UNIVERSITY OF OSLO
Department of Informatics

Survey of Controlled Software Engineering Experiments with Focus on Subjects

Cand Scient thesis

Ove Hansen

5th August 2004

Abstract

Empirical research within the Software Engineering field is a fairly new discipline. This calls for the development of new methods and new standards for how to perform and report experimental procedures and results. For research to be of interest for the community we need to know how the research has been performed, and within which context. Through the CONTEXT project, employees and M.Sc. students at Simula Research Laboratories have performed a survey on all controlled experiments reported in 12 leading journals and conferences during the period 1993 to 2002; a total of 118 experiments from 107 papers. Our aim has been to find out how the experiments are reported – what information is supplied, and to some extent the quality of this information.

The focus of this thesis is on the subjects participating in the experiments. What information is provided regarding who they are and their backgrounds, and to what extent variables that might be of relevance for the validity and generalizability are described. This gives an indication of the maturity of this research.

Almost all controlled experiments in our field are aimed at the professional population, yet only 26 % of the experiments in our survey used professionals as subjects. Only 19 % of the experiments using students as subjects generalize their findings to professionals. The heterogeneity of the subjects is generally not given much attention, as differences between them are rarely discussed and background variables often described in little detail. We have found that whether or not it is appropriate to use students as subjects is highly dependent on the issue studied and which populations the research is targeting. The terms “student” and “professional” are in many cases misleading, as these may not be relevant for the task at hand. It may often make more sense to talk about novices and experts.

The overall impression of controlled experiments in the SE field is that this discipline is developing, but not yet mature. Our analysis shows that there is a need for standards and guidelines that authors should adhere to when publishing their work. Today many papers provides so little information that it is difficult for the reader to get an impression of to what extent the results are valid, and if they are – to whom.

Acknowledgements

This thesis was written in association with the CONTEXT project at Simula Research Laboratories for my Cand Scient degree at the Department of Informatics, the University of Oslo.

First of all I want to thank Anette C. Rekdal, who by introducing me to the Software Engineering group at Simula Research Laboratories and the CONTEXT project let me escape from years of frustration trying to get my Cand Scient thesis on the way. I also want to thank my advisor, Dag Sjøberg, and the members of the CONTEXT project group: Vigdis By, Jo Hannay, Amela Karahasanovic, and my fellow M.Sc. student Nils-Kristian Liborg. Cooperating with these people has been inspiring. Marit-Solveig Finset has been most helpful providing us with statistical data generated in SAS. I would also like to thank Barbara Kitchenham for sharing valuable insight during her Oslo visit.

Finally I want to thank my family and all others who supported me and had faith in me as years ran by and I was working my way through dead ends and endless frustrations. A big smile goes to my former colleague, Thomas Strandenæs, for supplying me with the picture of "the bitterest person in the world" – the grad school dropout (described on the next page). This was a true "inspiration" during the darkest moments.

Some background sections in this thesis are in part collected from a paper in progress, reporting results from the project. These sections are mainly in chapter 1.1, 2.1 and chapter 3. Examples and quotes from the source material are referenced anonymously, as the material remains confidential as long as it is still subject of investigation.

Oslo, August 2004

Ove Hansen


MEET THE BITTEREST PERSON IN THE WORLD
THE GRAD SCHOOL DROPOUT
SPENT 4 LONG YEARS LIVING WRETCHEDLY
WORKED LATE INTO THE NIGHT
REPRESSED ALL ANGER
HAD A SERIES OF LOW-PAYING DEGRADING JOBS
PLOWED THROUGH THOUSANDS OF TEDIOUS BOOKS
FINALLY GAVE UP AFTER BEING UNABLE TO FINISH THESIS
CURRENTLY UNABLE TO ENJOY ANYTHING

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There is an increasing understanding in the software engineering (SE) community that empirical studies are needed to develop or improve processes, methods and tools for software development and maintenance [1-8]. An important category of empirical study is controlled experiments, whose conduct is the classical scientific method for identifying cause-effect relationships.

1.1 Research in the Software Engineering field

Research in the software engineering field is difficult as more or less unique solutions are developed every time. It is problematic to generalize to large populations, thus it is important to keep in mind what populations you wish your results to apply to. Each piece of research is interesting in itself, but as pieces in a puzzle it is more interesting when seen in relation to other pieces - the big picture is in some respects greater than the sum of the pieces. Thus, it is of interest to get an overview of what research is conducted within a certain field, and what quality this research holds.

At Simula Research Laboratories we have conducted a survey that characterises the controlled software engineering experiments published in a sample of nine journals and three conference proceedings in the decade from 1993 to 2002 [9]. The survey is an attempt to systematize all controlled experiments reported in leading SE journals and conferences by analyzing the experiments in detail, and give an overview of how controlled experiments actually are reported. Researchers and master students have analysed these papers with respect to different variables within different focuses. These focuses are technology, tasks, environments, topics and subjects.

Some guidelines and recommendations for improving the quality of SE experimentation have been given [10]. We believe that the state-of-the-art description of formal experimentation in software engineering provided by this survey further helps identify appropriate guidelines to make software engineering a more mature scientific discipline.

1.2 Focus of this Thesis

The essential purpose for controlled experiments is to study cause and effect relations. You alter one variable, usually referred to as the independent variable or the treatment, and observe how the variation of this variable affects some other variable, referred to as the dependent variable. In some sciences this can be done mathematically, without considering potential confounding effects, but as in all other sciences studying human behaviour, SE experiments need to consider very carefully what other causes (except from the variation of the independent variables) can influence the variations in the dependent variable. Due to the diversity of different people and backgrounds, qualitative methods (as opposed to quantitative methods) are often necessary, just like they are in humanistic disciplines.

Much research in this field is aimed at refining work processes and methods in terms of gaining more efficient work practices. People, process and technology are three different aspects that affect software engineering [11]. Process and technology are aspects it is possible to control; the people involved, however, are quite heterogeneous. This makes research more complicated.

This thesis focuses on the role of the subjects participating in the experiments and how this is reported in the papers. Who are they? What backgrounds do they have? To what extent do the authors address the heterogeneity of the subjects due to their very different backgrounds, and what implications does this have for the experiments and their generalization?

One of the largest challenges is perhaps, to recruit a sufficient number of representative subjects, that is, subjects that are drawn from the actual population about which we wish to make claims [10]. It may also be difficult, and involve high costs, to maintain the cooperation of the subjects for a sufficient amount of time to enable them to perform tasks of a realistic size.

Most experiments are performed using student subjects, and a recurring question is whether these are justifiable with respect to the population we want our findings to apply to. Students are commonly used as subject because they are easier accessible than professionals. They are cheap, more flexible regarding time-issues, and some times experiments can be run as a part of courses they are taking. Can results produced by students be generalized to apply for professionals?

Our focus has not been to look at the qualitative aspects of the way the experiments are performed, but to analyze the way these experiments have been reported in the published papers. Experimenting in the SE field is a rather new discipline, and one of our main questions is whether this research is mature or not.

1.3 Organization of This Thesis

This thesis is organized as follows. Chapter 2 relates our survey to other relevant surveys and looks into earlier work regarding students versus professionals as subjects. Chapter 3 outlines the research methods for this survey. Chapter 4 presents the results, while chapter 5 discusses some of these. Chapter 6 discusses threats to validity of this survey, and chapter 7 draws conclusions and suggests future work.

Chapter 2

Related Work

Surveys that are comparable to ours have been conducted earlier. 2.1gives an overview of these. Little research has been published addressing the issue of students as subjects, but one such study is described in 2.2

2.1 Surveys

Table 2.1 describes the purpose, scope and extent of sampled papers in four major surveys as well as our survey. Tichy *et al* [5] compare the amount of experimental work published in a few computer science journals and conference proceedings with the amount of experimental work published in one journal on artificial neural network and one journal on optical engineering. In total, 403 articles are surveyed and classified into the five categories: formal theory, design and modelling, empirical work, hypothesis testing and "other". Zelkowitz and Wallace [8] propose a taxonomy of empirical studies in software engineering and report a survey in which 612 papers are classified within this taxonomy. Glass *et al* [12] investigate 369 articles with respect to topics, research approaches, research methods, reference disciplines and level of analysis.

The above surveys give a comprehensive picture of research methods used in software engineering. They differ in purpose, criteria for selection of papers and taxonomies of empirical studies. Their results, nevertheless, suggest the same: The major part of published papers in computer science and software engineering provide little or no experimental validation; the proportion of controlled experiments being particularly low. The surveys also propose means to increase the amount of empirical studies and their quality.

The major difference between those surveys and ours is that they describe the extent and some characteristics of all empirical studies, while we provide an in-depth study of controlled experiments only. We have narrowed our field of interest, but widened our sample. The survey by Zendler [13] also focuses on experiments. He reports the results of 31 experiments with the aim of developing a preliminary software engineering theory. Shaw [14] has categorised the research reported in papers submitted and accepted for ICSE 2002.

Table 2.1: Surveys that reports empirical studies

	(Tichy *et al.* 1995)	**(Zelkowitz *et al.* 1997)**	**(Glass *et al.* 2002)**	**(Zendler 2001)**	**Our survey**
Purpose	Comparing the extent of empirical studies in computer science with other fields	Classifying empirical studies in SE and to validate the taxonomy of empirical studies proposed by the authors	Surveying topics, research approaches, research methods, reference disciplines and level of analysis.	Developing a preliminary SE theory from the results of various SE experiments	Surveying topics, subjects, tasks, environments, and generalisation of controlled experiments in SE
Scope	Comp. Sci., incl. SE	SE	SE	SE	SE
Journals	ACM (random publications), TSE, PLDI Proc., TOCS, TOPLAS	ICSE Proc., IEEE Software, TSE	IEEE Software, IST, JSS, SP&E, TOSEM, TSE	Various journals and conference proceedings	EMSE, ICSE, IEEE Computer, IEEE Software, ISESE, IST, JSME, JSS, METRICS, SP&E, TOSEM, TSE
Sampling of papers	Partly random 1991-1994; one to four volumes per journal, random selection of work published by ACM in 1993	All papers in 1985, 1990 and 1995	Random in the period 1995-1999	Not reported	All papers in the period 1993-2002
Number of investigated papers	403	612	369	49 papers assessed, 31 papers analysed in depth	5453 papers scanned, 107 papers analysed in depth

In addition to the general surveys described above, there are of course many surveys within sub-disciplines of software engineering, for example, object oriented technology [15], testing techniques [16], and software estimation [17].

2.2 Students as Subjects

Runeson is one of few who have looked into the question of using students as experimental subjects [11]. His hypothesis was that "there are small differences between graduate students and industry people on one hand, while there are significant differences between graduate students and freshmen students on the

other hand." He conducted an experiment in the context of the Personal Software Process with freshmen and graduate students, and later on also industry people. Industry data was not available to perform comparisons with the student groups, but he concluded that there were substantial differences between freshmen and graduate students, and even stated that "freshmen students should not be used as subjects for software engineering experiments" [11]. This conclusion must be viewed with respect to how valid the results are for a professional population in the industry.

Chapter 3

Research Methods

This chapter outlines how this survey was conducted. Section 3.1 presents our definition of a controlled experiment. 3.2 and 3.3 describe the selection of papers included in our survey, while 3.4 and 3.5 give a brief description of the analysis and data storage.

3.1 Controlled Experiments in Software Engineering

The common attribute in all experiments is control of treatment, though control can take many different forms. Shadish, Cook, and Campbell [18] provide the following definitions:

Experiment: A study in which an intervention is deliberately introduced to observe its effects.

Randomized experiment: An experiment in which units are assigned to receive the treatment or an alternative condition by a random process such as the toss of a coin or a table of random numbers.

Quasi-Experiment: An experiment in which units are not assigned to conditions randomly.

Correlation study: Usually synonymous with non-experimental or observational study; a study that simply observes the size and direction of a relationship among variables.

This survey focuses upon experiments in which individuals or teams (the experimental units) apply a process, method, technique, language or tool (the treatments) to conduct one or more software engineering tasks. (An organisation or company could also be an experimental unit, but we found no such cases in our survey.) The insistence of treatment excludes empirical studies such as pure correlation studies, re-sampling studies and other studies that are solely based on calculations on existing data. Moreover, usability experiments are not included since we regard those as part of another discipline (human computer interaction). Articles that focus on methodological issues but that still describe experiments and

articles that only summarise experiments are also not included; our survey focuses on articles that provide the main reporting of an experiment.

In addition to randomized experiments, we include quasi-experiments. General random assignment of experimental units to treatments may not always be feasible, e.g., for logistic reasons. Laitenberger *et al*. [19] report an experiment in which units are imported into the experiment from intact training groups in a company. Randomised assignment would in this case have disturbed the training process. See also [20]. Since the term *experiment* is inconsistently used in the software engineering community (often used synonymously with empirical study), we use the term *controlled experiment* to emphasize the control of application of treatment.

3.2 Selection of Journals and Conferences

It is not possible to cover all published articles in a survey like this, thus a selection had to be made. Our sample is based on what is acknowledged to be leading journals and conferences by the SE community (also based on earlier work like Glass etc). Whereas Zelkowitz/Glass were sampling (not covering all papers), we have examined all papers published in our selection of journals the last decade.

The following 12 journals and conferences provide the basis for our material: IEEE Transactions on Software Engineering (TSE), Empirical Software Engineering (EMSE), Journal of Information and Software Technology (IST), Journal of Systems and Software (JSS), Journal of Software Maintenance and Evolution (SME), Software: Practice and Experience (SP&E), ACM Transaction on Software Engineering Methodology (TOSEM), IEEE Computer, IEEE Software, The proceedings of the IEEE International Symposium on Software Metrics (METRICS), The IEEE International Symposium on Empirical Software Engineering (ISESE) and The International Conference on Software Engineering (ICSE).

3.3 Identification of Articles Reporting Controlled Experiments

To identify the controlled experiments in the different journals and conferences selected, one person systematically read the title and abstract of the 5453 scientific articles published in the selected journals and conference proceedings for the decade 1993-2002. If it was unclear from the title or abstract whether a controlled experiment was described, the whole article was read by both the same person and another person in the project team. In the end, 107 articles were selected. Note that identifying the right articles was not straightforward since the terminology in this

area is confusing. For example, several authors claimed they described experiments even though no treatment was applied in the study.

Among the 5453 articles, 107 (2.0 %) reported controlled experiments in which individuals or teams conducted one or more software engineering tasks according to our definition.

3.4 Analysing Articles

The survey data was stored in a relational database (MS SQL Server 2000). Some information was specific to an article, some was specific to an experiment and some information concerned the combination of article and experiment. Moreover, one article could describe several experiments and one experiment could be described in several articles, then typically with a different focus. Consequently, a data model with the entities Article, Experiment and Focus (combination Article-Experiment) were defined with a corresponding set of attributes relevant to our survey. Figure 3.1 shows the relation between these entities.

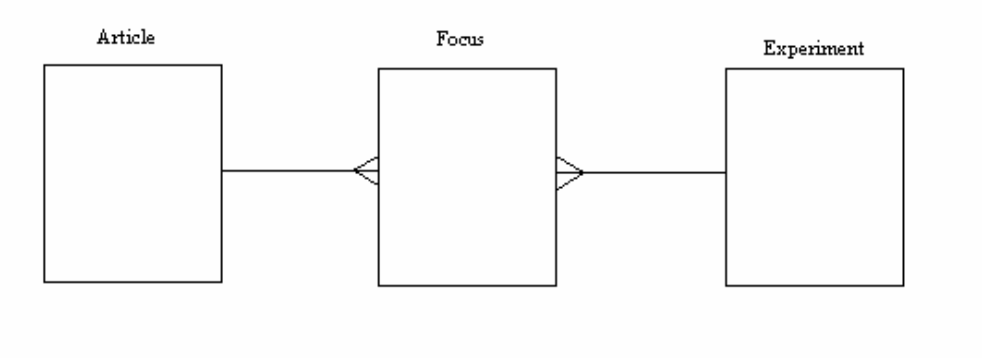

Figure 3.1: Relation between article, experiment and combinations of these

Six researchers or research assistants analysed the articles, focusing on certain aspects. Each aspects, corresponding to a set of attributes of the database, was analysed independently by at least two people to ensure correctness. The data from the different analyses were merged into resulting tables to make one single dataset. Note that some fields at the time present are analyzed only once, because they were introduced during the second analysis phase, conducted by other people than in the first phase. The main analysis tool was SAS, and some of the tables produced with this tool can be viewed in appendix B.

At some point during the original analysis it was decided to send requests to authors regarding information that was unclear or not stated, but since the response varied substantially and the amount of unclear points kept growing it was decided

to focus merely on the papers themselves and the way things were reported in them. Hence, information collected from other sources than the papers had to be disregarded. This also resulted in us having to oversee facts we knew about papers written at Simula Research Laboratories, due to that this information was not stated in the papers.

3.5 Database

Table 3.1 gives an overview over which variables have been analyzed and described regarding subjects. Appendix A contains more details about these. Most variables were identified and defined during the research assistants' identification and analysis of papers, but the interpretation may have changed during our second analysis and the merging of data. A few fields are added during the second analysis.

Table 3.1: Variables analysed and reported in database

Fieldname	**Description**
Total_number_of_participants	How many subjects participated originally in the experiment
Active_participants	How many subjects actually participated and were included in the analyses
Students	How many students participated
Undergraduate_students	How many undergraduate students participated
Graduate_students	How many graduate students participated
PhD	How many PhD students participated
Participating_scientists	How many scientists participated as subjects
Professionals	How many professionals participated
Replication*	Is the experiment a replication?
Individual_or_team	Did subjects work individually, as team or both
Selection_of_participants*	Who the participants are
Information_about_participants*	Information about subjects background
Recruitment*	How the participants were recruited
Paid_rewarded*	Whether the subjects where paid or rewarded in some way.
Mandatory*	Whether experiment participation was mandatory or voluntarily
Differences_of_group_members	Are there different results within the same subject category?
Categories_of_subjects	Which combination of subject categories are represented in the experiment
Study_unit**	Which experimental unit data is collected and analyzed for. Individual, team or both
Number_of_teams**	The number of teams that participated in the experiment
Different_results_between_categories	Are there different results between the different subject categories that participated?
Generalizations_from_students*	Whether the students in the samples are generalized to professionals

* This field was also analyzed by a different member of the project team
** This field was analyzed by only one person

Chapter 4

Subjects in SE Experiments

The data collection and analyses generated information that can be viewed in numerous combinations. In this chapter some of them are presented.

4.1 Subject Categories

Two main categories of subjects participating in software engineering experiments are *students* and *professionals*. In 92 (78 %) of the 118 experiments investigated, students participated, either alone or in company with professionals and/or scientists. Professionals took part in 31 experiments (26 %). Table 4.1 shows descriptive statistics for the various categories of subjects that participated in the experiments. As participating subjects we count those who took an active part in the experiment and were included in the analyses performed.

Table 4.1: Categories of participants in experiments

Category of participants		Experiments		Participants					
		N	%	Mean	Std	Min	Median	Max	Sum
Students only	Undergraduates only	42	35.6	63.0	61.3	10	45	266	2644
	Graduates only	15	12.7	25.1	11.1	9	24	48	377
	Undergraduates and graduates	17	14.4	57.4	57.5	6	31	208	976
	Students, type unknown	8	6.8	65.5	70.3	13	43	231	524
		82	69.5	55.1	56.7	6	36	266	4521
Professionals only		24	20.3	18.3	13.9	4	17	68	439
Mixed type of Participants	Undergraduates and professionals	2	1.7	75.0	35.4	50	75	100	150
	Graduates and professionals	2	1.7	45.0	4.2	42	45	48	90
	Graduates and scientists	1	0.8	34.0	-	34	34	34	34
	Students, type unknown and scientists	1	0.8	12.0	-	12	12	12	12
	Undergraduates, graduates and scientists	1	0.8	18.0	-	18	18	18	18
	Undergraduates, graduates, scientists and professionals	1	0.8	20.0	-	20	20	20	20
	Students, type unknown, professionals and scientists	1	0.8	120.0	-	120	120	120	120
	Undergraduates, graduates and professionals	1	0.8	36.0	-	36	36	36	36
		10	8.5	48.0	35.4	12	39	120	480
Unknown		2	1.7	21.5	17.7	9	22	34	43
Total		118	100.00	46.5	50.9	4	30	266	5483

27 experiments reported that subjects from more than one category were used. In 17 of these the distribution between the different categories was not reported, but which categories the subjects came from was (with exception from two experiments where the type of students used was not reported, but since other categories were included as well we still know there were more than one category of subjects). In addition, eight experiments used only students without specifying whether they were undergraduates, graduates or a combination of both. In two experiments absolutely nothing was stated about which categories of subjects were used; merely the number of subjects was reported.

51 % of the experiments where students were the only participating category used only undergraduate students as subjects. These were also the experiments using largest number of subjects. The eight experiments where we do not know whether the students were undergraduates or graduates show similar statistics, but as the number of experiments in this category is low and we know that undergraduates were represented in at least three of these, it seems fair to conclude that graduates appear in considerable less numbers.

Table 4.2: Distribution of experiments and subject categories over time

Category	**Year**									
	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Undergraduates only	1	0	0	2	3	6	7	8	6	9
Graduates only	1	2	1	2	2	2	0	3	2	0
Undergraduates and graduates	1	0	5	3	0	1	0	4	1	2
Students, type unknown	0	0	0	0	1	2	1	1	2	1
Professionals only	0	1	0	4	5	4	2	2	6	0
Undergraduates and professionals	0	0	0	0	0	0	0	1	0	1
Graduates and professionals	0	0	0	0	1	0	0	1	0	0
Graduates and scientists	0	0	0	1	0	0	0	0	0	0
Students, type unknown and scientists	0	0	0	0	0	0	0	1	0	0
Undergraduates, graduates and scientists	0	0	0	0	0	0	1	0	0	0
Undergraduates, graduates, scientists and professionals	0	0	0	0	0	0	1	0	0	0
Students, type unknown, professionals and scientists	0	0	0	0	0	1	0	0	0	0
Undergraduates, graduates and professionals	0	0	0	0	0	0	0	0	1	0
Unknown	0	0	0	0	1	0	1	0	0	0
Total	**3**	**3**	**6**	**12**	**13**	**16**	**13**	**21**	**18**	**13**

Table 4.2 shows the distribution of experiments and mixtures of participants over time. Note that in the years 1993 through 1995 professionals took part in only one experiment altogether.

4.1.1 Student categories

The reported subject categories “undergraduate students”, “bachelor students”, “third and fourth year students”, “last year students”, “honors students”, “juniors and seniors”, “majors” and “students following undergraduate courses” are categorized as *undergraduate students* in this survey. The terms “juniors” and “seniors” caused us a bit of trouble as it was not clear whether these terms related to undergraduate students only or also for graduate students, but after examining those articles that described this in detail and checking with people familiar with the American university system we learned that these terms relate only to undergraduates. “Graduate students”, “students following graduate courses” or “Master programs”, “MSc” and “PhD” students, are all included in the category *graduate students*. ‘Students in computer science’ or merely “students” are categorized as *students, type unknown* as these terms are undeterminable. The category ‘Students, type unknown’ contains experiments were we do not know anything else than that the subjects were students, that some of them were from one category but not necessarily all, or that the information supplied does not make it clear what category of students to count them as. When a country specific term was used (for instance, the German “Vordiplom”, which is aquired after two years), we categorized the students based on the number of years of study. In nine experiments (eight with only students and one mixed) information about type of students was not possible to extract. In two cases, subject information was not provided at all.

4.1.2 Professional and scientist categories

The category *professionals* includes “developers”, “practitioners”, “software engineers”, “analysts”, “domain experts”, “business managers”, “facilitators” and “professionals”. “Professors”, “post-doctorates” and “staff members” of educational institutions were categorized as *scientists*.

Figure 4.1 shows the relative proportion of experiments that include professional subjects compared to the total number of experiments performed each year. This also includes seven experiments that use both students and professionals as subjects. Note that the number of experiments conducted each year is in magnitude 3 to 21, and particularly in the early years the number of experiments was in the lower scale (see table 4.2 for details). Thus, small changes in numbers make considerable impact on the graph. For more details on number of experiments conducted each year, see figure 5.1.

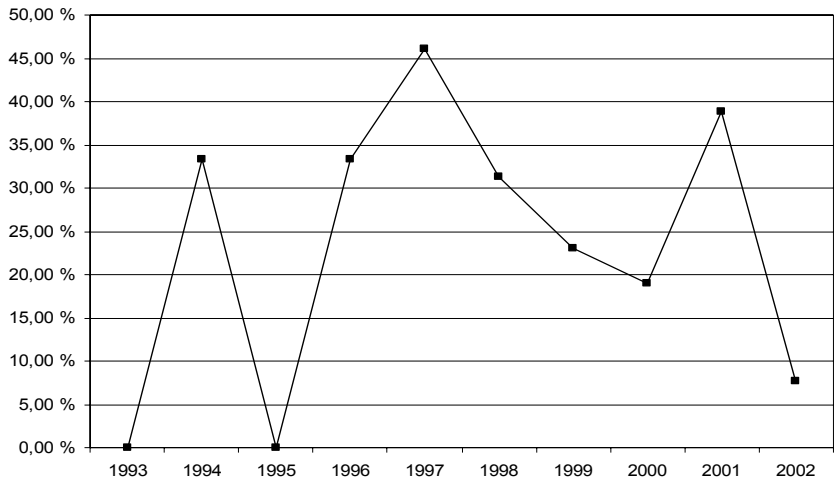

Figure 4.1: Use of professional subjects over time

Seven papers describe experiments using both students and professionals. In addition, four series of experiments have replications where in each series at least one of the replications involves professionals where students were involved earlier, or vice versa. The use of students versus professionals in software engineering experiments will be discussed in chapter 5.

4.2 Number of Subjects

The total number of participants was reported in all the papers, either explicit or implicit so that it could be roughly calculated (for instance by stating that 10 teams with average four subjects participated). In one case the paper stated that "more than hundred subjects participated". In this case the total number of participants was for simplicity recorded as 100 in our database. In 26 experiments mortality among the subjects were reported. Altogether approximately 5611* subjects participated, and 5483 of these were counted as active participants. This means that the total mortality among the subjects was 2.3 %. It is interesting to note that even in experiments with as many as 266 subjects (as well as many other experiments with a relatively high number of subjects), no mortality was reported. This should mean that no outliers were removed, and that no subjects failed to complete the experiment. Unless the experimental design was made to not depend on how many subjects were meant to participate, the number of experiments reporting mortality seems low. One article states that "Non-random drop-out of subjects has been

* Due to the above stated approximations

avoided by the experimental design, i.e. assignment of groups only on the second day of the experiment, i.e. directly before the treatment, and not before the pre-test already on the first day of the experiment", but most papers say nothing about how they avoided mortality.

Table 4.3: Known subjects from different categories

Category	Unknown	Number unclear	Zero	Known	Total
Undergraduates	9	17	42	50	2843
Graduates	12	14	68	24	599
Professionals	2	2	85	29	536
Scientists	2	3	111	2	74
Total					**4052**

Table 4.3 shows how many participants from the different categories we were able to extract from the material we have studied. The column "unknown" shows the number of experiments where we do not know whether the category was represented or not, while the "number unclear" column shows how many experiments we know subjects from the category participated without knowing the exact number. We can account for 4681 participating students, meaning that we have additional information regarding 1239 subjects where we do not know whether they were undergraduates or graduates as they were merely presented as students in general terms. The remaining 802 subjects we are not able to categorize at all.

4521 subjects participated in the 82 studies where only students participated as subjects (an average of 55.1 subjects in each experiment), whereas 439 professionals participated in the 24 experiments (average 18.3 per experiment) where professionals were the only subject category present. These numbers are skewed by some student experiment with hundreds of subjects, but with more than 80 experiments to calculate the mean from the difference is still of great magnitude. Undergraduates are also much more used than graduate students, with 2644 undergraduates in 42 experiments (average 63.0) versus 377 graduates in 15 experiments (average 25.1) where these categories respectively were the only one present.

4.3 Information about Subjects

The papers analyzed vary a great deal on how they report background information about the participants. The fields “Selection of participants” contains information about who the subjects were and where they came from, while “Information about participants” contained general or detailed background information as working experience, age, programming experience etc. To be able to say anything about the way this information was reported we had to look for variables that reoccurred in different papers and categorize according to these. Thus we ended up with 17 different labels, listed in table 4.4. One or more of these apply to each experiment as each paper can contain information belonging to more than one category of information. The level of details reported varies to a large extent. An example of detailed information on programming experience is: “On average subjects programming experience was 7.5 years, using 4.6 different programming languages with largest program of 3510 LOC. Before the course 69 % of the subjects had experience with object-oriented programming, 58 % with programming GUIs”. An example of a general description is: “Some of the students had industrial programming experience”.

15 experiments (13 %) give no information about the subjects beyond categorizing them as students or professionals. Gender and age were addressed for eight and seven experiments respectively. For 83 experiments (70 %), task related information about participants was provided. Explicit or detailed information was reported in 41 cases and more general information in 42 cases. Working experience and programming experience were reported for 25 (21 %) and 38 (32 %) experiments respectively. Training in association with the experiment was addressed for 39 experiments (33 %). This could be explicit training preceding the experiment, training as part of a course or an experiment performed as part of a training session.

Table 4.4: Information about subjects

Description	N	%
Explicit info about gender	6	5.1
General info about gender (no numbers)	2	1.7
Explicit age	6	5.1
General info about age (no numbers)	1	0.9
Explicit programming experience (numbers or particular info)	20	17.0
General programming experience	18	15.3
Explicit working experience (numbers or particular info)	14	11.9
General working experience	11	9.3
General grade info	5	4.2
GPA, grades etc.	1	0.9
Explicit task experience (detailed or extensive info)	41	34.8
General task experience	42	35.6
General demographic info	3	2.5
Explicit information about training	8	6.8
General info about training	29	24.6
Training as part of course	2	1.7
Experiments with no info	15	12.7

Table 4.5 shows what kind of information was provided regarding where the subjects came from. For the 92 experiments using students as subjects, in 62 (67 %) cases it was stated something about which university these came from. The universities were named in 49 cases and anonymous in 13 cases. The anonymous universities were either explicitly stated to be anonymous or described in terms of "a large US university" or "a large urban university". For 70 experiments, it was stated which field or class the students came from, and for 51 it was stated what kind of course the subjects were recruited from (for instance, "students, taking a graduate course on software engineering"). For the 31 experiments using professionals as subjects, in 21 (68 %) it was stated something about the companies the subjects came from. Five were anonymous, 13 were named and three stated that subjects came from various companies and/or organizations. In only five cases it was stated that the professionals came from more than one company, that is, the subjects came from "several" companies or organizations; the exact number was not given.

Table 4.5: Selection of participants

Description	N	%
Named university	49	41.5
Anonymous university	13	11.0
Varied universities	1	0.9
Named companies	13	11.0
Anonymous companies	5	4.2
Varied companies	3	2.5
Class(es)	70	59.3
Course(s)	51	43.2
Varied courses	1	0.9
None	9	7.6

4.4 Recruitment of Subjects

Whether participation is voluntarily or mandatory may have implications for subject effort generalization. If subjects are volunteers from some population they may be particularly interested in the topic and thus not be a representative sample. Volunteers may also have different motivation for participating compared to non-volunteers:

"For the volunteer subject to feel that he has made a useful contribution, it is necessary for him to assume that the experimenter is competent and that he himself is a "good subject"…Viewed in this way, the student volunteer is not merely a passive responder in an experimental situation but rather he has a very real stake in the successful outcome of the experiment." [21]

Studies have been performed on what characteristics apply to the volunteer subjects compared to the non-volunteers. Volunteers seem to be better educated, more intelligent, more approval-motivated, more sociable, and come from higher social class [21].

For 43 experiments (36 %), this issue was addressed. For 12 experiments the participation was mandatory, for 25 it was not, for three experiments participation was mandatory for some subjects and voluntary for others, and for three experiments the issue was not relevant due to that subjects were not aware that they participated in an experiment.

One paper stated the following: "Subjects were advised that the quality of their work was not being assessed and that their data would be treated confidentially. The intention was to invoke as natural a response as possible under the artificial condition of being videoed." Alleviating subjects' concerns when under close surveillance is, however, not without penalty: their responses may not be typical of the responses they would give in a commercial setting with a rigorous quality control program. Thus, whether or not one wants the subjects to be stressed or not might depend on what settings you want your results to apply to.

A motivation issue is whether the subjects were paid or in other ways rewarded to participate in the experiment. Table 4.6 shows the reported main stimulus used to recruit subjects to take part in experiments. Note that this issue was addressed for only 35 % of the experiments. "No reward" means that the paper explicitly states that no reward was given. One experiment using students only was categorized as "part of job". This was due to students having part time work as staff, and participating in the experiment as staff members.

For the experiments involving students only, in two cases payment was used, in 10 cases participation affected their grade, in nine cases some other kind of reward was given. When the experiment was part of the job (this was usually not explicitly stated) no payment to the company in question was reported. This does not necessarily mean that the company was not paid, but quite often it seems that the company's motivation was that their employees would receive training or gain knowledge.

Table 4.6: Subject categories and type of reward reported

Subject category	Reward	Part of job	No reward	Unknown	Total
Stud only	21	1		60	82
Prof only		14		10	24
Mix type	2		3	5	10
Unknown				2	2
Total	**23**	**15**	**3**	**77**	**118**

How subjects were recruited, is interesting from both a practical and methodological point of view. Recruiting subjects is generally a non-trivial task. In practice, the most convenient approach is to encourage students from the classes taught by university researchers. Hence, the majority of subjects are students. 70 % of the experiments where students participated reported that at least some of the subjects were recruited from one or more courses. The grades of the students were affected by the participation in some cases, and they received extra credits in other cases. In one case they were sponsored a trip to an exhibition. In about 2/3 of the student experiments, the participation was voluntary; in 1/3 it was mandatory. Altogether, in only three cases students were reported to receive payment.

More surprising is that payment to professionals was reported for no experiments. Typically, the experiments with professionals were organized as part of an ordinary project or a training program, and thus were paid implicitly by their ordinary employer. Hence, it seems that in no case the researcher or research team provided extra money to pay the companies or individual for experiment participation. This is in contrast to the model of Simula Research Laboratory, where consultants are paid to take part in experiments [22].

From a methodological point of view, the way subjects are recruited will influence the sample characteristics and thus the population to which one can generalize the findings.

4.5 Individual or Team

Whether subjects work alone or in teams, and what the experimental unit is with respect to analysis and data collection, are two different issues. In 79 experiments subjects *worked* alone on the experimental tasks; see table 4.7. This was in part explicitly stated, but in 49 experiments it was implicit from the nature of the tasks or by the results being presented at an individual level. In eight experiments the tasks were solved in teams, while in the remaining 31 experiments the subjects worked both individually and in teams. Working both individually and in teams could mean that some subjects worked in teams while others worked alone, or it could mean that the same subjects worked first alone and then in teams. The latter applied typically to experiments regarding inspections.

Table 4.7: Subjects working as individuals or in teams

Individual/Team	N	%
Individual	79	66.9
Team	8	6.8
Both	31	26.3
Total	118	100.0

We also analyzed whether data was *collected and analyzed* on individual level or as teams. The results are presented in table 4.8. As expected, the numbers are quite similar. However, even though the numbers are almost identical, the experiments that fit into the different categories in the two tables are not identical although most of them are the same. All experiments where subjects worked alone are in our terms also individual experimental units. In some experiments subjects worked both alone and in teams, but results were only collected on team basis – thus, the experiment would be categorized as "both" in table 4.7 and as team in table 4.8. In others, subjects worked in teams while data was collected both on individual and team basis, and on some occasions subjects worked in teams while data was collected on individual basis. In three papers analysis was performed on virtual (also called nominal) teams, in addition to individually. These are counted as individual when related to experimental units as well as working units, as we have chosen to leave out artificial permutations pulled together from individual results. The number of teams participating in the experiments was reported in all nine experiments with team as experimental unit, whereas in seven of the 27 experiments categorized as "both" it was unclear how many teams participated. This could for instance be due to that the results were specified as number of responses instead of teams when they also specified that some teams had participated in several inspections. Looking into the experiments using virtual teams did not make sense in this respect; in addition to being disregarded as teams by us, they either operated with different number of teams for different analyzes or combined the subjects into close to indefinite number of teams.

Table 4.8: Data collected and/or analyzed on individual or team level

Individual/Team	N	%
Individual	82	69.5
Team	9	7.6
Both	27	22.9
Total	118	100.0

4.6 Differences Between Categories of Subjects

27 experiments used subjects from more than one category of subjects; among these, differences in results between categories were described in eight. In addition, four papers reported replications using a different subject category than in the original experiment. In seven, the differences were between students (mostly undergraduates) and professionals. In one, comparisons were made between students and scientists, and in the remaining four comparisons were made among subjects from the student categories undergraduate and graduate. In these the comparisons may not necessarily be between subjects from the two categories, it may just as well be "inter-student" comparisons made without having categories in mind.

In 81 experiments differences between categories were not relevant, due to that only one category of subjects participated. In 10 experiments it is not known whether there was one or more categories represented.

The experiments reporting differences between categories are not consistent, but they support the notion that students may provide a realistic measure when the experimental task is not complex and does not favor subjects with long experience. One of the papers using undergraduates and professionals as subjects in a maintenance experiment states that "It appears that novices may be somewhat less adept than experts at identifying classes for reuse". Another paper where professionals replicated an earlier study using students performing inspection tasks reported that "For the student population the performances of the Ad Hoc and Checklist methods were statistically indistinguishable. For the professional population, the performance of the Ad Hoc method was statistically superior to that of the Checklist method. For the students, but not the professionals, specification is also significant." Both examples use tasks that are of such a nature that it seems reasonable to assume that experience makes a difference.

4.7 Variation of Subjects within Categories

The variation between subjects from same category is not given much attention in the papers; see table 4.9. 26 % of the experiments address the issue of differences between subjects from the same category, although they may not have found any. All these experiments used only subjects from one category, suggesting that experiments using subjects from different categories are more interested in examining differences between the different categories than looking into differences within the same category.

Table 4.9: Experiments addressing differences between subjects from same category

Category	Address	Total	%	Differences	%
Undergraduates only	8	42	19.0	5	62.5
Graduates only	4	15	26.7	3	75.0
Students, type unknown	2	8	25.0	2	100.0
Professionals only	10	24	41.7	6	60.0
Total	**28**	**106**	**26.4**	**20**	**71.4**

Note that 42 % of experiments using professionals address the issue of differences between subjects, while only 19 % of experiments using undergraduate subjects do the same. Possible explanations may be that experiments using professionals are viewed as more relevant and hence investigated with more interest, or that undergraduates are regarded as a more homogeneous population than professionals.

An experiment using junior and senior professionals shows how experience and skill may influence results: "Still, ANGEL adds value to seniors. Despite the fact that ANGEL in some cases performs poorly, senior people using ANGEL perform well suggesting that seniors still benefit from the tool (compared to the results from the junior group)… Senior estimators are more likely to use their expertise to screen the results from ANGEL and discard the extreme errors. It takes human skill to identify when to trust the output and when to discard it as misleading. The more experienced, and probably more confident, practitioners seem to better judge on this matter…Not surprisingly, seniors estimate more accurately and more reliably than juniors.…We see that the standard deviation and the largest error are much larger for the junior group. This is most notable when estimating without any tools where we see that the junior group benefited most using the tools." These data resembles comparisons made between student and professional populations, and shows that investigating differences within categories may be just as interesting as viewing differences between categories.

4.8 Generalization from Students to Professionals

As most research is aimed at professional populations, we were interested in looking at how the papers using students treated the issue of generalization to the professional population. The results are presented in table 4.10. Note that generalization is discussed in terms of the combination of experiment and paper, as some experiments are reported in more than one paper and the question of generalization may be discussed differently in the different papers. Thus the number of elements in this table adds up to 125 instead of 118.

Table 4.10: Generalization from student populations to professional populations

Generalization	N	%
Generalized	19	15.2
Not generalized	42	33.6
Not relevant	24	19.2
Not discussed	35	28.0
Inconclusive	5	4.0
Total	**125**	**100**

The 24 experiments using professional subjects only are naturally not relevant in this respect. Out of the remaining 101 combinations using students only or a mix of students and professionals, only 19 generalizes the results to the professional population. 5 combinations are categorized as “inconclusive”, as they are discussed in a blurry way. Usually this means that they discuss the topic without taking a clear stand in either direction. Surprisingly, 34 % of the experiments discuss the issue *without* generalizing, and as many as 35 % do not even discuss the generalization issue whatsoever.

4.9 Replications using Students and Professionals

To be able to generalize and validate findings from empirical studies, replications are necessary. This is even more important when you want to generalize to different populations than the ones represented in your experiment.

The simplest form of integration occurs when an experiment is replicated without modification. In this case, the structure of the studies is similar and the integration is relatively straightforward - however, mistakes or biases in the original design will permeate all the studies. Partial replication occurs in cases where an experimenter identifies a problem with the initial experimental design and modifies it before proceeding. According to a lecture given by Barbara Kitchenham [23], we should not encourage the use of same experimental materials (packages etc.) for

replications, as this may make us make the same flaws as the replicated study. It is better to replicate the hypotheses with new material and new procedures.

Table 4.11: Series of replications that comprises use of students and professionals

Subject	Exp. No	Students	Prof	Support	Contradiction	Author
Reading techniques	1	X				
	2	X		X		Same
	3	X			X	Others
	4	X			X	Others
	5		X	X		Same
Perspective-Based Reading	1		X			
	2	X			X	Others
Database referential integrity metrics	1	X				
	2		X	X	X	Same
Maintenance process	1	X				
	2	X		X		Same
	3		X	X		Same

In our survey, 21 experiments are stated by the authors to be replications. Four series of replications have used both students and professionals, see table 4.11. One of the series contained five experiments - two of the four replications supported the original findings, while the other two did not. An interesting finding is that the experiments that were supportive were performed by the author that performed the original experiment. This is also the case for all three series of experiments that found supporting evidence in replication! Looking at it from the opposite side is also interesting – while no others have been able to support findings from the original experiment, only one of the replications performed by the same author has found partly contradictory results. The only exception is one replication finding both supportive and contradictory results. This clearly shows the importance of making independent replications. Not only should procedures and materials be changed; so should the experimenter.

Chapter 5

Discussion

In this chapter I will try to look behind some of the numbers described in chapter 4. The main focus is on the issue of using students vs. professionals as subjects.

5.1 Choosing Subjects

One important aspect of research is to decide whom we want the results to apply to; what the target population is. An experiment is conducted with a certain group of people as subjects, and the results have to be interpreted with this in mind. Your subjects merely represents a *sample* from some population, and in terms of research you never want to state something about just the actual subjects you use in the experiment. Of course, if you perform the study within the context of a company and only want the results to apply for the company this is perfectly OK. Then you can ignore the external validity (whether the results apply to others) without any hesitation. But if you perform the experiment, for instance, within an educational institution and want the results to be of interest for the academic society, you need to address the issue of context and generalization to define the target population. As researchers we want to say something more general, thus we want the results we find to apply to a wider population than just the subjects at hand.

There are differences in sampling between e.g. medical research and SE because biology is expected to be the same for all humans regardless of where they live, while research in SE depends on how you are taught, culture etc. Barbara Kitchenham refers to this as a *sampling problem* [24].

This imposes a great deal of effort on the researchers regarding choosing subjects that are representative of some population. In the software engineering field the population most researchers *want* to generalize to is professional developers working in the industry. It would seem logical that to generalize to such a population you would choose to use representative subjects from the very same population in your experiments, but this is rarely the case. Among the 108 papers we have analyzed, only 24 use professionals as subjects (in addition to seven who uses both students and professionals). In some experiments the professionals were

professionals in another field than software engineering, but since the experiment was related to their field they are still considered professionals.

The material we have studied shows that although the number of experiments performed each year is growing, the number of experiments using professionals as subjects is not; see figure 5.1. In fact, the use of professionals has decreased since its peek in the late 90's. Note that this figure does not include the seven experiments using both students and professionals, nor does it contain the two experiments where the subject categories represented is unknown.

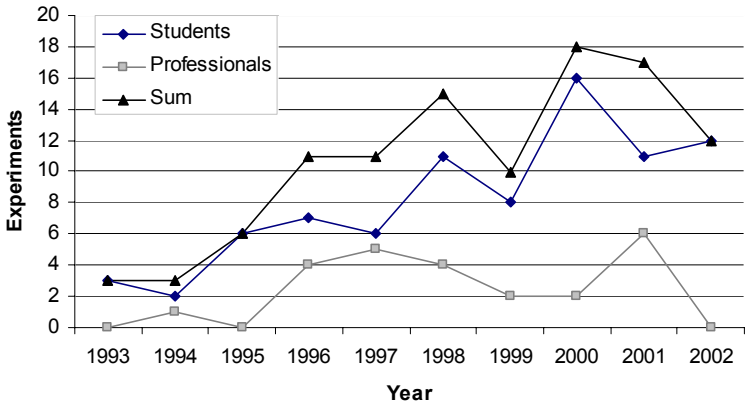

Figure 5.1: Professionals vs. students as subjects

Professional software developers are not very available for research. They claim high wages working in the industry, and their employers usually are unwilling to let them spend working hours doing activities that will not generate money for their company. Hence, researchers usually have to buy their time at regular hour basis or in some other way make this worthwhile for the company (e.g. giving a course or a workshop to educate and train the employees). Most researchers do not have the resources to do this, thus they look to other populations that might do similar use for much less money (or even for free). The natural population to look to is students in the software engineering field, often from classes taught by the researchers. They possess much of the same knowledge as professional developers do; the longer they have studied the more knowledge they have acquired. Students are usually much more available and flexible regarding the time issue. Usually students are less experienced than professional developers, thus, one important question is whether this lack of experience makes them less suited as subjects. Even if the use of students may be questionable, many experimenters use

university students to refine their experimental designs and measurement strategies before using them in industry to reduce costs. This is somewhat similar to using pigs or rats to check whether some medicine seems promising with respect on human beings. One article states:

"This approach also allows us to do a kind of bulk screening of our research hypotheses. That is, we can conduct several studies in university, but only rerun the most promising ones in industry. Intuitively, we feel that hypotheses that don't hold up in the university setting are unlikely to do so in the industrial setting."

It seems the authors feel that there is an asymmetric relation between students and professionals – results collected from a student population can not be generalized to a professional population without further notice, but hypotheses that do not hold for students are generalized to not hold for professionals either. This may imply that students are viewed as less complex than the professional population, stating that negative effects should be generalized while positive effects are questioned. Alternatively this supports the notion that negative results are not of much interest, and that the researchers do not want to spend time and resources confirming them.

Since the target population for most research in the SE field is professional developers working in the industry, one would expect the researchers to discuss how they suppose their results gained through experiments with student subjects also will hold for professionals. Still, almost one third of the papers analyzed do not discuss this. The question is whether this is because the authors do not view this as interesting, have not thought about it or implicit want the reader to infer generalization or not. It is tempting to question whether the authors, by not discussing the question, hope that the reader will view their results as globally more applicable than their results actually may deserve. This would fit in well with the notion that researchers want to report as positive results as possible.

5.2 Negative results

Quite often papers do not report results regarding questions one would expect them to look into, given the setting and issue at hand. This may be because they actually have not looked into them, but it seems reasonable to suspect that this is often due to the results not supporting the stand the researchers have taken. Positive results seem much more appreciated than negative results. There is a forum established to report experiments that did *not* find results supporting the hypotheses the authors wanted to claim, called Forum for Negative Results – apparently there has been *no* submissions to this forum! As Lutz Prechelt, one of the originators, states it in the "Announcement of the Forum for Negative Results (FNR)":

"Due to the current CS publication climate such negative results today are usually camouflaged as positive results by non-evaluating or mis-evaluating the research or by redefining the problem to fit the solution" [25]

Naturally, this takes away valuable insight as negative results should be appreciated just as much as positive results. Not finding the results you look for does not mean the results are not of interest for the community. This indicates that personal pride and self-esteem, being able to show that one's initial questions and hypotheses were correct, may come before the interest in contributing valuable information to the research community. Hence, one might also suspect that expected results not necessarily are looked into with the same scepticism as unexpected results. Positive and negative results should be viewed with the same sound scepticism. If the author will not do it, then at least the reader should be enabled to do it. This emphasizes the need for publishing experimental procedures as well as data material.

5.3 Generalizations

The issue of generalization (also called external validity) is quite often discussed among threats to validity, and experiments conducted with students often recommend that the study should be replicated with professional developers as subjects. However, there have been only a handful of studies comparing the differences in performance between students and professionals, and the results are ambiguous. Among the 107 papers we have analyzed, only seven use both students and professionals as subjects. Table 5.1 gives an overview of these experiments and the significant differences between students and professionals found in them. Only three of these really focus on explicitly comparing the results of these populations, the other four do not even make a point out of having different populations represented. If generalization is such a big deal, students are much more frequently used than professionals as subjects, and we want to say something about professionals working in the industry – why does not more research look into this issue?

There is no obvious answer to this question, except from the money issue. The research that has focused on comparisons between professionals and students shows different results. Some times students and professionals perform equally, some times the professionals perform better and some times the students deliver the best results. In rare cases the results have even shown negative correlations between experience and results, although plausible reasons for this have been given. An example from our material using professionals as subjects states:

"The computed T statistic indicated with a 95 % confidence that those subjects without previous functional testing experience achieved somewhat better test

coverage than those subjects with some functional testing experience. At the 99 % confidence level, no difference in performance was found. No strong conclusion can be drawn, but perhaps the inexperienced subjects were more careful when creating their tests."

What this tells us is that there is no simple answer to whether students are representative for the professional population or not. Some times they are, some times they are not. The question is not whether or not students are representative, but rather *when* they are representative and when they are not. Is it possible to abstract some common traits that make us able to predict when results can be generalized from students to professionals?

It seems that the real differences are usually between inexperienced students (undergraduates) and professionals. As long as the students have a few years of experience from academia (maybe combined with some experience from industry), their results are mostly comparable to those from industrial professionals with minor or no differences to talk about. This depends somewhat on the actual tasks performed. The more abstract or complex the task is, the more likely it seems to be that the professionals will perform better than the students. When tasks are less complex students seem to perform quite similar to the professionals. In most experiments the tasks at hand are smaller and less complex than they are in the real world, due to practical issues like time and money, and due to the fact that the researchers need to keep the setting as controlled as possible. This should favour the students when comparisons to professionals are made.

Sometimes the students actually perform better than the professionals. This could be due to the professionals "surfing" on their experience and routine, thus being a bit "sloppy", while the students are more true to following techniques or guidelines by the book. The same may also apply within categories. One experiment investigating Perspective Based Reading used professionals as subjects. The paper describes the differences like this: "Reviewers with more experience do not perform better than reviewers with less experience. Subjects with less experience seem to follow PBR more closely, while people with more experience were more likely to fall back to their usual technique." It could also be related to the fact that professionals are not professionals in every aspect of the field; sometimes their experience can even be outdated or forgotten because they work with other things.

An experiment using students without stating whether they were undergraduates or graduates shows that also within the student population the less experienced group may perform better: "As expected, subjects in the naive group may have felt that they could adequately represent the causal relationships in their diagrams, and in fact they did a better job of doing so overall than did the knowledgeable group. Subjects in the knowledgeable group may have ignored the causation present in the scenario because causation is not one of the types of relationships analysts are

trained to look for when creating semantic models." Similar comparisons may hold for students versus professionals as well.

It appears that the variance more often is less among the professionals than it is among the students. This may imply that industrial experience make the professionals a more homogeneous group, particularly when compared to students from different levels. There is a big span between students in their first or second year, and advanced students. We need to look behind the terms "student" and "professional" and see what kind of students and professionals we are dealing with when analyzing results.

A large number of studies on the effect of "evaluation apprehension", e.g. Sanders [26], show that an increased awareness of being evaluated seems to increase the effect of "dominant responses", letting instincts override reflection. This effect is basic even for lowlife creatures, as even cockroaches completing a maze performed poorer when other cockroaches were present [17]. This is analogue to subjects and their task abilities. Subjects who are experts at what they do tend to perform better when they are evaluated, whereas subjects who are less competent tends to perform worse when they know they are being evaluated. This may be an issue when selecting experimental subjects, particularly with respect to generalizations. This also means that subjects who are not aware that they are being evaluated may perform different from subjects who know that they are participating in an experiment, which again raises ethical issues about using subjects who are unaware that they are participating in an experiment. These ethical questions are outside the scope of this thesis.

It is interesting to note that many of the papers generalizing from students to professionals refer to the same sources, suggesting that the body of material supporting their generalization might not be substantial. N.K. Liborg's M.Sc. thesis looks into the topic of generalization and target populations in more detail [27].

Table 5.1: Experiments that uses both students and professionals

	Journal	Tot	Ugrad	Grad	Prof	Study	Significant results
Porter A A, Johnson P	TSE 1997	48	0	21	27	Code inspection with three different detection methods.	Not discussed
Burkhardt J, Détienne F, Wiedenbeck S	ESE 2002	50	20	0	30	Evaluate the effect of programmer expertise, programming task (reuse or documentation) and the development of understanding over time on program comprehension. Subjects developed situation models and program models.	*Documentation:* *Situation models:* Professionals significantly better *Program models:* No difference *Reuse:* No difference *Comment lines:* Difference in use
Höst M, Regnell B, Wohlin C	ESE 2000	42	0	25	17	Non-trivial software engineering judgment task involving the assessment of how 10 different factors affect the lead-time of software development projects.	*Conception:* No difference *Correctness:* No difference
Arisholm E, Sjøberg D , Jørgensen M	ESE 2001	36	?*	?*	?*	Changing a Java-program	Not discussed *Category represented, distribution unknown
Ramanujan S, Scamell R W , Shah J R	JSS 2000	100	50	Grad. + prof. = 50		Differences between undergrads vs. grads and professionals regarding software maintenance.	Professionals significantly better
Visaggio G	JSS 1999	90	60	0	30	Analyzes the characteristics of the Quick Fix and Iterative Enhancement paradigms as regards the level of comprehensibility of the resulting system from the maintainer's viewpoint.	Not discussed
Vinter R, Loomes M, Kornbrot D	METRICS 1998	120*	26, unclear if they were undergrad or grad.		22	Identifying linguistic properties of the Z notation which are prone to admit non-logical reasoning errors and biases in trained users.	Not discussed *72 staff subjects

5.4 Categorizing Professionals and Students

Who are students, and who are professionals in this context? Usually students come from higher educations like universities or similar institutions – in our survey, all information provided stated that the students came from universities. In a few cases no such information was given, beyond stating that the subjects were students. Professional is a highly general term, but in SE experiments it usually refers to full-time industrial software developers. Of course, this depends on what the topic for the experiment is. If you want to look into questions regarding e.g. financial planning, it would seem more natural to regard managers rather than developers as professionals. Several papers merely claim to have used professionals (or refer to them in other ways) without stating what they mean by this term.

It may seem simple to categorize students and professionals; students go to school, professionals go to work. However, in terms of research and the role as experimental subjects categorizing is not trivial. If it is going to make sense to use the terms student and professional, this must imply that there are semantic differences between these terms. This raises many questions: When does a recently graduated student convert into becoming a professional? Is it reasonable to differ between a student who is about to finish her studies and a recent graduate who has just started working as a developer in a company? How long experience should the subjects have to be called professionals? When can they be said to be different from the students they were until graduation?

Today it is more common for people to return to school after working a few years. Some students have years of industrial background before they take courses, or work part-time during their studies, so one might ask whether it is fair to view these as regular students. Both populations are quite heterogeneous when we look at their backgrounds, and this makes it even more complicated to generalize between them. There are no clear-cut labels; a student is not merely a student, and a professional not merely a professional.

Also, a professional within some area is not necessarily to be viewed as a professional regarding the task to be solved in the experiment at hand. Sometimes one can actually regard advanced students with recent experiences from the relevant task as more “professional” than the so-called professionals because of extensive and recent experience. One of the papers in our survey describes students as domain experts when estimating their own work. Even if the professionals have relevant experience, this can be outdated or forgotten due to non-use in their everyday work. The paper states that skills may not improve very much with increased experience when there is no proper learning environment.

The professionals' experience and expertise may be directed at the current issue, but it might just as well be general experience from the field. Thus it is important to collect information about the subjects' experience, and to describe this information and its use in material reporting the experiment. In other words, it is important to show how expertise is actually defined and taken into consideration when interpreting the results, particularly regarding generalizations. Generalizations have to be done with great care, and only to settings that resemble the setting the results were produced under.

The borderline between students and professionals is blurry. Maybe is the question not whether to use students versus professionals, but rather whether to use novices versus experts. One of our papers states it this way: "Probably, being a software professional does not imply that the experience matches with the skills that are relevant to the object of study". Thus it might make more sense to use terms as expert or novice related to the task at hand than regarding subjects merely as students or professionals per se.

5.5 When Should Students be Used as Subjects?

The conclusion with respect to the question of whether students or professionals should be used as subjects is that this is highly dependent on the issue studied. As long as the target population is professionals it will always be preferable to use professional subjects, but in some situations students may be more useful than in others. If the issue is basic and the complexity is low, students (in some cases even undergraduates) are likely to give a good measure for how professionals would perform and thus generalization may be well justified. This is also a question of whether you merely want to find some indication of which "treatment" is best, or whether you also want to state something about the relative magnitude of treatments. For more complex issues experience plays an important role, thus one should be very careful making such generalizations.

Sometimes, professional experience may actually be disturbing, because subjects have adopted techniques that they stick to even when they are presented to new methods. In such cases students may be preferred because they have a "pure mind" that can be shaped by the ideas of the researchers. Again, it all depends on which settings your results are aimed at. If you want to say something about how recently educated professionals adopts to certain techniques, you will use other subjects than if you want your results to apply to old timers who have been in the business for decades.

Chapter 6

Threats to Validity

"Researchers have a responsibility to discuss any limitations of their study" [10]. In all experiments there will be threats to validity that needs to be considered and addressed. This will also apply to surveys. The main threats to validity for this study are selection bias, inaccuracy in data extraction and misclassification. In 6.4 I describe some of the problems that occurred during the analysis, before some concluding remarks are stated in 6.5.

6.1 Selection bias

Our survey has focused on controlled experiments, and even though we regard our sample of journals and conference proceedings as leading in software engineering, much research is left out. Our selection of papers is also according to our definition of a controlled experiment; hence, others might end up with a different set of material. Papers have continuously been removed from our database, because they are found to not fit in with our definition and understanding of what is feasible to comprise.

6.2 Data extraction

Extracting data from papers is a non-trivial task, and the lack of a common terminology complicates this even more. Data that may seem obvious to us may thus be misinterpreted. Some data have not been stated explicitly enough, making approximations and best educated guesses necessary, while other data have been extracted between the lines. Our goal has been to analyze articles in an objective manner, but we have been forced to use more or less subjective opinions on several occasions. We have tried to meet this challenge by having at least two persons extracting most of the attributes and discuss their differences. If necessary we have also consulted other people in the project group.

6.3 Misclassification

The lack of a common terminology imposes a threat to validity also regarding classification of data. We know that the terminology is applied differently by different authors, thus making it easy to categorize data wrong. Classification is difficult when there are no standardized labels to sort elements into. We have in many cases had to come up with our own labels based on the type of information

contained in the analyzed papers. As this information is not standardized, we have extracted and classified it according to our own interpretations of what is written in clear text and between the lines.

6.4 Examples of Problems Regarding Analysis

After reading a few articles it became clear that what seemed to be clear and unambiguous definitions when we started out actually were not so clear, and that we would have to make the definitions more explicit and detailed as we kept running into situations where the definitions did not capture the data well enough. Reading articles made us build up a notion of what we could expect to find, and what material we had to fit into our fields and definitions.

Several of the data fields were common for at least two persons analyzing the papers, meaning that we analyzed with respect to these fields and filled in our personal interpretations of these. In addition, the research assistants had also performed an analysis of the papers before the Master students were included in the project. However, the project intends to end up with one resulting database with unique and singular values, requiring us to merge our data into one set of resulting tables. For most of the data this was quite straightforward, but in cases where we had different interpretations, we had to discuss our differences to end up with a common view we could agree upon. This way we were able to double-check most of the data collected, both on errors (typing errors or wrong data) and on "interpretation errors". The latter was useful in respect of getting conscious that not everything is as simple as it may seem when viewed through only one pair of eyes; see figure 6.1. This was an iterative process where new interpretations could force us to go through the field from the start again. Some times we also had to change our interpretations of other fields or add new fields with new interpretations as a result of this.

These discussions were in many cases highly time consuming. Usually the differences were not due to sloppiness or left-hand work, they were results of different interpretations of definitions or the way things were stated in the papers. In many papers the facts and discussions we were looking for were not stated clearly enough, and even if they might have been stated between the lines we were often not able to state these things as facts. Our goal is to at the largest extent report what is actually stated, not speculate what is written between the lines.

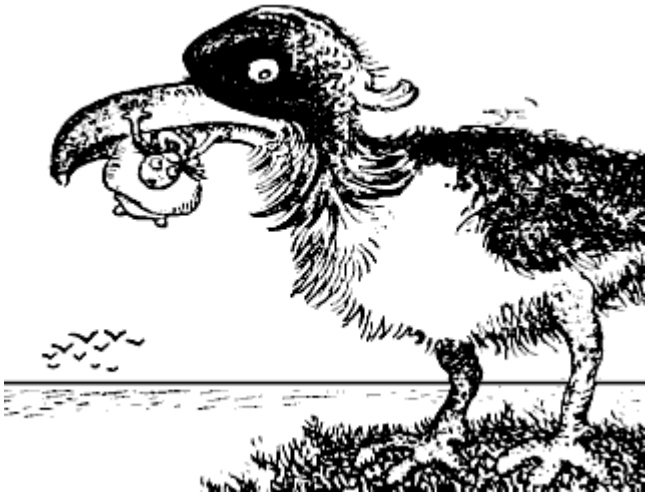

Figure 6.1: One drawing, but what you see and how you interpret it depends on what angle you look at it from.

We have analyzed the papers within different contexts. This means that although we have read the exact same documents, we have done so with different views or glasses. This makes us see the same material from different angles, and raises questions about how the different data fields should be interpreted. There are always several ways to interpret a phenomenon, and it is not necessarily so that one is right and the others wrong. Each member of our project team has based their interpretations on the field definitions the research assistants had produced, and adjusted these according to the other fields they have analyzed and the individual context they have done this in. We did not know the context and the adjusted setting our fellow project members were working within (these might also change along the way), thus our interpretations of the same field may float during our initial reading and analysis. Therefore, we needed to adjust our personal interpretations and agree upon a common definition and understanding to create a unified database. Due to the ambiguous nature of these data it is important to supply definitions and documentation on what our results are based on.

A different kind of problem arose when starting to merge the field "information about participants". This is a text-field without declared standardized alternatives, thus it is not formatted in a strict manner and therefore subject to different interpretations. As we have another field called "selection of participants", and this field contains some information about the participants, the question is whether the information stated in that field should also be inserted into the information-field. The research assistants had standardized the information in "selection of participants" and thus put much more information into "information about participants", while I had split the information in a less structured manner. This startled me, so I went to the field definitions to check whether I had totally

misinterpreted the meaning of the field, but it turned out that the initial analysis by the research assistants did not comply with their own definition. This shows the importance of documenting and updating definitions as they change. As the definitions change in one persons mind they remain clear and unambiguous for this person, but for everybody else this will be confusing. During merging it became an issue whether to use the research assistants' convention or my own. In my opinion, this depends on whether the fields in the database will be viewed separately or in light of related fields. The way I have reported information about participants, this field will in many cases state that no information about the participants were given. This does not necessarily mean that the paper did not state any information about the participants, but rather that it did not state information beyond what was stated in the field "selection of participants". This is an example of how it becomes important not to view each field individually without looking to related fields, and should be considered when using the database to create reports and statistics.

Merging the field "Recruitment" also gave some new experiences. The field was in part used in many different ways, and also inconsistent within each analyzer. The research assistants had in part mixed this up with information from the "Selection of participants" field, while the Master students had focused partly on different information. This was to be expected, as we had different focus. I used these fields as background information for the fields "Paid" and "Mandatory", thus my comments were focused in this direction whereas the other Master student split up information between the main field and the comment.

Even though most fields have been analyzed by at least two persons, the merged data may to some extent be the result of one single person's interpretations. Usually, one person performed the merge of each field alone. If the data did not match, the differences were often resolved by this person alone. In many cases it was quite clear which data was correct (e.g. typing errors or overlooking subtle information), but in some cases this was a question of interpretation. This could also be due to that the field definitions had changed. However, the differences made us conscious of differences of interpretation, and the fact that we found inconsistencies also shows the value of double-checking data.

6.5 Concluding Remarks

Although we are aware that there are threats to the validity of this survey, we feel that we have addressed these issues, and more important - taken actions to minimize them. We have had to use subjective opinions when selecting, extracting and analysing our material, but this has been subject to many discussions and careful evaluations. Others may disagree with us on singular categorizations, but in the big picture I feel confident that the overall results will remain the same. It is important to acknowledge that there are threats, and take them seriously. Being

conscious about them, keeping them in mind during the different phases of any research and to take the proper precautions is the best you can do, along with addressing them when publishing the results of your work.

Looking back, a “Difficult to categorize”-tag could have been valuable. This was also suggested by Barbara Kitchenham [24]. In most cases we have been able to categorize information, but it could be useful to be able to report how many experiments were difficult to categorize on different variables as a quantitative measure on the quality of reporting.

Chapter 7

Conclusions and Future Work

This thesis is a result of the work I have performed within the CONTEXT project. The project looks into different aspects of controlled experiments, and my contribution to the project has been related to issues regarding subjects. In this chapter I draw some conclusions of this work and outline some possible issues for future work.

7.1 Conclusions

Analyzing papers turned out to be more difficult than expected. Even after as much as three individuals had made up individual opinions about how to categorize pieces of information (and merged them into one final dataset), new readers could disagree on how to interpret them. Even elementary and assumed trivial information, as how many subjects participated, could be a question of interpretation. This shows that the way experiments are reported does not adhere to any standards, and indicates that such standards are necessary to clarify which information is to be supplied and the quality of this information. In our opinion many of the papers published reports at a level that is far from good enough, and as the papers analysed are published in major journals and conferences it also appears that the criterions for reviewing and publishing papers need to be revised.

Experimenting with human subjects within the SE field is not very different from social sciences. The differences between subjects and their backgrounds on relevant task variables within SE may often be just as diverging compared to what they may be within e.g. humanistic disciplines, and thus make human factors highly important. It is not always sufficient to report numbers alone, quite often it is necessary to dig deeper. As one of our analyzed papers states it: “It is important to do qualitative studies when our variables of interest are heavily influenced by human behavior (confounding factors). This is recommended in social sciences”.

The way the subjects’ background information is reported varies to such a great extent that it is hard to make general statements. The main categories of information reported was programming experience, working experience, task experience and task related training. The information most often provided is

information regarding task experience (reported in 70 % of the experiments), the second most reported variable was programming experience (32 %). In general, little background information was provided for most subjects.

Although most research aims to produce results that apply to the professional population, only one fourth of the experiments conducted use professionals as subjects, and this number has actually decreased the recent years. Undergraduate students are by far most used as subjects; in 36 % of the experiments they were the only category represented, 52 % of all subjects we can account for were undergraduates and we have good reasons to believe that the majority of the 26 % we have not been able to categorize are the same. This is disturbing, especially since only 19 % of experiments using students as subjects generalize their findings to the professional population while as much as 42 % do *not* generalize. For what is worse, 35 % of experiments using students as subjects actually do not discuss this issue at all.

The heterogeneity of the subjects is generally not paid much attention to in the papers analysed. Most of them do not seem to focus on the diversity in subject backgrounds, and only a few of them report on differences between the individual subjects or between categories of subjects. Runeson [6] suggests that you should not use undergraduates as experimental subjects. Our findings are not that categorical, although it shows that you can not use students as subjects and without further notice claim that the results they provide will also apply to a professional population. In general, the largest differences in performance seem to be between undergraduates and professionals, but this is heavily influenced by the experimental tasks. Our findings indicate that you should generally pay more attention to who you use as subjects; not only in the perspective of categories, but equally important in terms of individuals. The differences within the categories “student” and “professional” may in many cases be of the same magnitude as the differences between them, although there are some indications that the variation between less experienced subjects is greater than between more experienced subjects. Hence, it would in many cases make sense to exchange the use of the terms “students” and “professionals” with “novices” and “experts”.

It is very popular to perform research on inspections, and one might wonder why so many people are into an area that we know relatively much about. Is it because it is easy to explore within a field that there are earlier findings? Maybe because you need to produce results of good quality, or just simply publish *something*? Maybe is it even because you want to make a commercial profit on consultancy work. It would be naive to think that all research is conducted to serve the community.

Through this survey we have focused on certain traits and variables that in our opinion are essential when reporting controlled experiments, and thus should be

addressed. Failing to report on many of these variables may not be done on purpose, it is more an indication that there are no established and detailed standards on how to report the way experiments are run in the SE field. We are quite aware that there are limitations to what is possible to fit within a paper submitted for publication in a journal (some papers refer to other publications for more details on results, procedures etc.), but certain characteristics are required to claim credibility for the results provided. Scientific journals are read by scientists and other people with a critical eye; hiding bad design and lack of results by silence is not the best way of making an impression.

Some papers merely report their conclusive results without giving any description of the data these results are based on. Certain information is rarely provided, and often this is claimed to be due to commercial reasons. Quite often the companies participating requests confidentiality, and researchers often protect their data from being used by others. One may question whether researchers have the right motivation for publishing their findings. It should be in their interest to have others look into their work with a critical eye.

We are not at liberty to criticize the authors for not describing facts the way we would like to see them. We can, however, point the finger at the need for standards regarding what we feel is important information when reporting an experiment. As empirical research in the SE field is young, there is in most cases no established task taxonomy to classify by, and according to Dag Sjøberg [24], checklists are often made by the authors themselves. Some standards for how an experiment should be reported have been defined, but these are quite new and not yet substantially adhered to. It is surprising how few details some papers actually report – for instance, some articles do not state anything about who the subjects were.

The importance of replicating studies is clearly shown by the series presented in this study. The series of replications that used both students and professionals as subjects showed that *all* replications that supported the original experiment were performed by the author of the original experiment! They also showed that replications conducted by others than the original author always found contradicting results.

This should make it pretty obvious that independent replications are necessary. According to Barbara Kitchenham [24], the baseline for quality of reporting is as follows: “Can you replicate the study from the information provided in the paper?” In many papers this is not nearly the case, thus the software engineering community has work to do.

The bottom line seems to be that controlled experiments in the software engineering field are in a “crawling phase”; like a child who needs to learn to crawl

before it can learn to walk. Few controlled experiments have been conducted, and the lack of procedures and a common terminology makes it necessary to start building from the ground. The body of controlled experiments we have studied can serve as a very useful foundation to learn from and build on, and in this respect studies using students can serve as a starting point. I believe our work may contribute as well.

7.2 Future work

The analyses of our source material have produced a large amount of data, and the ways to view and combine them are endless. The CONTEXT project has not yet come to an end, and the data may be enhanced and used in different ways later. We have focused on the way the experiments are reported, not the quality of the experimental designs or the statistical analyses and validity of results. This survey has investigated papers that adhere to our definition of a controlled experiment. This accounts for as little as 2 % of all papers published in our selection of journals and conferences. Similar work can be done on case-studies, usability-studies and other kinds of empirical research.

Bibliography

1. V. Basili, R.W. Selby, and D. Hutchens, “Experimentation in software engineering”, *IEEE Transaction on Software Engineering*, July 1986

2. D. Rombach, “Experimental software engineering issues: Critical assessment and future directions”, *Dagstuhl Workshop*, September 1992

3. V. Basili, D. Rombach, and R.W. Selby, “The experimental paradigm in software engineering”, *Experimental Engineering Issues: Critical Assessment and Future Directions, International Workshop*, vol. 706, no. 8, 1993

4. N. Fenton, “How effective are software engineering methods?”, *Journal of Systems and Software*, vol. 22, no. 2, 1993

5. W.F. Tichy, P. Lukowicz, L. Prechelt, and E.A. Heinz, “Experimental evaluation in computer science: A quantitative study”, *Journal of Systems and Software*, vol. 28, no. 1, pp. 9-18, 1995

6. V. Basili, “The role of experimentation in software engineering: past, current, and future”, in *Proc. of the 18th International Conference on Software Engineering (ICSE)*, 1996

7. W.F. Tichy, “Should computer scientists experiment more? 16 excuses to avoid experimentation”, *IEEE Computer*, vol. 31, no. 5, pp. 32-40, May 1998

8. M.V. Zelkowitz and D. Wallace, “Experimental models for validating technology”, *Theory and Practice of Objects Systems*, vol. 31, no. 5, pp. 23-31, May 1998

9. D.I.K. Sjøberg, V. By, J.E. Hannay, O. Hansen, A. Karahasanovic, N.K. Liborg, A.C. Rekdal, “A Survey of Controlled Experiments in Software Engineering”, *paper in progress for submission to IEEE Transaction on Software Engineering*, 2004

10. 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 Transaction on Software Engineering*, vol. 28, no. 8, pp. 721-734, August 2002

11. P. Runeson, "Using students as Experiment Subjects – An analysis on Graduate and Freshmen Student Data", *Proceedings 7th International Conference on Empirical Assessment & Evaluation in Software Engineering*, 2003

12. R.L. Glass, I. Vessey, and V. Ramesh, "Research in software engineering: an analysis of the literature", *Journal of Information and Software Technology*, vol. 44, no. 8, 2002

13. A. Zendler, "A preliminary software engineering theory as investigated by published experiments", *Empirical Software Engineering*, vol. 6, no. 2, 2001

14. M. Shaw, "Writing good software engineering research paper: Minitutorial", in *Proc. Of the 25th International Conference on Software Engineering (ICSE)*, 2003

15. I.S. Deligiannis, M. Shepperd, S. Webster, and M. Roumeliotis, "A review of experimental investigations into object-oriented technology", *Empirical Software Engineering*, vol. 7, no. 3, 2002

16. N. Juristo, A.M. Moreno, and S. Vegas, "Reviewing 25 years of testing technique experiments", *Empirical Software Engineering*, vol. 9, March 2004

17. M. Jørgensen, "A review of studies on expert estimation of software development effort", *Journal of Systems and Software*, vol. 70, issues 1-2, pp 37-60, 2004

18. W.R. Shadish, T.D.Cook, and D.T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin, 2002

19. O. Laitenberger, K. El Emam, and T.G. Harbich, "An internally replicated quasi-experimental comparison of checklist and perspective based reading of code documents", *IEEE Transaction on Software Engineering*, vol. 27, May 2001

20. T.D. Cook and D.T. Campbell, *Quasi-Experimentation. Design & Analysis Issues for Field Settings*, Houghton Mifflin, 1979

21. R. Rosnow, R. Rosenthal, *People studying people: Artifacts and Ethics in Behavioral Research*, W.H. Freeman and Company, 1997

22. D. Sjøberg, B. Anda, E. Arisholm, T. Dybå, M. Jørgensen, A. Karahasanovic, E. Koren and M. Vokác, "Conducting Realistic Experiments in Software Engineering", *ISESE'2002 (First International Symposium on Empirical Software Engineering)*, Nara, Japan, October 3-4, 2002, pp. 17-26, IEEE Computer Society

23. B. Kitchenham, "Preliminary guidelines for empirical research in software engineering", *lecture held at Simula Research Laboratories*, May 2004

24. CONTEXT project presentation, arranged by D. Sjøberg during visit by B. Kitchenham, Oslo, May 2004

25. L. Prechelt, "Why we need an explicit forum for negative results", *Journal of Universal Computer Science*, vol. 3, no. 9, September 1997

26. G.S. Sanders, "Self-presentation and drive in social facilitation", *Journal of Experimental Psychology*, vol. 20, no. 4, pp. 312-322, 1984

27. N.K. Liborg, "A study of threats to validity in controlled software engineering experiments", *Master thesis at University of Oslo*, 2004

Appendix A

Field description for database

Fieldname	Description	Who
Total_number_of_participants	How many subjects participated originally in the experiment.	Ove
Comments_on_total_number_of_participants	Comments on the number of subjects, whether it was unclear, has been estimated etc.	Ove
Active_participants	How many subjects actually participated and were included in the analyses. <= Total_number_of participants.	Ove
Comments_on_active_participants	Comments on active_participants. Describes why active_participants is less than total_number_of_participants or whether it was unclear how many actually participated.	Ove
Students	How many students participated (supercategory including undergraduate_students, graduate_students).	Ove
Undergraduate_students	How many undergraduate students participated.	Ove
Graduate_students	How many graduate students participated.	Ove
PhD	How many PhD students participated. This is a subgroup to graduate students, meaning that this number is also included in graduate_students.	Ove
Participating_scientists	How many scientists participated as subjects. Scientists are professors, staff at educational institutions.	Ove
Professionals	How many professionals participated.	Ove
Replication	"Yes" if the experiment is a replication, "No" otherwise.	Ove and NK
Comments_on_replication	Comments to replication, e.g. which experiment is replicated	Ove and NK
Individual_or_team	"Individual" if subjects worked only alone, "Team" if subjects worked only in teams and "Both" if subjects worked both individually and in teams or there was a mix of subjects working individually and in teams.	Ove
Comments_on_individual_or_team	Supplementary information about why individual_or_team was categorized as it is.	Ove
Selection_of_participants	Who the participants are (students, professionals etc.), where they come from (university, company etc.), course/training session etc. (if this is relevant).	Ove and NK
Comments_on_selection_of_participants	Comments to selection_of_participants. Mostly supplementary information.	Ove and NK
Information_about_participants	Information about subjects background.	Ove
Comments_on_information_about_participants	Supplementary information	Ove
Recruitment	How the participants were recruited (e.g. as part of a course).	Ove and NK

Comments_on_recruitment	Comments to recruitment. Mostly supplementary information.	Ove and NK
Paid_rewarded	Whether the subjects where paid or rewarded in some way.	Ove and NK
Mandatory	Whether experiment participation was mandatory or voluntarily.	Ove and NK
Differences_of_group_members	Are there different results within the same subject category?	Ove
Comments_on_differences_of_group_members	Supplementary information.	Ove
Categories_of_subjects	Which combination of subject categories are represented in the experiment.	Ove
Study_unit	Which study unit is data collected and analyzed for. Individual, team or both	Ove
Comments_on_study_unit	Supplementary info about the choice in study_unit	Ove
Number_of_teams	The number of teams that participated in the experiment.	Ove
Comments_on_number_of_teams	Supplementary information about how many teams participated.	Ove
Different_results_between_categories	Different results between the different subject categories that participated.	Ove
Comments_on_different_results_between_categories	Supplementary information	Ove
Generalizations_from_students	Whether the students in the samples are generalized to professionals.	Ove and NK
Comments_on_generalization_from_students	Comments to generalizations_from_students. Mostly supplementary information and quotes from the article.	Ove and NK

Appendix B

SAS output

The SAS System

The MEANS Procedure

Analysis Variable : Active_participants

Categories_of_subjects	N Obs	Sum	Mean
Undergraduates only	42	2644.00	62.9523810
Graduates only	15	377.0000000	25.1333333
Undergraduates and graduates	17	976.0000000	57.4117647
Students, type unknown	8	524.0000000	65.5000000
Professionals only	24	439.0000000	18.2916667
Undergraduates and professionals	2	150.0000000	75.0000000
Graduates and professionals	2	90.0000000	45.0000000
Graduates and scientists	1	34.0000000	34.0000000
Students, type unknown and scientists	1	12.0000000	12.0000000
Undergraduates, graduates and scientists	1	18.0000000	18.0000000
Undergraduates, graduates, scientists and professionals	1	20.0000000	20.0000000
Students, type unknown, professionals and scientists	1	120.0000000	120.0000000
Type of participants unknown	2	43.0000000	21.5000000
Undergraduates, graduates and professionals	1	36.0000000	36.0000000

TABLE O1: Number of experiments with categories of subjects represented (all combinations of categories that are represented)

	Frequency	Percent
Undergraduates only	42	35.59
Graduates only	15	12.71
Undergraduates and graduates	17	14.41
Students, type unknown	8	6.78
Professionals only	24	20.34
Undergraduates and professionals	2	1.69
Graduates and professionals	2	1.69
Graduates and scientists	1	0.85
Students, type unknown and scientists	1	0.85
Undergraduates, graduates and scientists	1	0.85
Undergraduates, graduates, scientists and professionals	1	0.85
Students, type unknown, professionals and scientists	1	0.85
Type of participants unknown	2	1.69
Undergraduates, graduates and professionals	1	0.85
Total	118	100.00

TABLE O6: Number of subjects in experiments

	Frequency	%	mean	std	min	median	max
Students only	82	69.5	55.1	56.7	6	36	266
Professionals only	24	20.3	18.3	13.9	4	17	68
Students and professionals	7	5.9	59.4	36.4	20	48	120
Other	3	2.5	21.3	11.4	12	18	34
Unknown	2	1.7	21.5	17.7	9	22	34
Total	118	100.0	46.5	50.9	4	30	266

TABLE O8b: Reward for taking part in experiments

	Experiments		Participants	
	n	%	n	mean
Unknown	77	65.3	3512	45.6
Part of job	15	12.7	307	20.5
Grade	10	8.5	693	69.3
Extra credits	9	7.6	660	73.3
Unpaid	3	2.5	166	55.3
Paid	3	2.5	121	40.3
Other reward	1	0.8	24	24.0
Total	118	100.0	5483	46.5

TABLE 09: Number of mandatory participants in experiments

	Frequency	%	mean	std	min	median	max
??	75	63.6	44.3	51.5	4	26	266
For some	3	2.5	70.7	42.4	40	53	119
No	25	21.2	53.2	61.2	9	32	242
Not relevant	3	2.5	8.7	2.5	6	9	11
Yes	12	10.2	49.4	22.5	24	45	88
Total	118	100.0	46.5	50.9	4	30	266

TABLE 012:

Category	Year									
	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Undergraduates only	1	.	.	2	3	6	7	8	6	9
Graduates only	1	2	1	2	2	2	.	3	2	.
Undergraduates and graduates	1	.	5	3	.	1	.	4	1	2
Students, type unknown	.	.	.	.	1	2	1	1	2	1
Professionals only	.	1	.	4	5	4	2	2	6	.
Undergraduates and professionals	.	.	.	.	.	.	.	1	.	1
Graduates and professionals	.	.	.	.	1	.	.	1	.	.
Graduates and scientists	.	.	.	1	.	.	.	.	.	.
Students, type unknown and scientists	.	.	.	.	.	.	.	1	.	.
Undergraduates, graduates and scientists	.	.	.	.	.	.	1	.	.	.
Undergraduates, graduates, scientists and professionals	.	.	.	.	.	.	1	.	.	.
Students, type unknown, professionals and scientists	.	.	.	.	.	1	.	.	.	.
Type of participants unknown	.	.	.	.	1	.	1	.	.	.
Undergraduates, graduates and professionals	.	.	.	.	.	.	.	.	1	.
Total	3	3	6	12	13	16	13	21	18	13

TABLE O13: Use of study units

	Both	Ind+V	Individual	Team
Undergraduates only	9	.	31	2
Graduates only	4	.	11	.
Undergraduates and graduates	1	2	14	.
Students, type unknown	2	.	4	2
Professionals only	8	1	11	4
Undergraduates and professionals	.	.	2	.
Graduates and professionals	1	.	1	.
Graduates and scientists	.	.	1	.
Students, type unknown and scientists	.	.	1	.
Undergraduates, graduates and scientists	1	.	.	.
Undergraduates, graduates, scientists and professionals	1	.	.	.
Students, type unknown, professionals and scientists	.	.	1	.
Type of participants unknown	.	.	1	1
Undergraduates, graduates and professionals	.	.	1	.
Total	27	3	79	9

TABLE O14: Subjects paid or rewarded

	Credit	Credit for	Grade	Paid	Part of job	Reward	Unknown	Unpaid
Undergraduates only	6	1	4	.	.	1	30	.
Graduates only	.	.	3	1	1	.	10	.
Undergraduates and graduates	.	.	2	.	.	.	15	.
Students, type unknown	1	.	1	1	.	.	5	.
Professionals only	.	.	.	.	14	.	10	.
Undergraduates and professionals	1	.	.	.	.	.	1	.
Graduates and professionals	.	.	.	.	.	.	2	.
Graduates and scientists	.	.	.	.	.	.	.	1
Students, type unknown and scientists	.	.	.	.	.	.	.	1
Undergraduates, graduates and scientists	.	.	.	.	.	.	1	.
Undergraduates, graduates, scientists and professionals	.	.	.	.	.	.	1	.
Students, type unknown, professionals and scientists	.	.	.	.	.	.	.	1
Type of participants unknown	.	.	.	.	.	.	2	.
Undergraduates, graduates and professionals	.	.	.	1	.	.	.	.
Total	8	1	10	3	15	1	77	3

TABLE O14b: Mandatory participation

	??	For some	No	Not relevant	Yes	Total
						N
Undergraduates only	31	.	5	.	6	42
Graduates only	8	1	3	.	3	15
Undergraduates and graduates	9	2	4	.	2	17
Students, type unknown	4	.	3	.	1	8
Professionals only	17	.	4	3	.	24
Undergraduates and professionals	1	.	1	.	.	2
Graduates and professionals	2	.	.	.	.	2
Graduates and scientists	.	.	1	.	.	1
Students, type unknown and scientists	.	.	1	.	.	1
Undergraduates, graduates and scientists	1	.	.	.	.	1
Undergraduates, graduates, scientists and professionals	1	.	.	.	.	1
Students, type unknown, professionals and scientists	.	.	1	.	.	1
Type of participants unknown	1	.	1	.	.	2
Undergraduates, graduates and professionals	.	.	1	.	.	1
Total	75	3	25	3	12	118

TABLE O15: Information about subjects

Description	Frequency	Percent
Explicit age	6	5.08
Explicit grade info.	1	0.85
Explicit info about gender	6	5.08
Explicit info about training	8	6.78
Explicit programming experience (numbers of particular info)	20	16.95
Explicit task experience (detailed or extensive info)	41	34.75
Explicit workin experience (numbers of particular info)	14	11.86
General demographic info	3	2.54
General grade info	5	4.24
General info about age (no numbers)	1	0.85
General info about gender (no numbers)	2	1.68
General info about training	29	24.58
General programming experience	18	15.25
General task experience	42	35.59
General working experience	11	9.32
None	15	12.71
Training as part of course	2	1.69
Total	224	189.83

TABLE O16: Selection of participants

Description	Frequency	Percent
Anonymous companies	5	4.24
Anonymous institution	13	11.02
Class(es)	69	58.47
Course(s)	51	43.22
Named companies	13	11.02
Named institution	49	41.53
None	9	7.63
VCL	1	0.85
Varied companies	3	2.54
Varied courses	1	0.85
Varied institutions	1	0.85
Total	215	182.20

TABLE O18: Use of subject categories in experiments

Category of participants		Frequency	Percent	No of participants				
				Mean	Std	Min	Median	Max
Students only		82	68.91	55.1	56.7	6	36	266
	Undergraduates only	42	35.29	63.0	61.3	10	45	266
	Graduates only	15	12.61	25.1	11.1	9	24	48
	Undergraduates and graduates	17	14.29	57.4	57.5	6	31	208
	Students, type unknown	8	6.72	65.5	70.3	13	43	231
Professionals only		24	20.17	18.3	13.9	4	17	68
Mixed type of participants		10	8.40	48.0	35.4	12	39	120
	Undergraduates and professionals	2	1.68	75.0	35.4	50	75	100
	Graduates and professionals	2	1.68	45.0	4.2	42	45	48
	Graduates and scientists	1	0.84	34.0	.	34	34	34
	Students, type unknown and scientists	1	0.84	12.0	.	12	12	12
	Undergraduates, graduates and scientists	1	0.84	18.0	.	18	18	18
	Undergraduates, graduates, scientists and professionals	1	0.84	20.0	.	20	20	20
	Students, type unknown, professionals and scientists	1	0.84	120.0	.	120	120	120
	Undergraduates, graduates and professionals	1	0.84	36.0	.	36	36	36
Unknown		2	1.68	21.5	17.7	9	22	34

Total		118	99.16	46.5	50.9	4	30	266

TABLE O20: Individual or team

Category	Frequency	Percent
Both	31	26.27
Individual	30	25.42
Individual (implicit)	49	41.53
Team	7	5.93
Team (implicit)	1	0.85
Total	118	100.00

TABLE O21: Recruitment of subjects

Category	Frequency	Percent
From courses	3	2.54
No info given	35	29.66
Part of a course	59	50.00
Part of a course and volunteers	2	1.69
Part of course and professionals	1	0.85
Part of their work	6	5.08
Part of training	1	0.85
Part of training course	4	3.39
Part of workshop	3	2.54
Private invitation	2	1.69
Subjects were recruited in different ways	1	0.85
Volunteers	1	0.85

Total	118	100.00

TABLE O21: Recruitment of subjects
Table of Categories_of_subjects by ipart

Categories_of_subjects
ipart

Frequency Percent Row Pct Col Pct	Training as part of course	General info about age (no numbers)	General demographic info	General grade info	General programming experience	General info about gender (no numbers)	General task experience	General info about training	General working experience	Total
Students only	2 0.89 1.29 100.00	1 0.45 0.65 100.00	3 1.34 1.94 100.00	5 2.23 3.23 100.00	17 7.59 10.97 94.44	2 0.89 1.29 100.00	32 14.29 20.65 76.19	20 8.93 12.90 68.97	9 4.02 5.81 81.82	155 69.20
Professionals only	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.45 2.13 5.56	0 0.00 0.00 0.00	6 2.68 12.77 14.29	7 3.13 14.89 24.14	1 0.45 2.13 9.09	47 20.98
Students and professionals	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 0.89 13.33 4.76	1 0.45 6.67 3.45	0 0.00 0.00 0.00	15 6.70
Other	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.45 20.00 2.38	1 0.45 20.00 3.45	0 0.00 0.00 0.00	5 2.23
Unknown	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.45 50.00 2.38	0 0.00 0.00 0.00	1 0.45 50.00 9.09	2 0.89
Total	2 0.89	1 0.45	3 1.34	5 2.23	18 8.04	2 0.89	42 18.75	29 12.95	11 4.91	224 100.00

TABLE O21: Recruitment of subjects
Table of Categories_of_subjects by ipart

Categories_of_subjects
ipart

Frequency Percent Row Pct Col Pct	None	Explicit age	Explicit grade info.	Explicit programming experience (numbers of particular info)	Explicit info about gender	Explicit task experience (detailed or extensive info)	Explicit info about training	Explicit working experience (numbers of particular info)	Total
Students only	10 4.46 6.45 66.67	3 1.34 1.94 50.00	1 0.45 0.65 100.00	9 4.02 5.81 45.00	4 1.79 2.58 66.67	25 11.16 16.13 60.98	5 2.23 3.23 62.50	7 3.13 4.52 50.00	155 69.20
Professionals only	4 1.79 8.51 26.67	1 0.45 2.13 16.67	0 0.00 0.00 0.00	8 3.57 17.02 40.00	1 0.45 2.13 16.67	11 4.91 23.40 26.83	2 0.89 4.26 25.00	5 2.23 10.64 35.71	47 20.98
Students and professionals	1 0.45 6.67 6.67	2 0.89 13.33 33.33	0 0.00 0.00 0.00	2 0.89 13.33 10.00	1 0.45 6.67 16.67	3 1.34 20.00 7.32	1 0.45 6.67 12.50	2 0.89 13.33 14.29	15 6.70
Other	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.45 20.00 5.00	0 0.00 0.00 0.00	2 0.89 40.00 4.88	0 0.00 0.00 0.00	0 0.00 0.00 0.00	5 2.23
Unknown	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 0.89
Total	15 6.70	6 2.68	1 0.45	20 8.93	6 2.68	41 18.30	8 3.57	14 6.25	224 100.00