slightly less complicated. There was also an ordering effect in the questionnaire, which made the introduction of stereotypes in the second round easier and intuitional.

The expected result was achieved by the subject representing group 1. The opportunity of having a dialogue with the subject also proved that the subject perceived stereotypes as helpful in understanding. The subject indicated that one of the models was less complex (the same as the subject representing group 2).

As a result of the 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, the subject showed that the set of stereotypes was indeed helpful in understanding the model.

5.2. Experiment results analysis

The experiment was performed on a group of 44 students. After the initial data set reduction it was found that answers from 39 (20 in group 1 and 19 in group 2) subjects could be taken into consideration. Five subjects were removed due to some errors in given documents (two subjects solved two tests for the same system) and personal issues (not handing 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 (NRESP and TSEC) independently, relative times for a correct answer (TSEC/NRESP) and overall subject improvement. Each variable (TSEC, NRESP and TSEC/NRESP) was tested for the normal distribution using the Chi-2 test. And since the test indicated that none of the distributions could be classified as normal, a non-parametric test (Wilcoxon) was used to analyze the data. The test is recommended for this type of design by for instance [16], and its detailed description could be found in [22]. For each analyzed variable a bar plot is used as a presentation of the acquired data.

5.2.1 Number of correct responses. One of the two direct measures – the number of correct responses (NRESP) – allows judging how the introduction of stereotypes influences the understanding of models in terms of accuracy. The influence of stereotypes on the number of correct responses for each subject is summarized in Figure 2.

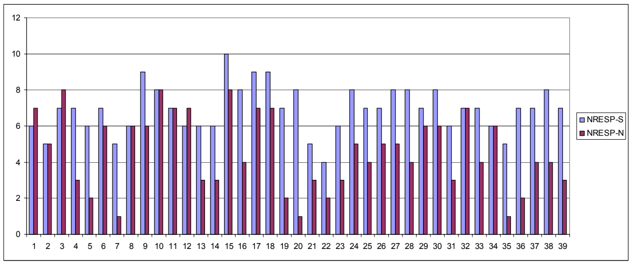

Figure 2 Number of correct answers for each subject. NRESP-S is the number of correct answers for the stereotyped model and NRESP-N is the number of correct answers in the non-stereotyped model.

The basic descriptive statistics indicated that the null hypotheses could be rejected. The results of the analysis are summarized in Table 3.

Treatment	Measure	Value	Percentages
NRESP-S	Mean	6.92	152%
	Standard deviation	1.29	
NRESP-N	Mean	4.56	100%
	Standard deviation	2.01	
Differences NRESP-S – NRESP-N	Mean	2.36	52%
	Standard deviation	1.87	

Table 3 Summary of analysis results for NRESP variable

The higher mean value and the lower standard deviation for the stereotyped model show that the subjects understood the model better and they were more consistent in their answers when the stereotypes were involved. The improvement can be measured in percentages, taking as the 100% the mean value of the number of correct answers for the non-stereotyped model. The value of the improvement is 2.36/4.56*100%= 52%.

The analysis by the Wilcoxon test showed that the null hypothesis can be rejected in favor of the alternative hypothesis with the significance level (p) of less than 0.0001.

5.2.2. Time spent for answering the questionnaire. The analysis of the time spent for answering the comprehension questionnaire (TSEC) allows judging the influence of stereotypes on the time spent for answering the questionnaire. The times acquired from the subjects are summarized in Figure 3.

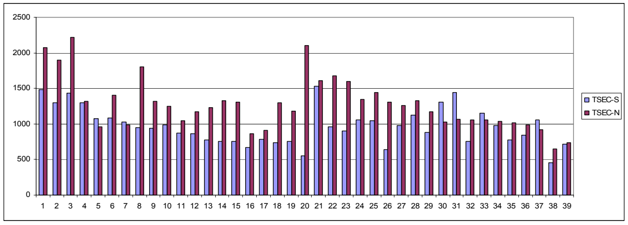

Figure 3 Time spent (in seconds) for answering the questionnaire for each subject. TSEC-S is the time spent for the stereotyped model and TSEC-N is the time spent for the non-stereotyped model.

The analysis with descriptive statistics indicated that there was an improvement. The results of the analysis are summarized in Table 4.

Treatment	Measure	Value (sec)	Percentages
TSEC-S	Mean	967	75%
	Standard deviation	260	
TSEC-N	Mean	1281	100%
	Standard deviation	367	
Differences TSEC-S – TSEC-N	Mean	-315	25%[1]
	Standard deviation	356	

Table 4 Summary of the analysis for the TSEC variable

Since the mean value for the stereotyped model was lower than the mean value for the non-stereotyped model, the average time required understanding the model was shorter. The subjects were also more consistent for the stereotyped model (lower value of the standard deviation). It shows that on average it takes less time to understand the stereotyped model than the non-stereotyped model. It is an interesting observation, since the subjects had to spend some time to understand the stereotypes (and this time was not required for the non-stereotyped model). The average improvement was on the level of 25%. As a basis for calculations, the mean value for the non-stereotyped model was taken (1281 = 100%).

The analysis using the Wilcoxon test showed that the null hypothesis can be rejected in favor of the alternative hypothesis with the significance level (p) of less than 0.0001.

5.2.3. Relative time required for a correct answer. An important correlation between the TSEC and NRESP variables is the relative time for a correct answer (TSEC/NRESP). 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 4 presents the summary of the results.

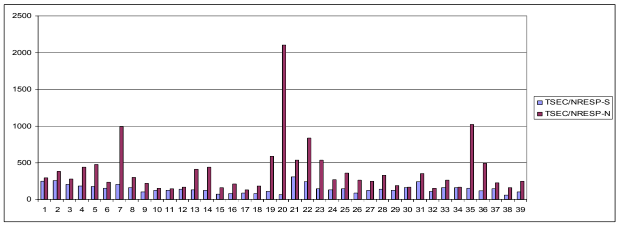

Figure 4 Relative time (in seconds) required for a correct answer. TSEC/NRESP-S is the relative time for the stereotyped model and TSEC/NRESP-N is the relative time for the non-stereotyped model.

The results of the analysis using descriptive statistics are presented in Table 5.

Treatment	Measure	Value (sec)	Percentages
TSEC-S/NRESP-S	Mean	147	38%
	Standard deviation	55	
TSEC-N/NRESP-N	Mean	389	100%
	Standard deviation	355	
Differences TSEC-S/NRESP-S – TSEC-N/NRESP-N	Mean	-242	62%[2]
	Standard deviation	355	

Table 5 Summary of the analysis for the TSEC/NRESP variable

The mean value for the differences indicates that on average the relative amount of time for a correct answer was shorter for the stereotyped model. It means that on average the subjects were more efficient in giving the correct answer for the stereotyped model.

An interesting observation from this variable was that there were no positive differences (c.f. Figure 4). It means that the relative times for a correct answer were better for every subject for the stereotyped model. Since the mean value is also lower for the stereotyped model, then a conclusion could be drawn that on average, the time required for one correct answer was shorter in the case of the stereotyped model. It gives a good indication of improvement since the average number of correct answers was higher for the stereotyped model. The average change was on the level of 62% of improvement in the relative time. Again for the calculations the mean value in the non-stereotyped model was chosen as a basis (100%),

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

[1] In this case the absolute value (314.69) was used in the calculations.

[2] In the course of calculations, the absolute value was taken.

were not identified as outliers in the analyses of the previous variables, because the data was more uneven distributed (larger standard deviation). It was found that the outliers had a significant influence on the analysis for the relative time. The summary table after removing the outliers is as follows:

Treatment	Measure	Value (sec)	Percentages
TSEC-S/NRESP-S	Mean	147	48%
	Standard deviation	56	
TSEC-N/NRESP-N	Mean	307	100%
	Standard deviation	157	
Differences TSEC-S/NRESP-S – TSEC-N/NRESP-N	Mean	-159	52%[3]
	Standard deviation	137	

Table 6 Summary of the analysis of the TSEC/NRESP variable after removing outliers

Removing the outliers resulted in a decrease of the improvement in this analysis compared to the results from Table 5. The improvement in this case was 52%.

The Wilcoxon test showed that the null hypothesis could be rejected at the significance level of less than 0.0001 for both variants of analysis – with and without outliers.

5.2.4. Overall improvements. The highly desired effect of the experiment was that there should be an improvement in both variables (number of correct responses and time) or in either of them (without influencing the other one). A strongly undesired effect would be that there is actually deterioration for both variables, i.e. fewer correct answers in a longer time. Between the extreme cases, there are results where one variable is improved and the other is deteriorated. An analysis of the effect when one of the variables was improved and the other is deteriorated must be done in the context of analysis of a relative time for a correct answer and the analysis of separate variables. Since the results indicated that the relative time was better (for each subject) for the stereotyped models (c.f. previous section), they may be regarded as an overall improvement. The results of the study from the overall improvement perspective are summarized in Figure 5.

The chart indicates that the improvement both in terms of time and number of responses was achieved by 62% of the subjects. Some kind of improvement (including improvements of only one variable) was achieved by 77% (62% + 15%). The improvement of only one variable (while deterioration of the other) was achieved by 23% of subjects. The improvement in at least one of the variables (not counting deterioration in the other) was achieved by 100% of the subjects (62% + 15% + 15% + 8%). There was no situation where the deterioration was achieved in both variables.

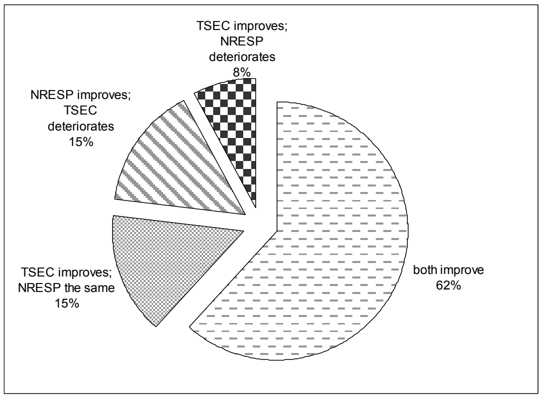

Figure 5 Overall improvement chart

The chart shows that in the majority, the introduction of stereotypes improved the performance of subjects (77%). The remaining 23% must be analyzed in the context of a relative time for a correct answer analysis – TSEC/NRESP. Because there was no deterioration in relative time for any subject, a conclusion can be drawn that the deterioration in the absolute time values were compensated by the number of correct answers (which were much higher for the stereotyped model). In this context, the 23% percent of the subjects, who achieved improvement for one variable and deterioration in the other, achieved the overall improvement. There were no other possible situations, i.e. deterioration in both variables, NRESP the same and TSEC deteriorates; NRESP improves and TSEC stays the same; TSEC stays the same and NRESP deteriorates.

5.3. Threats to validity of the study

As any empirical study, this study has threats to its validity. The threats are grouped as suggested in [16] and presented below.

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 in the second round (for non-stereotyped model). 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 can be seen in the analysis of times for solving the test.

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

[3] In the course of calculations, the absolute value was taken.

other hand, the prior preparation of the objects resulted in an equal complexity of the models (which was seen as a larger threat).

The main external validity threat is that the sample 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 (as described in section 3.2). 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.

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.

5.4. Discussion of results

Firstly, the results show clearly that introduction of stereotypes improves the understandability of UML models in terms of time required to answer the comprehension questionnaires for each model and the number of correct answers in them. Naturally there were some exceptional cases where the results indicated the negative effect (unlike the expected), but these cases formed a small number of all cases.

Secondly, the analysis of overall performance has a large importance from the practical perspective. It showed that a great majority of the subjects achieved an improvement of some kind, without deterioration in other aspects. All subjects acquired some improvement, but some of them decreased the other aspect (other variable).

Thirdly, despite the decrease in one variable, the overall performance of these subjects was positive, since the analysis of the relative time for a correct answer showed that for every subject, the relative time was shorter for the stereotyped model than for the non-stereotyped model.

Fourthly, the results showed a complete lack of the highly undesired effect of deterioration of the performance of the subject in terms of both variables. Although the sample was highly selective, the lack of such effect indicates that in the real-world environment its influence will most likely have the positive effect too.

Finally, the results show that the benefits of using the stereotypes when it comes to understandability are obvious. However, the costs of introduction of stereotypes (including their proper definition, introduction into tools, etc) are not studied. It requires further analysis.

6. Conclusions and further work

The empirical study of the role of stereotypes in understanding of UML models shows clearly that the stereotypes with graphical icons improve understandability. The improvement was achieved in the following three aspects. Firstly, an important measure is the number of correct answers in the tests checking the level of understanding. The improvement achieved in this category was 52%. Secondly, the improvement in time required for solving the test was on the level of 25%. Finally, the improvement in a relative time for a correct answer was achieved on a level of 62%. A notable fact is that an improvement of some kind was achieved by every subject in the study. Since the sample chosen for the experiment was a representative of the worst-case scenario sample, the results for professionals (who do not belong to the worst-case scenario group) are expected to be even more positive. Their background (knowledge of UML but not of the domain) means that the introduction of stereotypes could deteriorate the understanding of the models (because of the introduction of new, previously unknown elements).

The study presented in the paper is an introductory study. The further research in this field should be a case study on usage of stereotypes in their natural context, for instance a specific domain or 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 study is also to be performed with objects modified to contain stereotypes represented as guillements and not icons. The results of the study are a basis for evaluation of the cost of introduction of stereotypes of different kinds. Such analysis is to be done in the next step of our research.

References

[1] O. Object Management Group, *Unified Modeling Language Specification v. 1.4*, 2002.

[2] C. Atkinson, T. Kuhne, and B. Henderson-Sellers, "Stereotypical Encounters of the Third Kind," In the Proceedings of UML 2002, Dresden, 2002, Springer-Verlag, pp. 100-114.

[3] R. Wirfs-Brock, B. Wilkerson, and L. Wiener, "Responsibility-driven design: Adding to your conceptual toolkit," *ROAD*, vol. 2, pp. 27-34, 1994.

[4] M. C. Otero and J. J. Dolado, "An Initial Experimental Assessment of Dynamic Modeling in UML," *Empirical Software Engineering*, vol. 7, pp. 27-37, 2002.

[5] L. C. Briand, et al, "An Experimental Comparison of the Maintainability of Object-Oriented and Structured Design Documents," *Empirical Software Engineering*, vol. 2, pp. 291-312, 1997.

[6] M. Gogolla and B. Henderson-Sellers, "Analysis of UML Stereotypes in the UML Metamodel," In the Proceedings of UML 2002, Dresden, 2002, Springer-Verlag, pp. 84-99.

[7] L. Kuzniarz and M. Staron, "On Practical Usage of Stereotypes in UML-Based Software Development," In the Proceedings of Forum on Design and Specification Languages, Marseille, 2002, FDL, pp.

[8] A. Uhl and H. Lichter, "A UML Variant for Modeling System Searchability," In the Proceedings of Object Oriented Information Systems, Monpellier, 2002, Springer-Verlag, pp. 199-211.

[9] L. Kuzniarz and J. Ratajski, "Code generation based on a specific stereotype," In the Proceedings of Information Systems Modeling, Roznov, Chech Republic, 2002, MARQ, pp. 119-128.

[10] Rational, "Rose Data Modeling Profile," accessed on

[11] O. Object Management Group, "UML Profile for Schedulability, Performance and Time," ptc/02-03-02, www.omg.org, accessed on 2003-09-20

[12] O. Object Management Group, "UML profile for CORBA," www.omg.org, accessed on 2003-09-20

[13] D. Gornik, "UML Data Modeling Profile," Rational Corp., Whitepaper TP162 05/02, 2002.

[14] R. Wirfs-Brooks, B. Wilkerson, and L. Wiener, "Responsibility-driven design: Adding to your conceptual toolkit," *ROAD*, vol. 2, pp. 27-34, 2003.

[15] R. Wirfs-Brock, "Stereotyping: a technique for characterizing objects and their interactions," *Object Magazine*, vol. 3, pp. 50-3, 1993.

[16] C. Wohlin, P. Runeson, M. Host, M. C. Ohlsson, B. Regnell, and A. Wesslen, *Experimentation in Software Engineering: An Introduction*. Boston MA: Kluwer Academic Publisher, 2000.

[17] M. Staron, "Experiment on the role of stereotypes in UML based software development - materials," http://www.ipd.bth.se/mst/Experiment/index.html, accessed on 2003-09-03

[18] O. Object Management Group, *Object Constraint Language (OCL) Specification v. 1.4*: OMG, 2002.

[19] L. Kuzniarz, M. Staron, and C. Wohlin, "An experimental evaluation of the role of stereotypes in understanding of UML models," to be published as Technical report, Blekinge Instiute of Technology, Ronneby, Technical report Technical report, 2003.

[20] M. Höst, B. Regnell, and C. Wohlin, "Using Students as Subjects - A Comparative Study of Students and Professionals in Lead-Time Impact Assessment," *Empirical Software Engineering*, vol. 5, pp. 201-214, 2000.

[21] W. Tichy, "Hints for Reviewing Empirical Work in Software Engineering," *Empirical Software Engineering*, vol. 5, pp. 309-312, 2000.

[22] S. Siegel and N. J. Castellan, *Nonparametric statistics for the behavioral sciences*, 2nd ed. New York: McGraw-Hill, 1988.

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). All materials can be found in [16].

Artefact set A-S

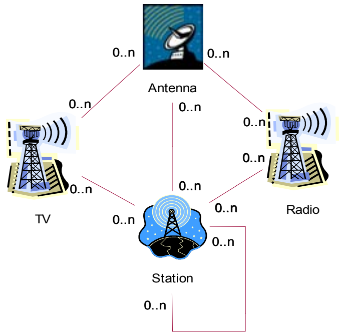

Figure 6 Excerpt from class diagram in artefacts set A-S (about 50% of the diagram)

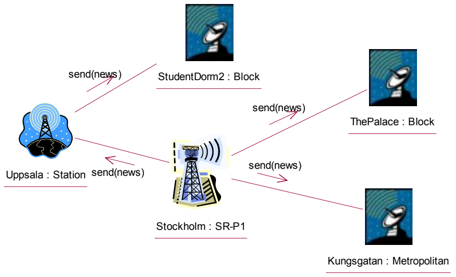

Figure 7 Excerpt from the collaboration diagrams from artefacts set A-S (around 20% of the diagram).

Artefact set A-N

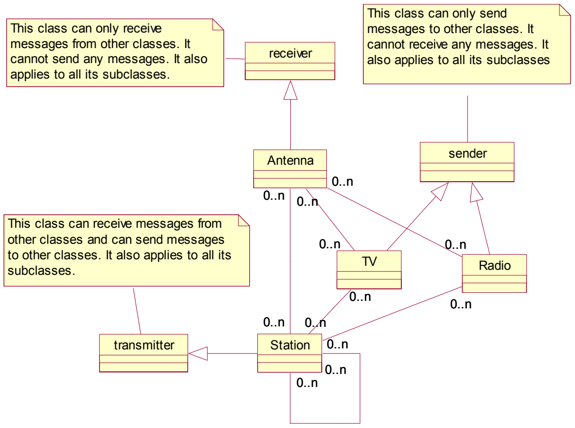
This class can only receive messages from other classes. It cannot send any messages. It also applies to all its subclasses.
receiver
This class can only send messages to other classes. It cannot receive any messages. It also applies to all its subclasses
Antenna
sender
This class can receive messages from other classes and can send messages to other classes. It also applies to all its subclasses.
TV
Radio
transmitter
Station
0..n
0..n
0..n
0..n
0..n
0..n
0..n
0..n
0..n
0..n
0..n
0..n

Figure 8 Excerpt from the class diagram in artefacts set A-N (about 50% of the diagram).

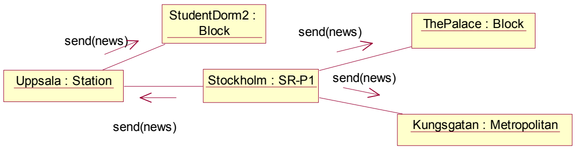

Figure 9 Excerpt from the collaboration diagram from set A-N (about 20% of the diagram).