Students as Study Subjects in Software Engineering Experimentation

Ludwik Kuzniarz, Miroslaw Staron, Claes Wohlin
Department of Software Engineering and Computer Science,
Blekinge Institute of Technology
Box 520 Soft Center
SE-372 33 Ronneby, Sweden
(lku, mst, cwo)@bth.se

Abstract

Experimentation using different types of subjects is an important issue in empirical software engineering. In particular, the use of students as subjects is many times questioned. This paper addresses this issue by dividing subjects into four types, where one type can be viewed as a worst case. It is discussed how students could be viewed as suitable subjects in the worst case. In particular, if it is possible to reject the null hypothesis then it is highly likely that the outcome is also valid for other types of subjects. An example experiment is presented that uses students as subjects and fulfill the criteria for being the worst case. It is concluded that students are good subjects under certain circumstances and that these types of experiments provide an opportunity for using students as subjects.

1. Introduction

Empirical methods in software engineering gain popularity largely due to the fact that the obtained results [1, 2] come from the "real world". The results obtained in the empirical way with the presence of human subjects are based on the human factor and give an estimation of the "real world" effect of the experiment object. The best estimation of the effect can be obtained by examining the whole population or the most representative sample from the sample frame [3]. In most cases the most representative sample is taken from industry professionals [2, 4], who have roughly the same experience as the whole population working in their field. The difficulty with these types of subjects is that they are "expensive" in terms of man-hours. It results in managers being reluctant to let their employees take part in studies instead of doing their primary tasks. To overcome the problem the experimenters enlarge the population to the students of the field and then take advantage of the strategies called purposive sampling or convenience sampling [3]. Although the results can be questioned sometimes, i.e. if the students really belong to the set of results of the given population, the advantage of these experiments is their low cost, since the students take part in the study without gratification. In some cases the experiments with student subjects provide results which cannot be easily generalized to the whole population, but the significance of the experimentation in software engineering requires (and also partially excuses [1]) the usage of such subjects. This paper presents an idea and an example that there are certain situations in which the usage of students as subjects are particularly beneficial. Thus, the main contribution is the identification of a situation when students may be better subjects for empirical studies than they are in general. Moreover, an example of study benefiting from this situation is presented briefly.

This paper examines the available types of subjects for a certain class of experiments – aiming at a comparison of two methods. Since the knowledge of methods that the subjects have varies, there are certain problems with generalizing the results to a broader population. Therefore, it is important to classify the subject into one of the possible type of subjects and then generalize the results based on the classification of the subject.

The material presented in this paper should not be treated as a definitive solution for a sampling strategy, but rather as a suggestion for choosing samples in certain situations based on their background and then generalizing the results based on that. A more complete coverage of the issue will be possible to present if the material presented in the paper is verified in a way outlined in further works section. The paper presents an example experiment – one of the situations in which student subjects may be particularly suitable.

The paper is structured as follows: firstly, a discussion of the possible types of subjects is presented; secondly, a sample experiment in which this strategy was used is outlined; and finally, the relationship between the general discussion and the specific experiment is discussed.

2. Types of subjects

In the experiments aiming at comparison of two methods – call them method 1 and method 2 – the subjects can be divided into four main types based on the knowledge of the methods they have. It is here assumed that we would like to evaluate whether method 2 is an improvement in comparison to method 1. These types are:

1. A subject has worse knowledge of method 1 than knowledge of method 2,
2. A subject has an equal knowledge of methods 1 and 2,

3. A subject has better knowledge of method 1 than knowledge of method 2, and
4. A subject has no knowledge of either method 1 or method 2.

The first type of subjects are subjects who may work with method 2, but do not use method 1 in their work. The subjects are usually industry professionals or students at study programs that involve method 2. The results obtained from these subjects would probably be much higher – better – for method 2, since they use the method they are already familiar with. They may not be the best types of subjects since the results could be confounded by the knowledge of method 2. Other types of subject may show a negative effect even if a sample of this type shows an improvement.

Subjects of the second type are subjects that already use method 1 and method 2. Since they know both methods the results obtained from them may not be generalized to other types of subjects. If method 2 is better, they would show the improvement, but this does not mean that all types of subjects would perform similarly. In particular, the results obtained from subjects of the third type may be lower.

The third type of subject is the worst-case situation from the perspective of introduction of the new method 2. Given the knowledge of method 1 the subject should be able to better use method 1 than method 2. This is the situation in which the introduction of the new method 2 could result in a negative effect for the subjects after the introduction of the method 2 if they used the method 1 so far, and which they are acquainted with. The introduction of the new method 2 requires an additional effort to learn it, whereas this effort would not be required if method 1 is used. This could have a negative effect on the introduction of the new method 2. However, it becomes very interesting if it is possible to show a positive effect of the introduction of method 2 given the background of the subjects.

The last – fourth – type of subjects can provide similar results of the experiment as the second type since both types of subjects have roughly the same amount of knowledge of both methods. If the effort put into the learning of the new method for both types of subjects could be estimated to be equal for both methods the types would show similar results. However, the subjects of type two and four should not be mixed together, because the difference in understanding of the method may be reflected in the time required to understand the methods and most probably will be different for subjects of type 2 and type 4.

From the analysis presented above it can be deducted that the most interesting type of subject from the perspective of the evaluation of the improvement after introducing the new method 2 is the subjects of the third type. They are the "worst-case scenario" type of subjects, because the introduction of the new method requires some learning effort for them in comparison to the alternative method, which they are already familiar with. As a result they are the most probable subjects that would not show a positive effect when introducing the new method 2. Thus, they are the sample for which the hypotheses are tested in the most rigorous way (as suggested in [4]). If they are able to give positive results, then it is highly likely that other types of subjects will have positive results. Naturally, if they give a negative result (deterioration), then perhaps the results of the experiment cannot be generalized for the population including this type of subjects and the new method provides benefits only for a narrower population. This strategy is one of the possible sampling strategies and it provides a lower approximation of the size of the improvement after introduction of the new method 2. Although there is no guarantee that the results of the other types of subjects will not be lower than the obtained (due to the "human factor" in the study), it gives a good estimation of the size of the improvement in case of positive results. An upper approximation of the size of the improvement could be obtained by choosing another type of subjects (i.e. the subjects that belong to the first type).

A special kind of subjects which can be classified as the third type, are students who take part in courses where they practice using method 1. Depending on the course they attend, they get some experience in using method 1 while they do not get any experience with using method 2. Naturally, they must be introduced to the new method 2 before the experiment, but usually this introduction is not the same as the practical classes when they practice using method 1.

The professionals in industry may be acquainted with both methods. Then they belong to the second type of subjects. However, it is important to capture the knowledge of the two methods through a survey as part of the empirical study to understand to what type the subjects belong.

3. Summary of experiment

As an example of using the third type of subjects in an experiment involving student subjects a controlled experiment of UML-descriptions is presented. The example experiment presented in the paper is described in four sections showing separately the design of the experiment, the subjects in the study, the conduction of the experiment and outline of the results.

3.1. Experiment design

The study presented in this paper is the empirical evaluation of the roles of stereotypes in UML-based software development (for a complete description of the study, please consult [5]). The study is designed in the form of an experiment which involved comparison of two ways of designing the software. One of them incorporated only the standard Unified Modeling Language (UML, [6]) constructs to design the software and the other included UML stereotypes to express the same design. The comparison between the two ways was done in terms of understandability of the designs. 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* 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 [7]. The treatments are model types, with two possible values - stereotyped model and non-stereotyped model (which correspond to methods 2 and 1 from the previous section respectively). 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 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.

The set of experimental objects (fully presented in [5, 8]) consists of four artifacts as follows:

- 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 and different types of antennas) and a corresponding object diagram describing one of possible situations (i.e. sending a news program across a country).

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) and a corresponding object diagram describing one possible situation of using the network (i.e. making phone calls at a given time).

The best solution would be to have the same models (A) in both rounds, but because of the learning effect in the second round it cannot be done (subjects could understand the model better in the second round simply because they examine the model for the second time). So another set of artifacts (B) is introduced, to avoid it.

The sets of experiment objects, which use stereotypes, are based on a telecommunication profile, introduced and fully defined in [5]. 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 [9]). Considering the classification of stereotypes presented in [10], the stereotypes are added to the elements at the model level, but their semantics concern 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.

Table 1. Summary of experimental design.

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

The subjects taking part in the study are divided into two groups, where members of the groups get different treatment.

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.

The experiment was designed to test the hypotheses. Falsification of the null hypothesis in favor of the alternative hypothesis (H_1) would mean that the introduction of stereotypes improves the understanding of UML models. 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;

For each subject, two measures are taken corresponding to the dependent variables of the study. The analysis of the results is done in three ways. Firstly, the results are analyzed with respect to each variable, which gives an overview of the improvement with respect to two factors – time and number of correct answers. Secondly, the variables are tied together; the relative time for a correct answer is computed and analyzed in the same way as the separate variables. Finally, the overall improvement is analyzed. The improvement is considered within a set of results. The most desired results are when a subject achieves an improvement in both variables, while the most undesirable result is when a subject achieves a negative effect for both variables. All the possibilities (i.e. with improvement of one variable and deterioration of the other) are between these two.

3.2. Subjects

In case of the example experiment, the students that took part in the study attended a UML course during which they get experience with UML but they had no contact with either the mobile telephony or TV/Radio transmissions domains. Naturally, they could have some experience with the domain outside of the university or by attending some joint programs. Therefore, the background questionnaire contained questions checking whether the students had any experience in designing systems in these domains, and their previous experience of UML and stereotypes. In this particular experiment they did not. The students were taking part in an object-oriented software development course, which consisted of theoretical lectures, practical exercises and individual projects. Since they were taught UML in a course, the knowledge of UML was sufficient to understand the given non-stereotyped model. They took the practical classes during which they used UML for software development, but were introduced to the stereotypes only during a short lecture before the experiment. The difference of experience between UML and stereotypes allows classifying these subjects to the third type of subjects. They were expected to give negative results (support the null hypothesis).

3.3. 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 [5]).

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 project group. Both groups were in the same room at 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 a background questionnaire to fill in.

3.4. Outline of the results

The highly desired effect of the experiment is improvement in both variables (number of correct responses and time) or in either of them without influencing the other one. A strongly undesired effect is a negative effect for both variables, i.e. fewer correct answers in a longer time. Between the extreme cases, there are results in which 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 the separate variables. The results of the study from the overall improvement perspective are summarized in Figure 1, which shows the percentage share of different types of results. The figure uses variable names as discussed in section 3.1.

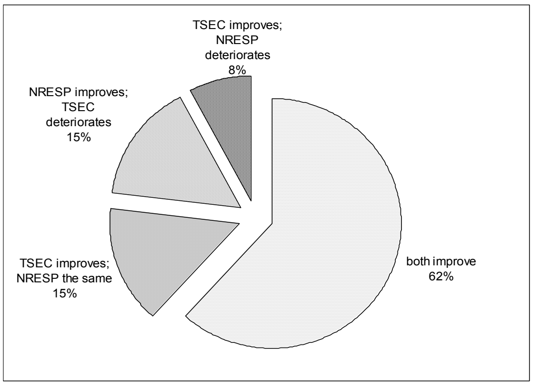

Figure 1. Overall improvement chart.

The chart indicates that an improvement for using stereotypes 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. 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 an overall improvement looking at the relative time for a correct answer. There were no other, although possible, situations of the changes in the variables:

- deterioration in both variables,
- deterioration of TSEC with no change in NRESP,
- improvement of NRESP with no change in TSEC, and
- deterioration in NRESP with no change in TSEC

The full analysis of the results of the experiment is presented in [5].

4. Discussion

The students can be found to be the very appropriate subjects in software engineering. They are accessible on courses at universities, which usually consist of a large number of students[1]. The use of students is of course always a validity threat. However, the objective of student experiments is often to target a specific issue and try to avoid confounding factors. One aim being that the relative difference between two treatments becomes comparable to the real situation. Moreover, it should also be remembered that the presentation here is primarily aimed at identifying situations when students may be better as subjects than they are in certain other situations. It is outside the scope of this paper to discuss the use of students as subjects in general. For the actual study presented, it is clear that the students have a much better knowledge of the use of UML without stereotypes than with stereotypes. This is similar to the situation in industry when one may want to introduce stereotypes in an organization that has previously used UML without stereotypes.

The main advantages of using students are their high availability, a possibility to redo the experiment on a different sample, low cost and the possibility to provide students with additional knowledge of the objects of the study. On the other hand, the experiments using students as subjects have multiple perspectives, as analyzed in [11].

Taking into consideration a special set of experiments, which consists of such experimental designs as paired-comparison designs, factorial designs, etc. aimed at comparison of methods the students as subjects are even the most desirable type of possible subjects, since they provide a good lower approximation for the size of improvement (or deterioration, depending on the experiment). In other words, given that they know the control method beforehand it is possible to evaluate the introduction of a new method. Since they know one of the methods from a university course, their knowledge is more directed than the knowledge of the professionals, who may have some other background experience in similar methods. This is an advantage, because the experiment may be designed in a way that takes into account this directed knowledge. The more the experimenter uses this technique, the more it is likely that the results of other types of subjects will not be below the results of the most desirable type of subjects.

However, this sampling strategy has also some drawbacks, among which the most important one seems to be that the experiment does not give the definite scale of the improvement. It is caused by the fact that the extreme case sampling strategy, with subjects from only one extreme case (third type), provides only the lower or upper approximation of the results. It does not contain results for the “average” subject. In case of the students, using them as subjects may not give results that could be expected from the industry professionals, but to some extent the industry professionals (assuming they belong to other types of subjects) should perform even better than the students, or at least not worse (for an example, please consult an example of another experiment [2] where students’ performance is compared to the industry professionals’ performance).

In the presented experiment, students (belonging to the third type of subjects) showed that the introduction of the new method (stereotypes) into the designs improved the understanding of UML models. Since this type of subjects, which was expected to show a negative effect showed an improvement, it is highly likely that other types of subjects will show an improvement as well. Moreover the improvement is likely to be even higher than the improvement of the investigated sample.

5. Conclusions

The paper elaborated on the subject of using students as extreme case sampling strategy in software engineering experimentation. The paper proposes a classification of the possible types of subjects in the experiments aiming at comparison of two methods. Along with the classification, the paper shows how to generalize the results to the whole population based on the classification of the subjects. An experiment in which the proposed classification and generalization has been used is briefly described. The explanation of the relationship between the general classification and the concrete experiment illustrated a way in which an experimenter can benefit from using students as subjects in such kind of experiments and results expected from the

[1] As compared to industry professionals, who usually work in smaller teams and are less accessible.

whole population based on the results from the student subjects in this case. The research that led to the presented material poses a research question of how experiments using students as subjects can provide general results. A further investigation of this research question is possible based on an industrial study – a re-run of the experiment using industry professionals as subjects and comparing the results. However, it is outside the scope of the paper to evaluate students as study subjects in general. The main contribution is the identification of one of the situations in which the students as study subjects are better than in other situations.

Although some indications on the differences between student subjects and professionals were given in [2], the results presented there are restricted to a special kind of experiment – based on a method which did not required additional studies (the Analytical Hierarchy Process – AHP) and the assumption in that study concerned a topic, which "is not taught in any courses, neither at the university nor in industry" [2]. In summary, the comparison in [2] was focused on general understanding of challenges in software engineering projects. As a further research, a re-run of the experiment presented briefly here with industry professionals as subjects is planned to verify the expectations discussed in the paper and to provide empirical data on this.

References

1. Tichy, W., *Hints for Reviewing Empirical Work in Software Engineering.* Empirical Software Engineering, 2000. **5**(4): p. 309-312.
2. Höst, M., B. Regnell, and C. Wohlin, *Using Students as Subjects - A Comparative Study of Students and Professionals in Lead-Time Impact Assessment.* Empirical Software Engineering, 2000. **5**(3): p. 201-214.
3. Robson, C., *Real World Research*. 2 ed. 2002, Oxford: Blackwell Publishing.
4. Kitchenham, B.A., et al., *Preliminary Guidelines for Empirical Research in Software Engineering.* IEEE Transactions on Software Engineering, 2002. **28**(8): p. 721-734.
5. Kuzniarz, L., M. Staron, and C. Wohlin, *An Empirical Evaluation of the Role of Stereotypes in UML-based Software Development*. 2003, to be published as a Technical Report, Blekinge Institute of Technology: Ronneby.
6. Object Management Group, O., *Unified Modeling Language Specification v. 1.4*. 2001.
7. Wohlin, C., et al., *Experimentation in Software Engineering: An Introduction*. 2000, Boston MA: Kluwer Academic Publisher.
8. Staron, M., *Experiment on the role of stereotypes in UML based software development - materials*. 2003.
9. Kuzniarz, L. and M. Staron. *On Practical Usage of Stereotypes in UML-Based Software Development*. in *Forum on Design and Specification Languages*. 2002. Marseille: FDL.
10. Atkinson, C., T. Kuhne, and B. Henderson-Sellers. *Stereotypical Encounters of the Third Kind*. in *UML 2002*. 2002. Dresden: Springer-Verlag, pp 100-114.
11. Carver, J., et al. *Issues in Using Students in Empirical Studies in Software Engineering Education*. in *9th International Software Metrics Symposium*. 2003. Sydney, Australia: IEEE Computer Society, pp 239-251.