M. Höst, C. Wohlin and T. Thelin, "Experimental Context Classification: Incentives and Experience of Subjects", IEEE Conference Proceedings International Conference on Software Engineering, pp. 470-478, St. Louis, USA, 2005.

Experimental Context Classification: Incentives and Experience of Subjects

Martin Höst
Dept. of Communication Systems,
Lund University, Sweden
martin.host@telecom.lth.se

Claes Wohlin
Dept. of Systems and Software Engineering,
School of Engineering,
Blekinge Institute of Technology, Sweden
claes.wohlin@bth.se

Thomas Thelin
Dept. of Communication Systems,
Lund University, Sweden
thomas.thelin@telecom.lth.se

ABSTRACT

There is a need to identify factors that affect the result of empirical studies in software engineering research. It is still the case that seemingly identical replications of controlled experiments result in different conclusions due to the fact that all factors describing the experiment context are not clearly defined and hence controlled. In this article, a scheme for describing the participants of controlled experiments is proposed and evaluated. It consists of two main factors, the incentives for participants in the experiment and the experience of the participants. The scheme has been evaluated by classifying a set of previously conducted experiments from literature. It can be concluded that the scheme was easy to use and understand. It is also found that experiments that are classified in the same way to a large extent point at the same results, which indicates that the scheme addresses relevant factors.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics

General Terms

Experimentation, Management

Keywords

Experimentation, Subject experience, Subject motivation

1. INTRODUCTION

Empirical methods as a means for conducting software engineering research have been stressed [27, 29, 25, 12]. One such method is controlled experiments. These experiments are often conducted in a laboratory setting, which is a challenge from an external validity point of view. Moreover, many experiments use students as subjects. This is many times questioned in reviews [26]. One example when the results from using students as subjects are similar to using professionals was reported in [9]. However, it is important to understand under which circumstances the results from using students become similar to using professionals or provide other interesting insights.

Understanding is crucial to ensure progress in experimentation in software engineering. In particular, it is important to understand whether it is not a too simplistic view to only look at students versus professionals. In a recent paper, it is observed that students conducting projects with industrial customers seem to behave more similar to professionals than students performing stand-alone experiments within a course [3]. This illustrates that a better understanding is needed of additional factors than just comparing students versus professionals. One potentially important factor is the motivation of the subjects regardless of any other division of subjects into categories, for example students and professionals.

Moreover, several studies, which seemingly are similar, present contradictory results. Basically, this indicates that the research community has not managed to capture the relevant underlying explanatory variables satisfactorily, i.e. we have not captured the context "good enough". This situation is of course very unfortunate since it limits the value of individual studies and it makes it hard to distinguish between experiment results that are based on chance and results that are relevant for a specific context. Thus, there is a need to better capture experiment context and understand better how results from different studies can be combined [11].

As mentioned above, a typical review comment when conducting experiments with students is that the results are not valid or even interesting due to the use of students. It is obviously easier to conduct experiments with students than involving professional practitioners in a study. Instead of just ignoring the opportunity of involving students in research studies, we claim that we have to understand when results from one type of subject could be generalized to another type of subject. The objective of this article is to contribute to this understanding by introducing and studying an additional factor namely motivation or incentive for conducting an experiment seriously, which makes the results from the experiment more trustworthy and hence interesting.

Traditionally, experience is often reported in software en-

ICSE'05, May 15–21, 2005, St. Louis, Missouri, USA.

gineering studies. There are two basic categories of subjects that often are mentioned: students and professional practitioners (e.g. [9]). It should, however, be noted that this simple categorization in many cases is too simplistic to be useful. In many cases the subjects in a study are not homogeneous and students may for example have industrial experience, as discussed by Sjøberg et al. in [23]. In addition, a well-motivated subject may perform better in an experiment than a poorly motivated subject, as discussed in [2]. A professional developer may be very motivated when the study is done with their real artifacts, but less motivated when artifacts are produced for the sake of the experiment.

This article addresses the problem by introducing a classification scheme that is not merely focused on the experience of subjects, e.g. whether they are students or working as professional engineers. The scheme is intended to be generally useful for all types of software engineering experiments when using humans as subjects, although a specific type of experiments is used here to evaluate the usefulness of the scheme. The scheme still includes the experience as a factor to consider, but the scheme is also focused on the incentive of the subjects of the study. That is, we argue that the validity of an empirical study is affected by the motivation of the subjects in the study. It would of course be possible to propose a more complicated scheme and including more factors than experience and incentive. It was however not seen as realistic to introduce more factors in this step. The scheme is evaluated by classifying a number of experiments on reading techniques. The classification is done both by the authors of this article based on reading the articles, and by the authors of the different experiments.

The remainder of the article is structured as follows. Section 2 presents the objective and applied methodology. The details of the classification scheme are presented in Section 3, and the evaluation using a set of experiments is discussed in Section 4. The article is concluded with a discussion in Section 5 and conclusions and some further work in Section 6.

2. OBJECTIVES AND METHODOLOGY

The objective of this article is to provide a classification scheme that can be used to decide whether the context factors of different controlled experiments have been the same. This, in turn, could be used when the transferability of results from different studies is evaluated. An iterative approach was used when the scheme was developed. The research was carried out in the following steps: planning and definition of initial classification scheme, execution, analysis and updating the initial scheme.

2.1 Planning and definition of an initial classification scheme

A first version of the scheme was derived. This version was with a few exceptions similar to the final version, which is presented in Section 3.

In order to prepare for the execution of the study, a questionnaire, which could be sent to authors of experiments, was developed. It consisted of an introductory text motivating the need for this type of classification and the participation of the reader in an evaluation of the scheme, a description of the (initial) classification scheme, and simple instructions to the participant for classification of an experimental study of him/her.

In order to evaluate the classification scheme, research articles resulting from perspective-based reading (PBR) experiments were used. 13 PBR experiment articles were found in the time frame from 1995 to 2002. This set of studies was chosen because it constitutes one of the largest sets of experiments on one single topic.

The research papers describing PBR experiments were identified by searching in the electronic library system at Lund University[1]. The library system consists of over 300 databases, and includes, for example, IEEE on-line, Science Online and ACM. Furthermore, the system consists of over 11,000 journals and conference proceedings, including Kluwer, IEEE, IEE, Elsevier, Wiley among others. The following search keys were used: "perspective-based reading" *or* "perspective-based inspection" *or* "reading techniques" After reducing duplicate experiments, and papers that do not describe PBR or do not describe an experiment, eight papers were left. In addition, two ISERN[2] reports and three theses[3] that we know of from personal contacts were also included. This procedure resulted in 13 articles describing PBR experiments in the time period.

During the definition, a draft of the scheme was presented and discussed at an ISERN meeting where some of the researchers that later were involved in the study participated. Comments from this discussion were used when the initial version was developed.

2.2 Execution

In order to evaluate the scheme with respect to understandability and usability for researchers in empirical software engineering it was sent via email to the first authors of the selected articles. The authors were asked to classify their articles and send the result of the classification back to the authors of this article. During the same time the authors of this article also classified the selected articles.

A standard procedure in questionnaire-based studies is reminding the participants after a certain time. This was also done in this study, and after this the response rate was 100%.

2.3 Analysis and updating the initial scheme

The scheme was evaluated by comparing the two different classifications with each other. When the results from the classifications were the same this was interpreted as indicating understandability of the schema. In the same way, when the classifications were different we investigated whether this could be because of low clarity or low usability of the scheme or due to that the required information was not clear in the published articles.

The two classifications were compared by analyzing for what articles there were differences. The differences between the classifications were also assessed quantitatively by calculating the Kappa value (e.g. [22]).

The Kappa value (K) can be used to assess the agreement when a set of raters classifies a set of objects into a set of classes, and it is calculated according to the following:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

[1]http://www.lub.lu.se/headoffice/elininfo.shtml
[2]http://www.iese.fhg.de/ISERN/
[3]two Ph.D. theses and one thesis corresponding to a higher degree than Ph.D.

Table 1: SPICE Software Process Assessment Kappa benchmark.

Kappa value	Strength of agreement
$K \leq 0.44$	Poor
$0.44 < K \leq 0.62$	Moderate
$0.62 < K \leq 0.78$	Substantial
$K > 0.78$	Excellent

where $P(A)$ denotes the proportion of times the raters agree and $P(E)$ denotes the expected agreement that would be present by chance if all ratings were made randomly. For a detailed description of how to calculate $P(A)$ and $P(E)$, refer for example to [6] (for two raters, as in this article) or [22] (for two or more raters).

A calculated Kappa-value is interpreted according to the following:

- $K = 1$: Complete agreement
- $0 < K < 1$: In this case there is an agreement and a higher value indicates a better agreement.

  There exist different schemes for interpretation of Kappa-values. For example, Hartmann suggested in 1977 that values larger than 0.6 should be considered good (e.g. [6]). In the same year, Landis and Koch suggested a more detailed benchmark stating, for example, that $0.21 \leq K \leq 0.40$ should be interpreted as "fair" agreement, $0.41 \leq K \leq 0.60$ should be interpreted as "moderate" agreement, and that $0.61 \leq K \leq 0.80$ should be interpreted as "good" agreement (e.g. [6]). In 1999 El Emam [6] presented a benchmark based on SPICE assessments as described in Table 1.

  In this article it was, based on the above presented benchmarks, decided to interpret Kappa-values larger than 0.6 as a "good-enough" agreement.
- No agreement, i.e., the agreement, that is present, is as could be expected by chance, i.e. $P(A) = P(E)$, or less, i.e. $P(A) < P(E)$.

In this article the raters, objects and classes were as follows:

- There were two different raters, or groups of raters, defined. The authors of the selected experiment articles carried out one rating. The other rating was carried out by the authors of this article.
- The objects in this study were the published experiment articles.
- Classification was carried out with respect to two different sets of classes, i.e. the two different dimensions (incentives and experience) of the classification scheme. One Kappa-value was determined for each dimension.

The results of the calculations are presented in Section 4.

The scheme was updated based on internal discussions that were held during the classification that was done by the authors of this article. The intention was also to update the classification scheme if the result indicated that the scheme was unclear. When differences between the two classifications were found, an analysis was carried out to understand the reason for the difference. In some cases the main author of the published article was contacted in order to clarify different understandings.

This procedure resulted in the scheme that is presented in this article. In Section 3 the resulting scheme is presented. It is, as described above, not exactly the same scheme as was used during the evaluation according to Step 2. The minor changes that were introduced according to Step 3 are presented in Section 3.4.

3. CLASSIFICATION SCHEME

According to the proposed scheme, classification is to be carried out with respect to the following two orthogonal factors:

- incentives
- experience of subjects.

We propose that experiments should be classified with respect to both of these factors. The factors are explained below.

3.1 Incentives

We argue that the validity of a study is affected by the motivation of the subjects. Consider, for example, an inspection experiment where a design document is inspected. It is not reasonable to believe that the same faults will be found if the only usage of the fault record from the inspection session is for analysis by the researcher, as in the situation when the faults actually are corrected and removed from the product. If a reviewer misses a fault in the first case this is probably not seen as a large problem by the reviewer, while it could be seen as a problem in the latter situation. In the latter case the reviewer knows that his/her work as reviewer will affect the quality of a product that is important for the developing organisation and for the customer of the organisation.

The following classes of incentives are proposed:

- I1: Isolated artifact
- I2: Artificial project
- I3: Project with short-term commitment
- I4: Project with long-term commitment

The classes are further described below.

3.1.1 Isolated artifact

This denotes the situation when a single object of study is chosen for the experiment. The subjects should not consider any relationship to any additional supporting material, and the subjects have in most cases no prior knowledge of the studied object. The object may, for example, be a requirements specification developed by the researcher or a document that is borrowed from an industrial organization.

An example of a study that is classified in this category is when a requirements specification is inspected in an inspection experiment. The participants have not seen the specification before and are given different techniques to inspect it. This class of experiment has been carried out with students as subjects and with engineers with several years of experience as subjects. The participants try to find as many faults as possible and their motivation consists of their will

or pride to find as many faults as possible and to do a good job as subject in the research study. Students may also be motivated by the grade that is given in a university course and professional engineers may be motivated by following a certain Code of Practice. The object of study denotes a piece of information without any additional material. For example, inspection of a requirements specification as described above corresponds to the situation of an isolated artifact. On the other hand, inspection of test cases based on the requirements of the system corresponds to a more complex situation (artificial project).

3.1.2 Artificial project

This denotes the situation when the subjects have to consider relationships to supporting material in the study. The subjects have typically no prior knowledge of the artifacts that they are working with. They may, for example, be working with a test specification and have access to a requirements specification as supporting documentation. The object and supporting material may be developed by the researcher or a complete set of documents from one project that is borrowed from an industrial organization.

Experiments that are classified in this category present a more complete environment for the subjects than in the isolated artifact case. However, as in the isolated artifact case, no more development of the artifacts is carried out than is necessary for the experiment. Another example of this type of experiment could be to compare different coding techniques. Subjects are given a requirements specification and a design and are asked to implement part of the system. After that, the system can be tested according to the requirements specification. In this case the object of study is the developed code using different coding techniques, and the requirements specification and the design are supporting material.

3.1.3 Project with short-term commitment

This denotes the situation when a real project is considered and the major objective of the subjects is not only to participate in the experimental study. The commitment of the participants to the project is limited to the time of the project. The subjects are motivated to perform the study because they, within the project, are affected by the results of the study. For example, if the study is concerned with an inspection, the subjects will be affected by faults that slip through the inspection. However, effects of low quality in the project will not affect the subjects after the project. They will, for example, not have to continue to work with maintenance after the experimental study because of errors they committed during the study.

An example of this kind of experiment is to carry out an inspection experiment in a student project. There may be a number of groups of students who develop the same program, including all project phases from specification to validation. The motivation for the students to find faults in the inspection experiment is mainly to remove faults that they otherwise have to remove later in the project to a larger cost.

In this type of experiment, the subjects carry out the tasks of the project in order to complete the project, and they carry out additional tasks, such as fault recording, time recording, giving answers to surveys, etc. in order to participate in the experiment. The tasks that are carried out in order to carry out the project must be "for real", and not carried out in order to participate in the experiment. For example, in an inspection the objective must be to find faults in order to obtain a product of higher quality. If faults in this case were injected, this would not be a situation that corresponds to a real project and the situation would in this case be classified as an artificial project.

It is not necessary for all participants to be directly involved in the project for the whole lifetime of the project. They can, for example, also be involved as expert reviewers and then not be directly involved in the rest of the project. They will, however, have an effect during a longer period of time.

3.1.4 Project with long-term commitment

This denotes the situation when a real project is considered and the major objective of the subjects is not to participate in the experimental study. The subjects participate in a project, but typically agree to also participate in a study. Decisions are taken based on the needs of the project and the empirical study is typically not allowed to affect the project to any significant extent if it is not judged as positive for the project.

An example of this kind of experiment is to carry out an inspection experiment in an industrial project where groups working on different sub-systems may use different inspection techniques. The motivation, for the subjects to find faults, is mainly to remove faults that they otherwise have to remove later, or even after the project, to a larger cost. Future effects of low quality in the project can for a long time in the future have a negative effect on the subject.

As in the project with short-term commitment, tasks carried out as part of the project should have realistic objectives, and the example with injected faults would also in this case result in a classification as artificial project. As in the project with short-term commitment, it is not necessary that the participants are directly involved in the project during the whole project.

3.2 Experience

The experience of subjects is commonly reported in published experiments, and there seems to be a consensus that this is one of the factors that affect the results of experiments. We propose the following classes of experience of subjects:

- E1: Undergraduate student[4] with less than 3 months recent[5] industrial experience.
- E2: Graduate student[6] with less than 3 months recent industrial experience
- E3: Academic[7] with less than 3 months recent industrial experience
- E4: Any person with recent industrial experience, between 3 months and 2 years

[4]Students being in their first three years at the university
[5]Less than two years ago
[6]Students having passed a bachelor exam (or equivalent) and aiming for a Master's degree
[7]E.g. faculty members, postdoctoral researchers and doctoral students

Table 2: Summary of classification scheme

Incentive	Experience
I1: Isolated artifact	E1: Undergraduate student with less than 3 months recent industrial experience
I2: Artificial project	E2: Graduate student with less than 3 months recent industrial experience
I3: Project with short-term commitment	E3: Academic with less than 3 months recent industrial experience
I4: Project with long-term commitment	E4: Any person with industrial experience, between 3 months and 2 years
	E5: Any person with industrial experience for more than 2 years

- E5: Any person with industrial experience for more than 2 years

The scale is intended to be ordinal, i.e. a higher value corresponds to more experience than a lower value. The definition of the scale is based on personal experience from conducting experiments. It could be argued that the classification could be done differently, e.g. that E1 should include subjects with less than 2 months experience instead of subjects with less than 3 months experience. However, this scheme was seen as a reasonable starting point which it is possible for many researchers to agree on.

Industrial experience denotes relevant experience, i.e. experience that is usable for the subject in the experiment, e.g. with respect to language used, tools used, application domain, etc. Industrial experience has precedence over education. For example, a graduate student with recent industrial experience for more than 3 months is classified in group 4 or group 5.

Notice that some persons could be hard to classify. For example, a university professor with 15 years old and 3 years long experience could be argued to be classified into E3 instead of E5, since the experience is rather old. In this type of case the classification should be done in the most intuitive way as possible, i.e. there is room for interpretations of the classes. This problem could have been avoided by narrowing the classes and, for example adding an "other"-class. However, this has intentionally been avoided in order to not complicate the scheme too much based on a few cases.

3.3 Summary of scheme

The classification scheme can be summarized with two orthogonal ordinal factors as seen in Table 2.

In the questionnaire that was sent to the authors of the published articles, the classification scheme is presented as a table where each cell represents a unique combination of incentive (column) and experience (row). We therefore define a classification cell (or simply "cell") as a classification of an experiment with respect to both incentive and experience.

3.4 Introduced changes

Based on the usage of the scheme some changes of it were introduced. This has mainly to do with clarification of situations that were not covered by the first version. For example, in the first version it was not clear enough what an "isolated artifact" means. This was identified during internal discussions when the authors of this article classified the published articles, and has been clarified in the presented version. In the final version it is also stressed more clearly that a study cannot be classified at a higher level than "artificial project" if all activities are not carried out "for real".

4. EMPIRICAL EVALUATION

4.1 Classification

In Table 3, our classification is compared with the authors' classification. In most of the cases, the classifications agree, but there are some discrepancies. The discrepancies are marked bold in the table.

In the classification of [4], the difference is in the incentive classification. The reason for this difference is that one sentence was missed when we read the article. Hence, this difference can be regarded as a human mistake, and is not due to the classification scheme.

In the case of [13] and [30], the exact information was difficult to find in their articles. This resulted in that we interpreted the subjects used as more experienced that they were. Since both classification E4 and E5, consider industrial experience, and the articles did not state the number of years with industrial experience, a misclassification has been made. Furthermore, in the case of [30], there is also a misclassification of incentives. The authors of this article interpret the project to be a long-term project, while it was a short-term project according to the authors of the article. Since the incentive classification did not exist before the article was written, the necessary information was not included in the article.

In conclusion, the classifications that were wrong could either depend on missing information in the articles or the authors thought it was difficult to interpret the classification scheme. In the case of [4], we can conclude that we have made an error, but in the other misclassified articles, it is more difficult to conclude the reason for errors. When checking the articles, we still think that it is difficult to classify it in the same cell as the author. Hence, assuming that the author has made a correct classification, it is more probable that the reason is missed information in the articles than wrong interpretation of the scheme, although the latter cannot the excluded.

4.2 Comparison and analysis

In order to make a quantitative analysis, the Kappa value is used [22]. The Kappa value compares two different classifications and results in a value of how well the comparisons agree. In this article, we use the Kappa value to compare the incentive and the experience classifications separately.

Regarding the incentive classification, the table shows two differences. In one of these we missed a sentence in the article; in the other one, we interpreted the project to be long-term instead of short-term. The incentive classification resulted in a Kappa value equal to 0.73. The Kappa value is above 0.6 and thus regarded as good-enough.

For the experience classification, there are two differences. In both these, we interpreted the experiment subjects to have more experiences than they actually had. The experience classification resulted in a Kappa value equal to 0.78.

Table 3: Comparison between the classifications.

Experiment	Our classification	Author classification
[1]	I1, E5	I1, E5
[4]	**I1**, E1	**I3**, E1
[5]	I1, E2	I1, E2
[7]	I1, E1	I1, E1
[13]	I2, **E5**	I2, **E4**
[14]	I2, E5	I2, E5
[15]	I2, E5	I2, E5
[16]	I1, E1	I1, E1
[18]	I1, E2	I1, E2
[19]	I2, E1	I2, E1
[20]	I1, E5	I1, E5
[24]	I1, E2	I1, E2
[30]	**I4**, **E5**	**I3**, **E4**

The Kappa value is above 0.6, which means that the experience classification also is good-enough.

In conclusion, there are some discrepancies in our classification and the authors' classification. In particular, it seems to be difficult to separate E4 and E5, since the research articles have only stated industrial experience and not the number of years in an industrial setting, which is required to separate E4 and E5. However, the Kappa analysis shows that there is a good agreement between our and the authors classifications. Moreover, if authors of articles use the scheme the difference between E4 and E5 would be documented.

4.3 Result analysis

In this section, an investigation of whether the experiments classified in the same classification cells show similar result is presented. In this analysis, we only consider the cells with two or more classified experiments. If the results of the experiments in the same cell point in the same direction, this indicates that the classification helps to provide an explanation of why the results differ among experiments. On the other hand, if there are differences, this shows that there are other factors (than incentives and experience) affecting the results of the experiments. In four cells, there are two or more experiments classified. In addition, there are four experiments classified in one cell each.

[7] and [16] are classified in I1-E1. Both these experiments show no significant results. Hence, they cannot conclude that PBR is better than checklist-based reading (CBR) when undergraduate students use isolated artifacts.

[5], [18], and [24] are classified in I1-E2. [24] compares two versions of PBR, and [5] and [18] compares PBR with AdHoc reading. [5] ran the experiment twice and detected significant differences once. The experiment in [18] showed no significant differences. Hence, they cannot conclude that PBR is significantly better than AdHoc when graduate students use isolated artifacts.

[14] and [15] are classified in I2-E5. These experiments evaluate PBR against AdHoc for code documents. Both these experiments show significant results in favor of PBR. Hence, they conclude that PBR is better than AdHoc in artificial projects by experienced subjects. It should be noticed that both studies are reported from the same researcher and that it could be argued that there is a need for replication by other research groups.

Table 4: Number of experiments in each cell.

Experience	Incentive			
	I1	I2	I3	I4
E1	2	1	1	-
E2	3	-	-	-
E3	-	-	-	-
E4	-	1	1	-
E5	2	2	-	-

[1] and [20] are classified in I1-E5. These experiments show significant results for PBR in comparison with AdHoc. Hence, they conclude that PBR is better than AdHoc when experienced subjects use isolated artifacts. A similar need for replication as for the experiments in I2-E5 may be identified, since the researchers come from the same research environment.

Furthermore, in some cells there is no experiment classified. In the I4 column, no experiments have been conducted and there are only few in the I3 column, see Table 4. This probably depends on that it is more expensive, takes longer time, and is more difficult to perform experiments for these classes. It illustrates that most experiments are run standalone, i.e. they are not part of a larger context such as a development project. For the experience classification, the same trend is not visible. However, in this case, there are no experiments within the E3 row. This shows that experiments have either been run with students or with professionals. The table helps pinpointing the need to run PBR experiments as part of a major project, i.e. experiments within the I4 column.

In summary, when analysing the results cell by cell, the results (in each cell) point in the same direction. This is not evidence that the classification can be used, since there are few experiments in each cell. However, this indicates, together with the Kappa analysis, that the framework and that the factors are important for experiment classification when analysing the results. Further, when we only consider one dimension of the classification, we cannot observe the same agreement among the experiments. Anyhow, the classification scheme will (if used) enforce a more coherent reporting of controlled experiments, which in the long run will improve the possibility to compare different experiments.

5. DISCUSSION

Experimentation is by no means easy, although support exists [28, 10]. A key problem is the external validity of controlled experiments performed in a laboratory setting. This is often materialized in the form of comments regarding the use of students as subjects [26]. However, it is, in our opinion, a too simplistic view to disregard experiments due to the use of students. Instead it is important to understand when students are suitable and how the results may be generalized. One step in this direction is the introduction of an additional dimension when classifying experiments. Other dimensions may very well exist, but a crucial dimension is the motivation of the subjects. Thus, this article introduced a way of trying to capture the motivation by looking at the situation in which the subjects participate in the experiment.

The evaluation of the proposed scheme indicates a number of interesting issues. First, the scheme is easy to use. This was measured by looking at the differences between the two classifications and they came out the same in most cases (10 out of 13). In some cases, there were minor differences that could be attributed to vagueness in the descriptions in the articles or minor misunderstanding by the authors of this article. In one case, there was a major deviation between the classifications. This was resolved by asking the author of the study to re-examine his classification. It turned out that his classification was correct and that the authors of this article made an oversight. Thus, overall the scheme worked well. Moreover, the quality of the classifications ought to improve over time, since authors of experimental articles get used to the scheme and the evaluation step presented in this article has also resulted in that we have tried to improve the descriptions of the different classes to ensure that the distinction between the classes becomes clearer.

Based on the success of the evaluation, the actual results of studies in different cells in the classification scheme were revisited. The hope was that patterns should emerge in terms of identifying similar results for studies appearing in the same cells, while studies in other cells very well may have different results. It turned out that relatively few studies could be attributed to each cell, which makes it difficult to draw any general conclusions. Having said this, it is still believed that the likelihood of finding similar results increases when classifying the studies in more than the experience dimension, i.e. students versus professionals.

However, a key issue for the future is of course to try to understand why the results in a cell are not the same (if that is the case) despite introducing one more classification dimension, i.e. motivation or incentive to produce a good result. Three potential explanations have been identified. First, it may simply not be sufficient to capture two dimensions in the classification. There may be other factors influencing the outcome than captured by the proposed classification scheme.

Second, an interesting issue raised when discussing with one of the authors of the studies investigated was that maybe it is still not fine grained enough to classify the studies; we may need to study the results of the individuals to see any patterns. In other words, it is not sufficient to take the average experience in a study instead it should be considered whether the subjects should be treated as individuals in the scheme. The main motivation for this being that in many studies there is a mixture of individuals, for example, students without any industrial experience and students with long industrial experience. This means that all individuals being classified in a specific classification cell should be compared rather than putting all subjects from one study into one cell.

Third, another possible explanation to why no consistent results emerge may be that the variation when comparing, for example, two competing methods is too large. In other words, single studies may come up with significant results, but different studies may have contradictory results. The best way to combine studies is debatable, including discussions regarding the use of meta-analysis [8, 17]. However, the point here is not to discuss the actual combination. Thus, the box plot in Figure 1 is used to illustrate the point and not to argue that studies should be combined using box plots. Anyhow, a potential reason for obtaining different results

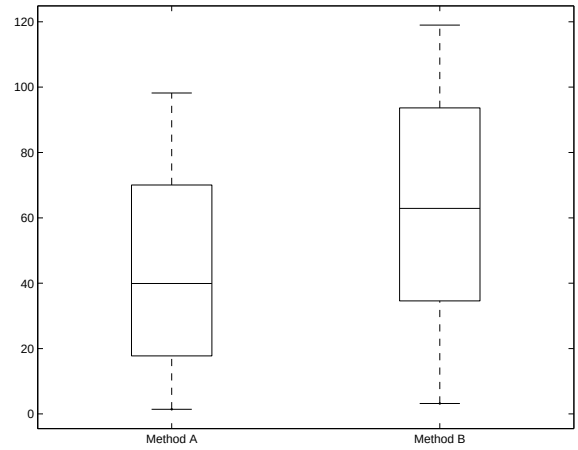

Figure 1: Box Plot for comparing two methods

when collating the results from different studies, and using a box plot, may be that the boxes overlap too much to provide any significant differences between the methods. An example illustrating this point can be found in Figure 1.

In this fictitious case illustrated in the box plot in Figure 1, the overlap of the box plots for Method A and Method B is considerable. Thus, in a single study we may obtain a good outcome (higher value in the example box plot in Figure 1.) for Method A and a poor outcome for Method B. This would most likely result in a significant result in this particular study. However, in a different study the results may indicate that Method A is better, and when combining all available studies may end up with the situation illustrated in Figure 1, i.e. that Method B probably is better, but that the overlap is large.

Three possible reasons for why significant results cannot be found when combining studies and also when using the proposed classification scheme have been highlighted. Anyhow, the classification scheme provides information about one more dimension than just considering experience (for example students vs. professionals). Hopefully, a better classification of controlled experiments would help us better understand how to interpret and combine results from different studies. Experimentation is an important tool in software engineering research [25], but the full benefit cannot be enjoyed unless the research community learn how to effectively combine studies.

An additional consideration, which also points to some interesting future work, is whether the focus is too much on being able to compare absolute results between different types of subjects, for example, when comparing two competing methods. It may be too ambitious to expect that it should be possible to transfer results from using students to professional engineers when it comes to expected improvements when evaluating two competing methods. However, the relative order of the two competing methods may be the same. An example of this situation may be that two competing design methods (A and B) are compared and an experiment with students in an artificial project shows that method A results in 20% fewer faults. The outcome that 20%

fewer faults are made using method A may not be possible to generalize, but maybe it is possible to generalize to that method A is likely to produce fewer faults than method B. This would be an example when the absolute value of 20% is not possible to transfer to another situation, but the knowledge that one method is likely to be better may very well be possible to transfer to another context.

6. CONCLUSIONS AND FURTHER WORK

The classification scheme presented in this article has been driven by an identified need to better understand results from controlled experiments, and in particular how results from different studies can be combined. The scheme addressed the important aspect of motivation of subjects participating in empirical studies in general and experiments in particular. The importance of this aspect is supported by a recent study on requirements prioritization [3].

It can be concluded that the scheme was easy to use and understand. This is concluded from that the authors of the specific studies found the scheme straightforward to use and the high correspondence between the classifications made by the authors of the studies and the authors of this article.

However, the classification did not help discern more clear patterns in terms of results from individual studies, which partially is due to the few studies ending up in the same classification cells. Some possible explanations to this outcome were identified, but the general conclusion is still that it is important to capture more context variables in controlled experiments. The proposed classification scheme contributes with structuring and systematizing the motivation of subjects participating in studies. Although, it does not provide the whole truth of what governs the outcome in an empirical study, it is believed to be one step in the right direction to better document controlled experiments to make them easier to compare with each other. Good documentation is a prerequisite to obtain the full potential of controlled experiments where replication is crucial to build a body of knowledge. Moreover, good documentation to make studies understandable and comparable is needed if moving towards evidence-based software engineering [11].

The further work should include investigating if more aspects have to be included, for example, the application domain. Moreover, it has to be investigated whether using the scheme on an individual level instead of on a study level would help us discern any patterns. Future work should also include work regarding the possibility to generalize relative order rather than absolute values.

However, this is a first attempt to include incentives and motivation as a complement to experience with the objective to improve the understanding of experiments in software engineering. Other aspects that have to be taken into account may include the actual development phase. For example, does it matter if investigating software inspections and reading techniques in the requirements phase or design phase? To facilitate studies and in particular replications other aspects have also to be taken into consideration. This may include aspects on, for example, tacit knowledge as discussed in [21].

7. ACKNOWLEDGMENT

We would like to thank all authors of the articles used in the evaluation for their cooperation and responsiveness. A special thanks to Dr. Forrest Shull for raising the issue of studying individuals rather than studies.

8. REFERENCES

[1] V. R. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull, S. Sørumgård, and M. V. Zelkowits. The empirical investigation of perspective-based reading. *Empirical Software Engineering: An International Journal*, 1(2):133–164, 1996.

[2] V. R. Basili, F. Shull, and F. Lanubile. Building knowledge through families of experiments. *IEEE Transactions on Software Engineering*, 25(4):456–473, 1999.

[3] P. Berander. Using students as subjects in requirements prioritization. In *Proc. 3rd International Symposium on Empirical Software Engineering*, pages 167–176, 2004.

[4] S. Biffl. *Software Inspection Techniques to Support Project and Quality Management*. Shaker Verlag, Austria, 2001. Habilitationsschrift.

[5] M. Ciolkowski, C. Differding, O. Laitenberger, and J. Münch. Empirical investigation of perspective-based reading: A replicated experiment. Technical Report 97-13, ISERN, 1997.

[6] K. El Emam. Benchmarking kappa: Interrater agreement in software process assessment. *Empirical Software Engineering: An International Journal*, 4(2):113–133, 1999.

[7] M. Halling, S. Biffl, T. Grechenig, and M. Köhle. Using reading techniques to focus inspection performance. In *Proc. 27th Euromicro Workshop on Software Process and Product Improvement*, pages 248–257, 2001.

[8] W. Hayes. Research synthesis in software engineering: A case for meta-analysis. In *Proc. 6th International Software Metrics Symposium*, pages 143–151, 1999.

[9] 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: An International Journal*, 5(3):201–214, 2000.

[10] N. Juristo and A. M. Moreno. *Basics of Software Engineering Experimentation*. Kluwer Academic Publishers, 2001.

[11] B. A. Kitchenham, T. Dybå, and M. Jørgensen. Evidence-based software engineering. In *Proc. International Conference on Software Engineering*, pages 273–281, 2004.

[12] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering*, 28(8):721–734, 2002.

[13] O. Laitenberger, C. Atkinson, M. Schlich, and K. El Emam. An experimental comparison of reading techniques for defect detection in UML design documents. *Journal of Systems and Software*, 53(2):183–204, 2000.

[14] O. Laitenberger and J. M. DeBaud. Perspective-based reading of code documents at Robert Bosch GmbH. *Information and Software Technology*, 39(11), 1997.

[15] O. Laitenberger, K. El Emam, and T. Harbich. An internally replicated quasi-experimental comparison of checklist and perspective-based reading of code. *IEEE Transactions on Software Engineering*, 27(5):387–421, 2001.

[16] F. Lanubile and G. Visaggio. Evaluating defect detection techniques for software requirements inspections. Technical Report 00-08, ISERN, 2000.

[17] J. Miller. Can results from software engineering experiments be safely combined? In *Proc. 6th International Software Metrics Symposium*, pages 152–158, 1999.

[18] B. Regnell, P. Runeson, and T. Thelin. Are the perspectives really different? - further experimentation on scenario-based reading of requirements. *Empirical Software Engineering: An International Journal*, 5(4):331–356, 2000.

[19] G. Sabaliauskaite, S. Matsukawa, S. Kusumoto, and K. Inoue. An experimental comparison of checklist-based reading and perspective-based reading for UML design document inspection. In *Proc. 1st International Symposium on Empirical Software Engineering*, pages 148–157, 2002.

[20] F. Shull. *Developing Techniques for Using Software Documents: A Series of Empirical Studies*. PhD thesis, Computer Science Department, University of Maryland, USA, 1998.

[21] F. Shull, V. R. Basili, J. Carver, J. C. Maldonado, G. H. Travassos, M. Mendonca, and S. Fabbri. Replicating software engineering experiments: Addressing the tacit knowledge problem. In *Proc. 1st International Symposium on Empirical Software Engineering*, pages 7–16, 2002.

[22] S. Siegel and N. J. Castellan. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Singapore, 1988.

[23] D. I. K. Sjøberg, B. Anda, E. Arisholm, T. Dybå, M. Jørgensen, A. Karahasanovic, E. F. Koren, and M. Vokác. Conducting realistic experiments in software engineering. In *Proc. 1st International Symposium on Empirical Software Engineering*, pages 17–26, 2002.

[24] S. Sørumgård. *Verification of Process Conformance in Empirical Studies of Software Development*. PhD thesis, Department of Computer and Information Science, The Norwegian University of Science and Technology, Norway, 1997.

[25] W. F. Tichy. Should computer scientists experiment more? *IEEE Computer*, 31(5):32–41, 1998.

[26] W. F. Tichy. Hints for reviewing empirical work in software engineering. *Empirical Software Engineering: An International Journal*, 5(4):309–312, 2000.

[27] W. F. Tichy, P. Lukowicz, L. Prechelt, and E. A. Heinz. Experimental evaluation in computer science: A quantitative study. *Journal of Systems and Software*, 28(1):9–18, 1995.

[28] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in Software Engineering – An Introduction*. Kluwer Academic Publishers, 1999.

[29] M. V. Zelkowitz and D. Wallace. Experimental validation in software engineering. *Information and Software Technology*, 39(11):735–743, 1997.

[30] Z. Zhang, V. R. Basili, and B. Schneiderman. Perspective-based usability inspection: An empirical validation of efficacy. *Empirical Software Engineering: An International Journal*, 4(1):43–69, 1999.