University of Oslo
Department of Informatics

Measuring time – A Systematic Survey of Controlled Experiments in Software Engineering

Espen A.Vikskjold

Master of Science Thesis

December 2005

Abstract

To conduct controlled experiments is the classical method for identifying cause-effect relationships. This helps researchers evaluate and validate their research results. This master thesis quantitatively systematically surveys the present practices of reporting and using time in 113 controlled software engineering experiments published in 103 scientific articles in the period 1993-2002. The survey can be regarded as an extension of the study of Sjøberg et al. [6], who investigated several other aspects of the same collection of controlled experiments.

The need for an analysis of time in controlled SE-experiments was recognized based on the results of an earlier survey conducted by Jo E. Hannay at SIMULA Research Laboratory. He identified that time often is reported as dependent variable. Due to the dependent variable's important role in experiments as provider of the effect construct in the causal relationship, it was relevant to find out how time is used.

The investigation of time in controlled SE-experiments focuses on the following aspects; the overall frequency of reporting time, the use of time, the recording of time, the specification of time in terms of time units, and the validity threats that time can constitute. These aspects were regarded as the most important features of time, and hence, would give a thorough and appropriate picture of time in controlled SE-experiments.

The main results of the survey of the aspects of time, is first of all, that time is reported in a substantial majority of controlled SE-experiments. Second, in most of the controlled SE-experiments time is used as dependent variable, or more precise, as measure of subjects' experimental tasks. However, in an extensive number of controlled SE-experiments time is not explicitly described as dependent variable, although it is used in that sense. Third, time measures are found to be recorded by subjects, experimenters and tools. An interesting finding in relation to this aspect is that the largest proportion of controlled SE-experiments did not explicitly report their time recording. Fourth, a majority of controlled SE-experiments measure the time in minutes. Last, but not least, very few controlled SE-experiments were found to address time in relation to validity threats.

The main conclusions and recommendations of the research, is first of all, that SE-researchers documenting controlled experiments should explicitly describe time as dependent variable in the experimental design, when time is used as dependent variable. This is obviously important due to the vital role of a dependent variable in experimentation. Second, SE-researchers should consistently describe how time is recorded in controlled experiments, due to validations of the experiment results and to possible replications. Third, SE-researchers should specify time measures in minutes, because it provides an appropriate granularity of the time data. And fourth, aspects of time constitute threats to both internal and external validity, and SE-researchers should therefore assess time in relation to possible validity threats of the controlled experiments.

The hope of this master thesis is that the results, discussions, conclusions and recommendations of the research provide useful insight to SE-researchers, so that they can improve the design of time in their controlled experiments. This is vital in order to advance a state-of-the art practice in empirical software engineering research.

Acknowledgements

When looking back on the work with this master thesis, it is clear that I owe a great deal of gratitude to certain people. I would especially like to thank my supervisor, Amela Karahasanovic, who has given me valuable guidance and support throughout the work with the thesis. Moreover, I am grateful to Jo E. Hannay for providing results about dependent variables. Last, but not least, a special thanks goes to my family and my girlfriend, Liva, for always being there and believing in me.

Oslo, December 2005
Espen A.Vikskjold

Table of Contents

List of Tables

List of Figures

1 Introduction

This section introduces the master thesis. Section 1.1 describes the context, the focus and the purpose of the research. Section 1.2 states the objective of the research. Section 1.3 briefly presents the project to which the master thesis is a contribution. Section 1.4 outlines the structure of the thesis.

1.1 Motivation

In the field of SE (Software Engineering), not enough attention has been shown in doing controlled experiments [2]. There are several reasons for this. Probably, the most fundamental one is that much of the SE community is not conscious of the need of discussing facts and claims supported by data, i.e. running experiments [2]. It is stressed that the SE community should focus more on experimentation. This is due to the fact that they are confronted with an increasing number of options for producing software, and that necessitates proof that a specific approach or technique is really better than another. As Juristo and Moreno [2] states, experimentation will help researchers within SE to increase their understanding of what makes software good and how to make software well. This can be achieved, since controlled experiments help researchers evaluate and validate their research results [8].

Controlled experiments are empirical methods. Other empirical methods are case studies and surveys. However, while case studies and surveys are both qualitative and quantitative methods, experiments are purely quantitative [9]. The reason is that they are based on measuring changes caused by different variables [2]. During these investigations, quantitative data are gathered, to which mathematical methods can be used to get formal results.

> "A randomised experiment or a quasi-experiment in which individuals or teams (the experimental units) conduct one or more software engineering tasks for the sake of comparing different populations, processes, methods, techniques, languages or tools (the treatments)."
>
> [6, p.4]

The quote above represents an operational definition of controlled experiments in SE. As it states, controlled experiments are conducted to compare various techniques, methods, working procedures, etc. Based on the outcomes of the comparisons, it becomes scientifically possible to state whether, for example, a method or tool is more effective than another method or tool.

The reason why these experiments are called controlled is that they are conducted in a laboratory setting, which gives a high level of control. The basic principles of an experiment are illustrated in Figure 1. The figure shows that the starting point of the experiment is that the researchers have a theory of a cause and effect relationship, which they formulate to a hypothesis. This hypothesis of a causal relationship is based on an assumption that the cause precedes the effect, that the cause is related to the effect, and that there is no reasonable explanation for the effect other than the cause [4]. The intention of the experiment is then to test this hypothesis, i.e. the experiment operation in Figure 1. This is done by manipulating the supposed cause (the treatment) and observing an effect (the outcome) afterwards [4]. In this way, the researchers can determine whether variation in the cause is connected to

variation in the effect, and thereby, make conclusions about the relationship between the cause and effect.

Figure 1 – Experiment principles [9, p.32]

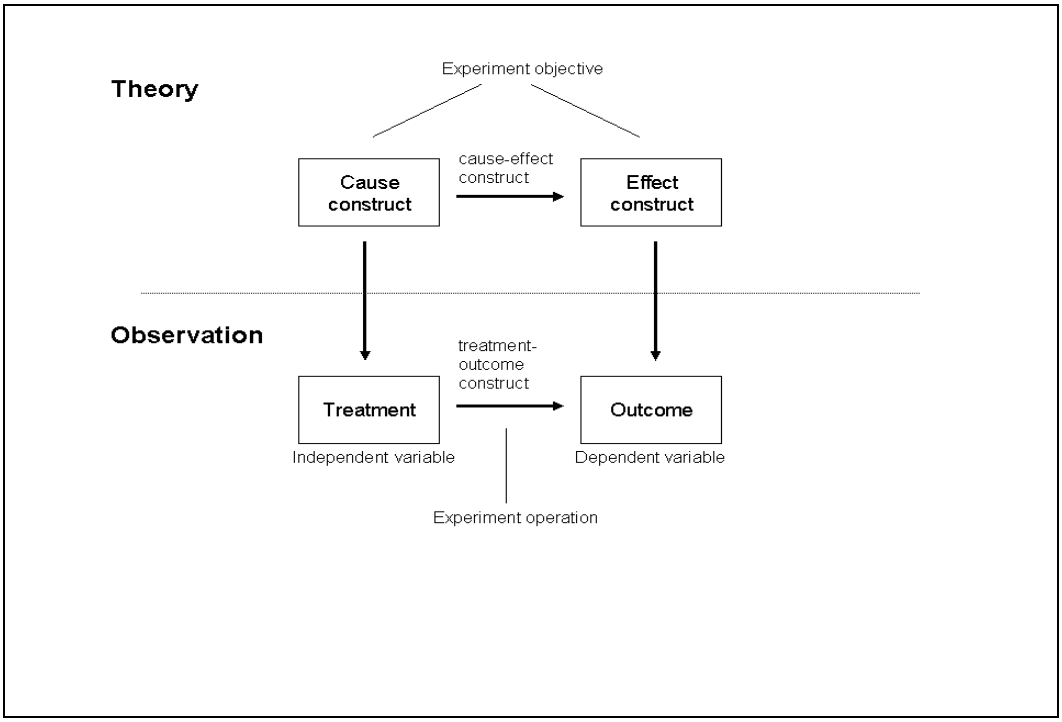

Figure 1 further shows that the treatment is referred to as independent variable and the outcome is referred to as dependent variable. The relationship between independent and dependent variables is more closely illustrated in Figure 2.

Figure 2 – Illustration of independent and dependent variables [8, p.922]

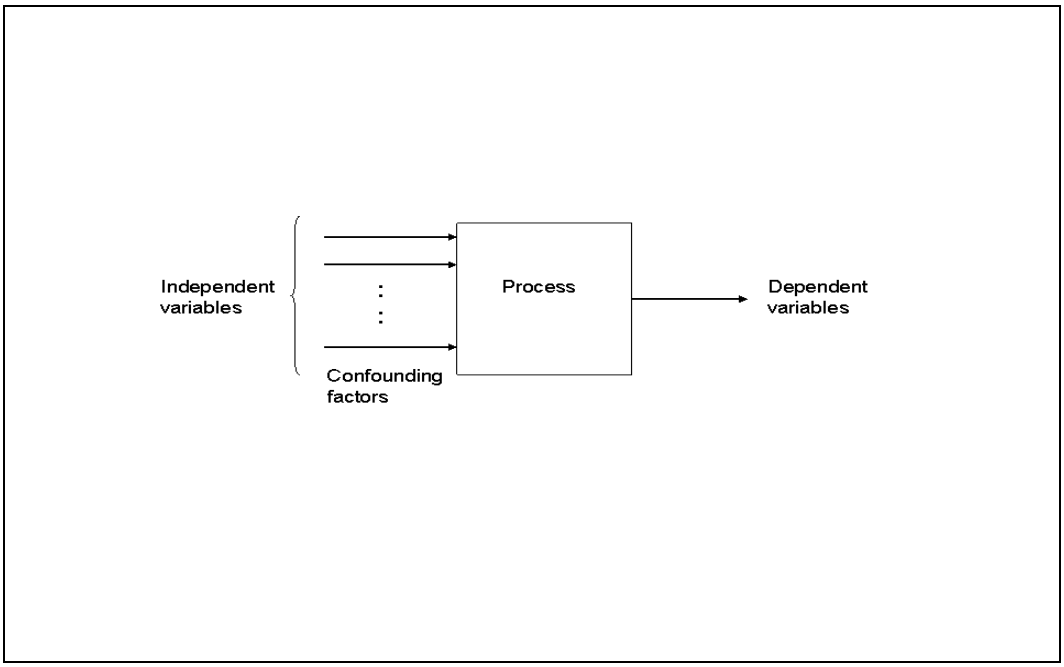

Figure 2 mentions independent variables, confounding factors, experiment process, and dependent variables. An independent variable is the treatment in the experiment process that is manipulated and controlled [9]. The main types of independent variables are:

- Population – e.g. evaluating/comparing novices' skills with experts' skills.
- Process – e.g. evaluating/comparing software process models
- Method – e.g. evaluating/comparing programming paradigms like recursive and iterative constructs.
- Technique – e.g. evaluating/comparing inspection techniques.
- Tool – e.g. evaluating/comparing testing tools.

A dependent variable, which is also called response variable [2], is the outcome that the conductors of the experiment want to study to see if it is influenced by the independent variable. Mainly, the dependent variable in a controlled experiment depends on the objective and hypothesis of the experiment in question [2]. Examples of dependent variables are:

- Time – e.g. completion time measured in minutes from the experimental task was given to its completion.
- Correctness – e.g. measured by assigning points and counting correct solutions from the experimental task.
- Effort – e.g. measured by man-hours from start to finish of the experimental task.
- Efficiency – e.g. measured by execution time of the experimental task.

Dependent variables can be divided into direct and indirect measures. A direct measure can be defined as a measure that does not depend upon a measure of any other attributes [9]. An indirect measure is a measure that is derived and calculated from direct measures. An example of an indirect measure is programmers productivity, which is derived from a calculation of numbers of line of code (direct measure), divided by the programmer's effort, i.e. time used (direct measure).

The following quote can be used as an example to clarify the independent and dependent variables, and how they interact: "…, if we want to quantify the time saving when using as opposed to not using a code generator, the response variable to be measured would be the time taken with both alternatives to program certain specifications" [2, p.70]. In this case, the alternatives of using a code generator and not using it in connection with programming tasks constitute the independent variables, and the time taken is the dependent variable that is measured to compare the most effective one (i.e. the most time-saving) of the two independent variables.

Confounding factors, which Figure 2 mentions, are variables that may affect the dependent variables without the knowledge of the conductors of the experiments [8]. The experiment process, which is represented by a box in Figure 2, comprises of the different phases of the experiment work. There are four main phases: 1) the definition of the objectives of the experiment, 2) the design of the experiment, 3) the execution of the experiment, and 4) the analysis of the results/data collected from the experiments [2]. In the first phase, the general theory is transformed into the hypothesis that the experiment is going to test. The second phase involves constructing a form of plan that the experiment will follow, often referred to as the experimental design. In this phase, the factors that represent the independent and dependent variables are defined. In the third phase, the experiment is run according to the design. The fourth phase presents the results of the data that were collected during the experiment.

The focus of the research of this master thesis was to quantitatively analyse how time was used in 113 controlled SE-experiments, which were described in 103 scientific articles published in the period 1993-2002. We only wanted to explore the use of time in the sense of a direct measure. This meant that we did not want to consider the use of time in cases where it was used in calculations with other measures, i.e. in indirect measures.

The specific aspects of time that we wanted to focus on were, first of all, the overall frequency of reporting time in controlled experiments. Here, we wanted to show the overall relevance of the theme. The second aspect of time was the use of time. Here, we wanted to describe how time was actually referred to and used, so that possible differences and similarities could be identified. The third aspect of time was time recording. The purpose of investigating that aspect was to give insight into how time measures were recorded in the experimental activities, in order to identify possible trends and problems. The fourth aspect of time was time units. The purpose of that investigation was to identify and evaluate the time units the controlled experiments specified their time in. The fifth and last aspect of time was time as validity threat. Here, we wanted to investigate how time was regarded as validity threat, in order to clarify how time could limit the validation of the research results. Overall, the reason why we chose to focus on these five aspects of time was that we regarded them as the main aspects of time in controlled experiments. In other words, our view was that this would provide a deep, and nearby complete understanding of the existing practices of using time.

The most important reason for the research of time must be seen in light of the results of an earlier analysis conducted by Jo E. Hannay at SIMULA RL (Research Laboratory). He summarised all the dependent variables in the 113 controlled experiments. The study showed that time, together with correctness, occurred most often as the dependent variables. This meant that measures of time often represented the outcomes of the experiment operations, and thereby, were of decisive significance to the conclusions regarding the research questions. The central role of time that the study of Hannay revealed necessitated a further investigation into the use of time.

Another reason for the research was that time as an aspect of controlled SE-experiments had not been analysed before, at least as far as we know. Several other aspects of controlled SE-experiments had been examined by Sjøberg et al. [6] (see section 2).

The outcomes of the research would be quantitative information about the practices of reporting and using time in controlled experiments, discussions about the identified practices, and our recommendations for better practices. A deeper understanding of the role of time in controlled experiments would most likely give SE-researchers a greater awareness of how they should deal with time in controlled experiments, and perhaps thereby, improve their design of controlled experiments.

1.2 Objective

The objective of the research of the master thesis was 1) to identify and quantitatively describe how controlled SE-experiments reported time, and 2) to discuss existing practices.

To approach the objective, the following aspects of time would be investigated:

- The overall frequency of reporting time in controlled experiments
- The use of time
- The recording of time
- The specification of time in terms of time units
- Time as validity threat

The master thesis was a short thesis. This meant that it lasted from the 22nd of August 2005 to the 19th of December 2005, and counted for 30 points in the master degree.

1.3 Research context

The master thesis was a part of the project "Research Methods and Support Tools for Conducting Empirical Research in Software Engineering" at SIMULA RL [5]. The purpose of this project is to advance the state-of-the art of empirical software engineering research. More precise, the research problem was how to develop infrastructures, apparatus and methods for conducting experiments and other empirical studies in software engineering that would significantly advance the state of the art. This was important in order for the private and public IT-industries to develop better IT-systems with fewer resources.

The contribution of this master thesis to the project described above was to add knowledge to methods, i.e. in terms of measuring time, that are used in the conducting of recent controlled SE-experiments. This would provide a foundation for a development of a state-of-the-art practice regarding use of time in controlled experiments.

1.4 Structure of thesis

The remainder of this master thesis has been structured in the following way:

- Section 2 presents related work.
- Section 3 explains the methodology of the survey. Moreover, some central sides of the execution of the survey are described.
- Section 4 presents and discusses the results of the survey.
- Section 5 deals with the validity of the survey.
- Section 6 summarises and concludes the research of the thesis, and gives proposals for future work.

2 Related work

Table 2.1 present four major surveys (the first four) of empirical studies in SE, and a recent survey of several aspects of controlled experiments (Sjøberg et al. 2004). The survey of this master thesis is included in the table in order to compare it with the others. The table summarises the purpose, scope, journals, sampling of articles (extent), and number of examined articles of all the surveys.

The first three surveys gave a broad overview of research methods applied in SE. Tichy et al. [7] compared the frequency of experiments (i.e. type of empirical study) published in a few computer science journals and conference proceedings with the frequency of experiments published in a journal on artificial neural network and a journal on optical engineering. Zelkowitz and Wallace [10] suggested a taxonomy of empirical studies in SE and described a survey in which 612 articles were categorised within this taxonomy. Glass et al. [1] surveyed 369 articles in terms of topic, research approach, research method, reference discipline and analysis-level. Although, these surveys varied in purpose, selection criteria and taxonomies, the following common conclusion could be drawn from them: most of the published articles in computer science and SE gave limited or no experimental validation, and the amount of controlled experiments was especially low. In Zendler [11], a survey of 31 experiments was reported. Here, the purpose was to work out an initial SE-theory.

While the first three surveys in Table 2.1 reported the extent and characteristics of different types of empirical study, the survey of Sjøberg et al. [6] gave a thorough study of controlled experiments. It focused on several aspects of controlled SE-experiments. These were the kind of technology that were investigated in the experiments (the topics of the experiments), the subjects that participated in the experiment, the tasks they carried out, the type of application systems on which these tasks were executed, the environments in which the experiments were run, the frequency of replications, and the degree to which external validity is discussed.

The survey of this master thesis can be viewed as an extension of Sjøberg et al. [6].

Table 2.1 – Summary of surveys of empirical studies in SE
(The original table presented in [6, p.3] and extended here by the outer right column)

	(Tichy et al. 1995)	(Zelkowitz & Wallace 1997)	(Glass et al. 2002)	(Zendler 2001)	(Sjøberg et al. 2004)	This survey
Purpose	Compared the frequency of empirical studies in computer science with other fields.	Categorised empirical studies in SE and validated the taxonomy of empirical studies suggested by the authors.	Surveyed topics, research approaches, research methods, reference disciplines and level of analysis.	Worked out an initial SE theory from the results of different SE experiments.	Surveyed aspects related to controlled SE-experiments. These aspects were topics, subjects, tasks, environments, replication, and external validity.	Surveyed aspects of time in controlled SE-experiments.
Scope	Comp. sci, incl. SE	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.	A variety of journals and conference proceedings.	EASE, EMSE, ICSE, IEEE Computer, IEEE Software, ISESE, IST, JSME, JSS, METRICS, SP&E, TOSEM, TSE	EASE, EMSE, ICSE, IEEE Computer, IEEE Software, ISESE, IST, JSME, JSS, METRICS, SP&E, TOSEM, TSE
Sampling of articles	1991-1994, one to four volumes per journal, random selection of work published by ACM in 1993.	All articles in 1985, 1990 and 1995.	Every fifth article in the period 1995-1999.	Not reported.	All articles in the period 1993-2002	103 articles from the period 1993-2002
No. of examined articles	403	612	369	49 articles considered, 31 articles analysed more thoroughly.	5453 articles scanned, 103 articles analysed more thoroughly.	103 articles (113 controlled experiments) inspected. 85 controlled experiments analysed more thoroughly.

3 Methodology

This section focuses on the methodology of the research of this master thesis. Section 3.1 clarifies the principles of the research method. Section 3.2 describes the process of identifying the articles that were investigated. Section 3.3 describes how the research method was applied.

3.1 Research method - systematic survey

The research method was a systematic survey, or a systematic review as it is also called. In conducting a systematic survey one evaluates, interprets, and summarises "all available research relevant to a particular research question or topic area or phenomenon of interest" [3, p.1]. Thus, either more general conclusions about the research question, topic area or phenomenon of interest can be drawn, or it can encourage further research activities within the field.

The most important advantage of doing a systematic survey is that they can provide knowledge about the outcome of some phenomenon across various situations and empirical methods [3]. Another advantage, which is related to quantitative studies, is the possibility of combining data using meta-analysis techniques. This may increase the probability of identifying real effects that individual smaller studies are not able to detect.

The activities of a systematic survey can be divided into three stages; planning, conducting and reporting the survey [3].

In the planning stage, the recognition of the need for a survey is established. Moreover, a survey protocol is developed. This protocol states the methods that will be applied to carry out the systematic survey. The purpose is to reduce the probability of researcher bias.

In the conducting of the systematic survey, the first activity is to identify the research, or, more precisely, the potentially relevant primary studies. Here, it is vital to determine and follow a search strategy for this identification process. The second activity is the selection of the primary studies that are actually relevant. When the relevance is decided upon, the third activity is to evaluate the "quality" of the primary studies. The fourth activity is the data extraction. Here, it is important to have designed a data extraction form in order to accurately record information. The fifth and last activity in the execution of the systematic survey is the data synthesis. This activity includes gathering and summarising the results of the chosen primary studies. Elements of the second, fourth and fifth activity should as far as possible be specified in the survey protocol in the planning stage.

In the reporting stage of the systematic survey, the data are interpreted and presented. Moreover, the generalization of the conclusions and limitations of the survey are debated, and recommendations for practices and research are stated.

3.2 Identification of articles that reported controlled experiments

Researchers at SIMULA RL [6] found the articles that were surveyed. The purpose of this process was to identify and extract controlled experiments. Originally, the process started out with 5453 articles. These scientific articles were published in a sample of nine journals and three conference proceedings in the period 1993 to 2002. The nine journals, which were regarded as leaders in SE, were ACM Transaction on Software Engineering Methodology (TOSEM), Empirical Software Engineering (EMSE), IEEE Computer, IEEE Software, IEEE Transactions on Software Engineering (TSE), Information and Software Technology (IST), Journal of Systems and Software (JSS), Software Maintenance and Evolution (SME), Software: Practice and Experience (SP&E). The three conference proceedings were the International Conference on Software Engineering (ICSE), IEEE International Symposium on Empirical Software Engineering (ISESE), and IEEE International Symposium on Software Metrics (METRICS). In addition, partially included amongst the conference proceedings was the conference Empirical Assessment & Evaluation in Software Engineering (EASE), since ten selected articles from EASE appeared in special issues of JSS, ESE, and IST.

A researcher systematically went through the 5453 articles' titles and abstracts. If the title and abstract did not clearly indicate that the article described a controlled experiment, the researcher and a second person in the project team read the whole article. In the end, 103 articles were selected. These articles reported a total of 113 controlled experiments. In twelve of the 103 articles, more than one controlled experiment was described, and four of the 113 controlled experiments occurred in several articles.

3.3 Application of the systematic survey

3.3.1 Purpose of the survey

The purpose of the systematic survey was to discover and summarise different aspects related to time in 113 controlled SE-experiments. Several aspects, i.e. extent, topic, subjects, task and environment, replication, and external validity, of the controlled experiments in the 103 articles had been studied by SIMULA RL, but time remained to be analysed.

3.3.2 Preparing the survey

Before conducting the systematic survey of the 113 controlled experiments, the attributes of time (referred to as aspects of time in other section of the thesis) had to be defined. Table 3.1 presents the attributes and the possible values. The attribute that first and foremost needed to be examined was "Time reported". If time was reported, the other attributes would be investigated.

Table 3.1 – Attributes and values of time

Attribute	Values
Time reported	Yes/no.
Use of time	Dependent variable, measure of effort, measure of effectiveness, measure of efficiency, time limit.
Time unit	Seconds, minutes, hours, days, weeks.
Time recording	Tools, subjects, experimenters.
Time as threat to validity	Threats to internal/external validity.

In order to systematically organise the information that was found during the analysis, a database-table was made which can be viewed as an extension of the Context-database. Table 3.2 represents the database-table with attributes and corresponding descriptions. This embodied the data that it was going to be searched for and gathered from the articles.

Table 3.2 – Attributes of time in the extension of the Context-database

Attribute	Description
Author	The author and the title of the article.
Year	The publication year of the article.
Topic category	A general topic classification scheme of Glass et al. [1] by which a study at SIMULA RL [6] had structured the controlled SE-experiments.
Use of time	The controlled experiment's use of time. A previous investigation at SIMULA RL, conducted by Jo E.Hannay, had examined and registered all dependent variables of the 113 controlled experiments into the Context-database. In the survey of this thesis, this was used as guidance.
Time recording	The controlled experiment's means of recording time measures.
Time units	The controlled experiment's specification of time in terms of time units.
Validity	The controlled experiment's assessment of time in regards to the validity of the experiment.
Comments	Some possible additional comments about time in the controlled experiment.

Prior to the execution of the survey, a tentative timetable was defined. The plan was to use one month, i.e. 20 working days, on the analysis. This meant that six articles needed to be investigated per day in average.

3.3.3 Performing the survey

Only one person, i.e. the author of this master thesis, conducted the systematic survey. Because of the strict time limit of the master thesis, the whole content of the articles could not be read. Therefore, as a pilot study, the entire content of five articles was studied at first in order to get an impression of how they were structured. Based on this knowledge, it was decided upon that the articles' abstracts, experimental designs (or similar chapters), and conclusions would be the focus of the reading for the rest of the articles. In addition, the usefulness of the search function in Acrobat Reader for scanning the articles for relevant terms was recognized. The most important search terms that were used were "time", "dependent variable", "outcome", "measure", "effectiveness", "effort", "efficiency", "seconds", "minutes", "hours" and "validity".

4 Results and discussion

This section presents and discusses the results. Section 4.1 presents the overall frequency of the controlled experiments that reported time. Section 4.2 describes how time was referred to and used. Section 4.3 describes how time was recorded. Section 0 describes how time was specified in terms of time units. Finally, section 4.5 shows how time was considered as validity threats.

4.1 Overall frequency of reporting time

Table 4.1 – Overall categorisation of controlled experiments: Time reported/not reported

Overall experiment category	Topic category	No.	%
Time reported in controlled experiments	Software life-cycle/engineering	35	30.9
	Methods/techniques	34	30.0
	Project/product management	7	6.2
	Tools	2	1.8
	Database/warehouse/mart organization	2	1.8
	Methodologies	2	1.8
	Measurement/metrics	1	0.9
	Personnel issues	1	0.9
	Programming languages	1	0.9
	Computing research	-	-
		85	**75.2**
Time not reported in controlled experiments	Software life-cycle/engineering	15	13.3
	Methods/techniques	7	6.2
	Project/product management	3	2.6
	Tools	1	0.9
	Database/warehouse/mart organization	1	0.9
	Methodologies	-	-
	Measurement/metrics	-	-
	Personnel issues	-	-
	Programming languages	-	-
	Computing research	1	0.9
		28	**24.8**
Total		**113**	**100.0**

Table 4.1 shows the overall categorisation of the 113 controlled experiments that were described in the 103 articles. The outer left column is titled “Overall experiment category”. This column comprises of two categories; the controlled experiments that reported time and the controlled experiments that did not report time. The second column from left is titled “Topic category”. This column consists of the main topics of the articles. The topics were based on the classification scheme of Glass et al. [1], and represent categories of controlled experiments in general SE research. The organization of the 103 articles into topics was done in the work of Sjøberg et al. [6], so this was our source of information. When we talk about topics of the articles in the master thesis, we refer to this topic classification. The columns to the right in the table represent the numbers of controlled experiments and the corresponding percentage.

The table shows that 85 (75.2%) of the 113 controlled experiments reported time while the rest, 28 controlled experiments (24.8%), did not report any use of time. The large proportion of controlled experiments that reported time should indicate that time was a frequently used factor in SE-experimentation.

The topic categorisation in Table 4.1 shows that the largest numbers of controlled experiments that reported time were in the fields of "Software life-cycle/engineering" (35 controlled experiments; 30.9%) and "Methods/techniques" (34 controlled experiments; 30%). Amongst the controlled experiments that did not report time, "Software life-cycle/engineering" (15 controlled experiments; 13.3%) was the dominating topic. Compared to the controlled experiments in "Methods/techniques" that reported time, the numbers of controlled experiments in "Methods/techniques" that did not report time, were proportionally quite a bit fewer. This may tell us that it is more common to use time as measure of the subjects' experimental activities in controlled experiments evaluating/comparing object-oriented design methods, which is a prominent category within "Methods/techniques", than in controlled experiments evaluating code inspections and walkthroughs, which is a prominent category within "Software life-cycle/engineering".

4.2 Use of time

This section describes how time was used in the 85 controlled experiments that reported time.

Table 4.2 – Categorisation of controlled experiments: Use of time

Experiment category	Topic category	No.	%
Time as dependent variable	Software life-cycle/engineering	18	21.2
	Methods/techniques	22	25.9
	Project/product management	1	1.17
	Tools	1	1.17
	Database/warehouse/mart organization	1	1.17
	Measurement/metrics	1	1.17
	Programming languages	1	1.17
		45	**52.9**
Time as measure of outcome	Software life-cycle/engineering	13	15.3
	Methods/techniques	12	14.1
	Project/product management	4	4.7
	Database/warehouse/mart organization	1	1.17
	Personnel issues	1	1.17
		31	**36.5**
Time as time limit only	Software life-cycle/engineering	4	4.7
	Project/product management	2	2.4
	Tools	1	1.17
	Methodologies	2	2.4
		9	**10.6**
Total		**85**	**100.0**

Table 4.2 shows that we categorised the 85 controlled experiments into three groups according to how they explicitly reported and used time. These groups are "Time as dependent variable", "Time as measure of outcome", and "Time as time limit only".

The first experiment category which we call "Time as dependent variable" included the controlled experiments that explicitly described time as dependent variable in their

experimental designs. This meant that they measured the duration of the subjects' execution of the experimental activities in order to evaluate/compare the effect on the independent variables (inspection techniques, programming paradigms, design methodologies etc.).

The second experiment category which we call "Time as measure of outcome" comprised of the controlled experiments that did not explicitly report time as dependent variable in their experimental designs. But, they used time as measure of the subjects' execution of the experimental tasks, and thus, used time for the same purpose as the first experiment category.

The third experiment category which we call "Time as time limit only" consisted of the controlled experiments that only reported time in terms of time limits. This group of controlled experiments differed from the others, since they did not use time as a measure of the experimental activities. They simply used time to limit the execution time of the activities. In the following sections, 4.2.1-4.2.3, the three categories of controlled experiments is presented in more depth.

As Table 4.2 states, a majority of the 85 controlled experiments reported time as categorised in the experiment category "Time as dependent variable" (45 controlled experiments; 52.9%). "Time as measure of outcome" was the second largest experiment category (31 controlled experiments; 36.5%). The experiment category "Time reported as time limit only" was substantially smaller than the other two, constituting of just nine controlled experiments (10.6%).

When it comes to the topic categories, Table 4.2 shows that "Software life-cycle/engineering" and "Methods/techniques" were the prominent topics in the two experiment categories "Time as dependent variable" and "Time as measure of outcome". In "Time as dependent variable", "Methods/techniques" was the largest topic (22 controlled experiments; 25.9%), followed closely by "Software life-cycle/engineering" (18 controlled experiments; 21.2%). In "Time as measure of outcome", the numbers for the topics "Software life-cycle/engineering" (13 controlled experiments; 15.3%) and "Methods/techniques" (12 controlled experiments; 14.1%) were almost the same. A majority of the controlled experiments in the experiment category "Time as time limit only" dealt with the topic "Software life-cycle/engineering" (four controlled experiments; 4.7%). The other controlled experiments in this experiment category were more evenly divided between the other three topics that were represented here.

4.2.1 Time as dependent variable

This section explains how time was referred to and used amongst the controlled experiments in the experiment category "Time as dependent variable". As stated above, the 45 controlled experiments in this experiment category had in common that they explicitly described time as dependent variable. This meant that time was measured to provide an outcome (effect) of the experimental activities, which could be used to determine the hypothesis of a causal relationship (see section 1.1 for more about independent and dependent variables). Another similarity amongst the controlled experiments in this experiment category was that the statements of time as dependent variable were found in the sections concerning the experimental designs.

Although the overall purpose of measuring time was the same, there were differences in regards to how they actually referred to and used time. In Table 4.3, we have categorised the

various kinds of time measures that were registered amongst the controlled experiments in the experiment category "Time as dependent variable".

Table 4.3 – Time measures in controlled experiments

Time measure category	No.	%
Change/modification/maintenance time	10	20.4
Inspection/review time	9	18.4
Designing/coding/testing/debugging/recording time	5	10.2
Time needed/taken/spent	4	8.2
Answering/questionnaire time	4	8.2
Completion time	3	6.1
Performance time	3	6.1
Response time	3	6.1
Development time	2	4.1
Comprehension time	2	4.1
Preparation time	2	4.1
Search time	1	2.0
Meeting time	1	2.0
Total	**49**	**100**

As the table tells us, 13 different kinds of time measures were found, and a total of 49 controlled experiments were registered. The reason why there are 49 controlled experiments in Table 4.3, when Table 4.2 showed that there were only 45 controlled experiments in all that reported time as dependent variable, is that three controlled experiments reported and used several time measures as dependent variables. Two out of the three controlled experiments used two time measures in the following combinations: "Change/modification/maintenance time" and "Comprehension time", and "Inspection/review time" and "Preparation time". The last of the three controlled experiments used three time measures: "Inspection/review time", "Answering/questionnaire time", and "Preparation time". In the following, the time measure categories in Table 4.3 will be described. At the end, the findings will be discussed.

The largest category of time measures was "Change/modification/maintenance time", comprising of 10 controlled experiments (20.4%). This group of time measures represented controlled experiments that measured the time of tasks related to identifying places for modification in software and/or doing changes in software. The purpose of the measuring of time was to compute the change/modification/maintenance effort of the software.

In the time measure category "Inspection/review time", there were nine controlled experiments (18.4%). These had in common that they investigated inspection techniques by measuring the duration of performing inspection/review tasks in order to determine the most efficient one. The controlled experiments varied in terms of whether the tasks were carried out individually or in groups.

As Table 4.3 shows, the time measure category "Designing/coding/testing/debugging/recording time" was the third largest with five controlled experiments (10.2%). This category included controlled experiments that measured the time of designing, coding, testing, debugging and/or recording tasks, so that they could

resolve the effort of these operations in relation to different software development techniques or methods.

"Time needed/taken/spent" was a more complex time measure category that comprised of four controlled experiments (8.2%). These controlled experiments had different experimental tasks but were placed in the same category, because they explicitly said that time needed, taken or spent (which can be viewed as synonyms) were the dependent variables. Time was measured due to cost or effort assessments.

Also the time measure category "Answering/questionnaire time" constituted of four controlled experiments (8.2%). In these controlled experiments, the duration of experimental tasks related to answering questions were measured. The controlled experiments differed in regards to whether they measured the time of answering each question or the total time of answering all the questions. The reasons for measuring the tasks, i.e. effort or efficiency, were not explicitly stated in the experimental design of these controlled experiments.

The three controlled experiments (6.1%) that reported "Completion time" as a dependent variable, measured the time of completing the experimental tasks. The purpose was to define the efficiency of the tools, methods/techniques or programming languages that were investigated.

The three controlled experiments (6.1%) in the time measure category "Performance time" measured the duration of the experimental assignments, so that they could determine the task performance efficiency.

Similar to "Completion time" and "Performance time", three controlled experiments (6.1%) were categorised in the time measure category "Response time". These controlled experiments stated response time as a dependent variable. In the context of the experimental activities, this meant that they measured the time from which the subjects were asked a question to they answered it. The intention was to measure comprehensibility.

There were two controlled experiments (4.1%) that measured time according to the time measure category "Development time". These controlled experiments had in common that they measured the total working time of developing a program. Total development time was reported as a measure of effort.

Also the time measure category "Comprehension time" included two controlled experiments (4.1%). In these controlled experiments, the time the subjects used to understand the system or code was measured. As mentioned earlier in this section, three controlled experiments were found to declare more than one time measure as a dependent variable in their experimental design. One of the two controlled experiments in the time measure category described here used two time measures, namely "Comprehension time" and "Change/modification/maintenance time". The measuring of comprehension was done before the time measuring of the maintenance activities. The purpose of combining these time measures was probably to assess how the effort of doing maintenance tasks was affected by adding a period for comprehension first.

Two controlled experiments (4.1%) reported "Preparation time" as dependent variable. These controlled experiments measured the time of reading documents related to the experimental tasks before performing the tasks. As stated earlier in this section, "Preparation time" was

used together with other measures in two controlled experiments. That means that the two controlled experiments mentioned here, reported several time measures as dependent variables in their experimental design. One used two time measures, “Inspection/review time” and “Preparation time”, and the other used three time measures, “Inspection/review time”, “Answering/questionnaire time” and “Preparation time”. Although the representation is small, this can indicate that time measures concerning preparation activities, which “Preparation time” points to, are accompanied by other time measures.

The least frequently used time measures were “Search time” and “Meeting time”. They only occurred in one controlled experiment (2%) each. The controlled experiment that used the time measure “Search time” measured and compared the search time of four different methods to examine the efficiency of them. The controlled experiment that expressed “Meeting time” as a dependent variable, measured the time the subjects spent in meeting during the experiment.

As the various time measure categories prove, the controlled experiments that reported time as dependent variable, referred to the measuring of time in different ways. One could say that in most cases the names of the time measures reflected the experimental activities. For a reader, this is obviously beneficial in order to understand exactly what time is measuring. On the other hand, the reader may question why the duration of the experimental activities is measured. As the descriptions above shows, most of the controlled experiments reported the purpose of measuring time. To resolve the effort or efficiency of the experimental tasks related to the treatment in question, were most often presented as the purpose of the time measures. This indicates that the controlled experiments also issued why time was used.

However, our view is that time measures should be more uniformly labelled when they are stated as dependent variables in the experimental designs. For example, the time measure could be named according to the purpose, i.e. as an efficiency or effort measure, and then a more thorough description of the time measure could follow. This would probably provide a clearer comprehension of the reason for using the time measure, and hence, increase the understanding of the influence the time measure has on the conclusions regarding the research questions (the hypothesis of the controlled experiment). A more consistent terminology of time measures in controlled experiments would especially be beneficial for possible replications. If time measures are vaguely described in a controlled experiment, they may be used differently in a replication. This may result in other outcomes causing different conclusions regarding similar research questions. Obviously, this is something we do not wish.

4.2.2 Time as measure of outcome

This section explains how time was referred to and used amongst the controlled experiments in the experiment category “Time as measure of outcome”. These controlled experiments had in common that they used time as measure of the subjects’ execution of the experimental activities, in order to provide an outcome for determining the independent variable. This use of time is identical with the use of time in the controlled experiments in the experiment category “Time as dependent variable” described above. However, the 31 controlled experiments addressed here, differed from the controlled experiments in the section above, by not explicitly describing time as dependent variable. Moreover, the descriptions of the time measures were to a large extent found in other sections of the articles than sections

concerning the experimental design. That is why they these two groups of controlled experiments were divided.

Similar to the controlled experiments in section 4.2.1, the terminology and use of time measures amongst the controlled experiments in "Time as measure of outcome" differed. In Table 4.4, we have summarised the time measures in 16 categories according to how time was referred to. A majority of the categories are only represented by one controlled experiment. In the following, each time measure category will be described briefly. At the end of the section, the findings will be discussed.

Table 4.4 – Time measures in controlled experiments

Time measure category	No.	%
Completion time	6	19.4
Project work time	4	12.9
Inspection/review time	4	12.9
Defect discovery time	2	6.45
Development time	2	6.45
Time spent/taken	2	6.45
Efficiency measure	2	6.45
Comprehension time	1	3.22
Coverage speed	1	3.22
Error correction time	1	3.22
Meeting time	1	3.22
Work flow time	1	3.22
Requirements analysis/module design time	1	3.22
Verification time	1	3.22
Basic metric	1	3.22
Disturbing factor	1	3.22
Total	**31**	**100.0**

As the table shows, the largest time measure category was "Completion time" comprising of six controlled experiments (19.4%). Typical for these controlled experiments were that they measured the speed of completing the experimental tasks.

Four controlled experiments (12.9%) were found to measure time in the sense of "Project work time". This meant that they measured the time of performing activities related to project phases. The purpose was to investigate the work-effort of each phase.

There were also four controlled experiments (12.9%) in the time measure category "Inspection/review time". These controlled experiments had in common that they measured time of inspection or review tasks. Both effort and efficiency estimations were stated as causes for measuring time.

As the name "Defect discovery time" implies, this time measure category consisted of two controlled experiments (6.45%) that measured the time the subjects used to find defects in programming code. The reason for measuring time was to determine the effectiveness of the subjects.

"Development time" included two controlled experiments (6.45%) that measured time in connection with designing, coding and debugging activities. Effort measures were stated as the intention of measuring time in these activities.

The two controlled experiments (6.45%) registered in the time measure category "Time spent/taken", said that they measured the time spent or taken on performing the experimental tasks. One of the controlled experiments stated that the purpose of measuring time was to measure the effort.

One of the two controlled experiments (6.45%) in "Efficiency measure" measured the time of the subjects performing queries in third and fourth normal form data organization for relational databases respectively, in order to compare the efficiency differences between them. The other one explicitly declared the time measure as efficiency measure.

Table 4.4 shows that the remaining nine time measure categories comprised of only one controlled experiment (3.22%) each. The controlled experiment in "Comprehension time" measured the amount of time needed to comprehend and answer questions. The time measure of the controlled experiment that was referred to as "Coverage speed" measured how fast the subjects performed the experimental tasks. The intention was to measure the efficiency. The controlled experiment in "Error correction time" measured the average time it took the subjects to correct each error. Here, the measuring of time was also stated as a performance statistic. As the name "Meeting time" indicates, the controlled experiment in this time measure category measured the time subjects spent in meeting. The controlled experiment in "Work flow time" measured the time at which work sessions began and ended, and recorded the time at which activities occurred. The controlled experiment in "Requirements analysis/module design time" measured the duration of tasks related to requirements analysis and module design. Here, time could be regarded as connected to efficiency and effort measures. The controlled experiment categorised within the time measure category "Verification time" measured the time it took to carry out the entire verification process for all techniques that were examined. The controlled experiment in "Basic metric" called the total required time for the experimental tasks a basic metric for each subject's task. The controlled experiment in "Disturbing factor" stated and treated time measures as a disturbing factor. Disturbing factors are the same as confounding factors, which meant that it could affect the dependent variables without the knowledge of the experimenters.

Similar to the time measures in the experiment category "Time as dependent variable", the descriptions above of the time measures in the experiment category "Time as measure of outcome", show that in most cases the time measures were referred to according to the specific activities that were measured. For example "Project work time" which was a measure of project work. In addition, to determine the efficiency or effort of the experimental activities was in most controlled experiments used as the purpose of the time measures. This tells us that the same arguments as in section 4.2.1, in favour of a more uniform reporting of time measures, also could be used here.

However, probably a more important aspect to issue in relation to the time measures described in this section was the fact that the time measures were not explicitly described as dependent variables. Except for the time measure categorised as "Disturbing factor", the time measures were clearly used as dependent variables. That is because their purpose was to provide outcomes that could be used to determine the treatment in question. For example in

the case of the time measure category "Efficiency measure", one of the controlled experiments measured the duration of the subjects performing queries in third and fourth normal form in order to determine the most efficient one. Here, the time measures provided information for determining the efficiency of the normal forms, i.e. the treatments (independent variables).

A finding that could explain why many of the controlled experiments in this experiment category did not explicitly report time measures as dependent variables, was that a large majority, 24 out of the 31 controlled experiments, actually did not describe any dependent variables, nor any independent variables. Furthermore, two out of the remaining seven controlled experiments used only the term independent variable, without applying the equivalent term dependent variable. These referred to time measures as completion time, and in addition to the term measurement. The rest, i.e. five controlled experiments, described both the independent and dependent variables, but without explicitly referring to time measures in that regard. The time measures of these controlled experiments were categorised under "Inspection/review time", "Development time", "Efficiency measure", "Meeting time" and "Disturbing factor" respectively.

4.2.3 Time as time limit only

The third experiment category "Time as time limit only" comprised of nine controlled experiments that were found to only limit the time the subjects had to execute the experimental activities. This meant that time was not used as a measure of the experimental activities, and therefore, could not be regarded as having the role of a dependent variable. It must be stressed, though, that many of the controlled experiments in the other two experiment categories explicitly reported time limits as well. However, in those controlled experiments, the measuring of time must be viewed as the primary use of time. That is why time limits were not commented on in relation to those controlled experiments, and why they are not included in this section.

Table 4.5 – Time limits in controlled experiments

Topic category	Time limit
Software life-cycle/engineering (4 controlled experiments)	- 1 hour - 1 hour - 4 hours - 100 min
Project/product management (2 controlled experiments)	- 100–500 min - 1 hour
Tools (1 controlled experiment)	- 2 hours
Methodologies (2 controlled experiments)	- 245 min - 21 weeks

Table 4.5 summarises the time limits for each of the nine controlled experiments. Except for one controlled experiment that lasted 21 weeks, all the controlled experiments operated with time limits that ranged from one hour to approximately eight hours, with a median of two hours.

The reason why a controlled experiment reported a varying time limit of 100-500 minutes was that it used different time limits for the various experimental assignments. The controlled experiment that reported the time limit of 245 minutes, limited the total time for all activities

in the controlled experiment. The same applies to the controlled experiment with the time limit of 21 weeks.

If one considers the time limits in the light of the topic categories, Table 4.5 shows that the controlled experiments in "Methodologies" lasted the longest. However, there are so few instances in this experiment category, so we should be cautions in drawing conclusions.

4.3 Time recording

This section describes how the 85 controlled experiments that reported time recorded the time measures of the experimental activities.

Table 4.6 presents how time was recorded in the controlled experiments. The controlled experiments are grouped according to the three experiment categories that were presented in section 4.2, and the topic categories of Glass et al. [1] that were introduced in section 4.1. In addition, we have grouped the controlled experiments into time recording categories. These reflect the actual findings in relation to time recording. As the "Time recording category" shows, three different ways of recording time were found in the controlled experiments. These were by time recording tools, by subjects and by experimenters. The "Tools"-category comprises of electronic or computerized means that automatically recorded the time in the controlled experiments. The "Subjects"-category includes the students or professionals who participated and performed the experimental assignments in the controlled experiments. The "Experimenters"-category consists of the people that conducted the controlled experiments, i.e. most often the researchers. While the time recording of "Tools" could be characterised as being automatic, the time recording of "Subjects" and "Experimenters" could be characterised as being manual, for example by means of a manual clock. In addition to these three categories, the controlled experiments that were not found to describe their time recording were grouped in the time recording category "Not reported".

Table 4.6 shows that overall for the three experiment categories, tools (23 controlled experiments; 27%) and subjects (24 controlled experiments; 28.3%) recorded time most frequently. Experimenters were reported as the ones who recorded the time in only two controlled experiments (2.4%). In the majority of the controlled experiments (36 controlled experiments; 42.4%), however, it was not explicitly reported who/what recorded the time. This number is surprisingly high. In our view, controlled experiments should explicitly state how they have recorded time. This is important due to possible replications of the experiments, and to considerations regarding the validity of the experiments.

Table 4.6 – Categorisation of controlled experiments: Time recording

Experiment category	Time recording category	Topic category	No.	%
Time as dependent variable	**Tools**	Software life-cycle/engineering	8	9.4
		Methods/techniques	7	8.2
		Database/warehouse/mart organization	1	1.17
		Programming languages	1	1.17
			17	**19.9**
	Subjects	Software life-cycle/engineering	4	4.7
		Methods/techniques	6	7.1
		Project/product management	1	1.17
		Tools	1	1.17
		Measurement/metrics	1	1.17
			13	**15.3**
	Experimenters	Software life-cycle/engineering	1	1.17
			1	**1.17**
	Not reported	Software life-cycle/engineering	5	5.9
		Methods/techniques	9	10.6
			14	**16.5**
			45	**52.9**
Time as measure of outcome	**Tools**	Software life-cycle/engineering	1	1.17
		Methods/techniques	5	5.9
			6	**7.1**
	Subjects	Software life-cycle/engineering	5	5.9
		Methods/techniques	2	2.4
		Project/product management	4	4.7
			11	**13.0**
	Experimenters	Software life-cycle/engineering	1	1.17
			1	**1.17**
	Not reported	Software life-cycle/engineering	6	7.1
		Methods/techniques	5	5.9
		Database/warehouse/mart organization	1	1.17
		Personnel issues	1	1.17
			13	**15.3**
			31	**36.5**
Time as time limit only	***Not reported***	Software life-cycle/engineering	4	4.7
		Project/product management	2	2.4
		Tools	1	1.17
		Methodologies	2	2.4
			9	**10.6**
Total			**85**	**100.0**

In the following, the results regarding each of the three experiment categories will be described in section 4.3.1-4.3.3. In section 4.3.4, the time recording tools that were found will be presented in more detail. And, in section 4.3.5, possible trends in the time recording will be discussed.

4.3.1 Time as dependent variable

In the experiment category "Time as dependent variable", Table 4.6 shows that "Tools" appeared most often (17 controlled experiments; 19.9%) as the means, by which time measures were recorded. Also the subjects were responsible for recording the time in a substantial amount of the controlled experiments (13 controlled experiments; 15.3%). The time recording category "Experimenters" was only represented by one controlled experiment (1.17%). Perhaps the most interesting finding in this experiment category is that as many as 14 controlled experiments (16.5%) were not found to report how time measures were recorded.

When it comes to the topic categories, the most prominent ones, "Software life-cycle/engineering" and "Methods/techniques", were quite evenly represented amongst the time recording categories "Tools" and "Subjects". In relation to the time recording category "Not reported", it is worth noticing that "Methods/techniques" was represented by almost twice as many controlled experiments (nine controlled experiments; 10.6%), as "Software life-cycle/engineering" (five controlled experiments; 5.9%).

4.3.2 Time as measure of outcome

Table 4.6 shows that the largest time recording category of the controlled experiments in the experiment category "Time as measure of outcome" was "Not reported" (13 controlled experiments; 15.3%). The most prominent topic categories "Software life-cycle/engineering" and "Methods/techniques" were both highly represented amongst these controlled experiments, with six (7.1%) and five controlled experiments (5.9%) respectively.

Amongst the controlled experiments that reported how time measures were recorded, "Subjects" was distinctively the largest time recording category (11 controlled experiments; 13%). Compared to "Time as dependent variable" which comprised of 17 controlled experiments that reported on tools, "Time as measure of outcome" only included six controlled experiments (7.1%) that reported on tools. Regarding the topic categories, the two most dominating, "Software life-cycle/engineering" and "Methods/techniques", were not as evenly divided between "Tools" and "Subjects" as in "Time as dependent variable". Five controlled experiments (5.9%) within "Software life-cycle/engineering" reported on subjects, while one controlled experiment (1.17%) within the same topic category used a tool to record the time measures. For "Methods/techniques", it was almost the opposite. Five controlled experiments (5.9%) described tools as the method of recording the time measures, while two controlled experiments (2.4%) expressed that it was the subjects.

4.3.3 Time as time limit only

As we can see in Table 4.6, none of the controlled experiments in the experiment category "Time reported as time limit only" reported any time recording means. This is not surprising, due to the fact that these controlled experiments only applied time in the sense of time limits, and thus, did not measure the duration of the experimental activities. Therefore, there was no need for recording time measures in these controlled experiments.

4.3.4 Descriptions of time recording tools

Table 4.7 specifies the various time recording tools that were used in the controlled experiments. Overall, 11 types of time recording tools were found amongst the 23 controlled experiments that reported on time recording tools. The table indicates that there is a great variety of tools in both of the experiment categories "Time as dependent variable" and "Time as measure of outcome". As the table shows, some time recording tools were used in controlled experiments in both of the experiment categories and in several topic categories. In the following, each type of time recording tool will be described.

Table 4.7 – Categorisation of controlled experiments: Tools

Experiment category	Topic category	Tools	No.	%
Time as dependent variable	**Software life-cycle/engineering**	- Computer log files	3	13.0
		- Facilities in CSRS (the Collaborative Software Review System)	3	13.0
		- Facilities in the UNIX-system	1	4.35
		- PROMPTS (Program Maintenance Performance Testing System)	1	4.35
	Methods/techniques	- Computer log files	2	8.7
		- Casio HS-3 digital chronometer	1	4.35
		- Shell scripts	2	8.7
		- SuperCard program	2	8.7
	Database/warehouse/mart organization	- VAX/VMS operating system	1	4.35
	Programming languages	- Facilities in the Motif programming environment	1	4.35
			17	**73.9**
Time as measure of outcome	**Software life-cycle/engineering**	- Computer log files	1	4.35
	Methods/techniques	- Computer log files	1	4.35
		- Facilities in the UNIX-system	1	4.35
		- Video-tape	2	8.7
		- Web-based tool	1	4.35
			6	**26.1**
Total			**23**	**100.0**

Overall, the most commonly used type of time recording tool was computer log files. Seven (30.4%) out of the 23 controlled experiments that reported on time recording tools, explicitly reported on computer log files. As the name indicates, the computer log files were typically files in the experiment monitoring environment, which logged all interaction with the computer program, and thereby also the clock times of the activities of the subjects.

The second largest type of time recording tool was CSRS (the Collaborative Software Review System). This tool was found in three controlled experiments (13%). This system used a timestamp-based mechanism to automatically record the time spent by each reviewer that had logged into the system.

Four different types of time recording tools were found in two controlled experiments (8.7%) respectively. These were “Facilities in the Unix-system”, “Shell scripts”, “SuperCard program” and Video-tape”.

The first of the two controlled experiments in “Facilities in the Unix-system” used the Unix-system to collect each programmer’s time log. The second controlled experiment reported that the subjects worked on the experimental tasks while using a specific Unix-account, which provided an automatic monitoring infrastructure, and which non-intrusively registered login/logout times and all compiled source versions with timestamps.

In the time recording tool “Shell scripts”, the two controlled experiments reported that each subject was required to start a custom designed shell script, which provided a workstation prompt with their login name and the time. This allowed the researcher to recognize how long the subject spent on a particular problem. Another shell script automatically copied each file with a timestamp to a backup directory, while compiling the subject's files to generate the executable.

In the time recording tool “SuperCard program”, the two controlled experiments reported on a program that was written in SuperCard 1.5. This program enabled the recording of time measures of the experimental activities by timing the subjects from when they clicked the button “Next” until they clicked the button “Done”.

In the two controlled experiments that reported on the time recording tool “Video-tape”, the experimenters videotaped the experimental activity-sessions. From these videotapes they could extract the time the subjects used on the activities.

The controlled experiment (4.35%) that reported on the time recording tool “PROMPTS (Program Maintenance Performance Testing System)” stated that PROMPTS was an especially designed system for this study. For each program (written in C) shown to the subject, PROMPTS recorded the total time spent reading the program, reading the task and successfully maintaining the program.

A second time recording tool that was only reported by one controlled experiment (4.35%) was the “Casio HS-3 digital chronometer”. Here, each subject made use of a Casio HS-3 digital chronometer to measure the time elapsed to answer each question. This tool differed from the other time recording tools due to the fact that it was a manual clock. That meant that it was not a part of a computer environment, and the subjects had to manually use it. Thus, one may maintain that it was actually the subjects that recorded the time. However, we have included it in the “Tools”-category, because it was explicitly referred to in the controlled experiment.

The controlled experiment (4.35%) that reported on the tool “VAX/VMS operating system” used this tool to automatically record the total elapsed time for each query construction task.

The controlled experiment that used “Facilities in the Motif programming environment” to record time measures, reported that the Motif programming environment captured all program versions submitted for compilation along with a timestamp and the messages produced by the compiler and linker. These data, in addition to a timestamp for the start and end of the work phase for each problem, was written to a protocol file in the programming environment.

The last time recording tool that was listed in Table 4.7 was the "Web-based tool". The controlled experiment (4.35%) that reported on this tool only stated that the subjects recorded the time they spent on the experimental project in a Web-based tool. The controlled experiment did not describe the tool in more depth.

4.3.5 Trends

In this section, we present the time recording categories in relation to the publication year of the controlled experiments (i.e. the articles), in order to try to identify some trends in the time recording. It was especially interesting to find out whether the use of time recording tools is a recent and growing practice within controlled SE-experiments. This may be assumed due to the increase in the use of computerized tools in many areas over the last years. Hence, one might believe that this has also happened in the field of controlled SE-experiments. Because the investigated controlled experiments were published in the period 1993-2002, the study obviously does not provide information about time recording in the most recent years. The experiment category "Time as time limit only" has been excluded from this investigation, since time recording was not an issue in that experiment category. Therefore, only 76 controlled experiments have been analysed in this context.

Table 4.8 shows the results. In the experiment category "Time as dependent variable", the table shows that there are no marked peaks in any of the time recording categories "Subjects", "Experimenters", and "Not reported". The controlled experiments in these three time recording categories were distributed over all the current years except for 1993, with fairly even numbers. However, when it comes to "Tools", the results are quite interesting. As many as ten out of the 17 controlled experiments in the time recording category "Time as dependent variable", were published in the years 1996 and 1997. As we see it, this must be regarded as a significant peak. In comparison, only two controlled experiments that used tools to record time were published in 2001 and 2002. Not only do these findings show that time recording tools were most frequently used in 1996 and 1997, but also that the use of time recording tools actually have decreased in the latest years. The largest proportion of the most recent controlled experiments reported that the subjects were the ones who recorded the time.

Table 4.8 – Categorisation of controlled experiments: Trends

Experiment category	Time recording category	Publication year of article (1993-2002)	No.	%
Time as dependent variable	**Tools**	1993	1	1.32
		1996	4	5.3
		1997	6	7.9
		1998	1	1.32
		2000	3	3.9
		2001	1	1.32
		2002	1	1.32
			17	**22.4**
	Subjects	1995	3	3.9
		1997	2	2.6
		1998	2	2.6
		2001	4	5.3
		2002	2	2.6
			13	**17.1**
	Experimenters	1997	1	1.32
			1	**1.32**
	Not reported	1994	1	1.32
		1996	1	1.32
		1997	1	1.32
		1998	2	2.6
		1999	3	3.9
		2000	1	1.32
		2001	2	2.6
		2002	3	3.9
			14	**18.4**
			45	**59.2**
Time as measure of outcome	**Tools**	1993	1	1.32
		1996	2	2.6
		1999	2	2.6
		2000	1	1.32
			6	**7.9**
	Subjects	1996	1	1.32
		1997	1	1.32
		1998	3	3.9
		2000	1	1.32
		2001	5	6.6
			11	**14.5**
	Experimenters	1998	1	1.32
			1	**1.32**
	Not reported	1994	1	1.32
		1995	1	1.32
		1996	1	1.32
		1997	1	1.32
		1999	2	2.6
		2000	3	3.9
		2001	1	1.32
		2002	3	3.9
			13	**17.1**
			31	**40.8**
Total			**76**	**100.0**

The findings in the experiment category "Time as measure of outcome" indicate the same trend as identified above. A majority of the newest controlled experiments (published in 2001 and 2002) (five controlled experiments; 6.6%) reported the subjects as the ones who recorded the time. In fact, none of the controlled experiments published in 2001 and 2002, reported on time recording tools.

Based on the results above, we cannot identify any trend that indicates a recent and growing practice in using time recording tools in controlled SE-experiments. The results rather point to the opposite. The reasons for this surprising trend are unclear. It may be that the authors of articles describing controlled experiments have become less accurate in reporting time recording tools. Or perhaps it was too expensive or too complicated to use tools for recording time measures in controlled experiments. As we see it, a discussion of this possible trend is beyond the scope of this thesis, but should be investigated more thoroughly in future work.

Another finding that is worth noticing from the results in Table 4.8, is that as many as nine out of 22 controlled experiments published in 2001 and 2002 in both of the experiment categories, were found not to explicitly report how they recorded their time measures. In comparison, only four out of 21 controlled experiments published in 1996 and 1997, did not report any time recording means. We suspect that the experimenters themselves recorded the time in most of the nine controlled experiments in the time category "Not reported" and published in 2001 and 2002. Only two (2.6%) out of 76 controlled experiments in Table 4.8, reported on the experimenters as the ones who recorded the time measures. We believe that the experimenters record the time way more frequently than our findings indicate. Therefore, this may represent another trend in relation to time recording in controlled SE-experiments. Obviously, this is not a positive trend. The authors of articles describing controlled experiments should explicitly state how time measures are recorded, due to possible replications and to validations of the results of the controlled experiments.

4.4 Time units

This section describes how the 85 controlled experiments that reported time specified the time in terms of time units. Table 4.9 summarises the results from this investigation. As the outer left column in Table 4.9 indicates, we still treat the controlled experiments according to the categories we described in section 4.2. Furthermore, the controlled experiments are categorised into time unit groups, time units and topics. The column "Time unit group" represent the categories "One time unit reported", "Two time units reported" and "Time unit not reported", and specifies whether the controlled experiment reported only one kind of time unit, two kinds of time units, or none time units at all. The column "Time unit" represent the time units that were used. The "Topic category"-column represents the SE-topics classified by Glass et al. [1], which are used in the previous sections.

Table 4.9 – Categorisation of controlled experiments: Time units

Experiment category	Time unit group	Time unit	Topic category	No.	%
Time as dependent variable	***One time unit reported***	**Seconds**	Software life-cycle/engineering	1	1.17
			Methods/techniques	7	8.2
				8	**9.4**
		Minutes	Software life-cycle/engineering	10	11.8
			Methods/techniques	13	15.3
			Project/product management	1	1.17
			Tools	1	1.17
			Database/warehouse/mart organization	1	1.17
			Measurement/metrics	1	1.17
				27	**31.8**
		Hours	Software life-cycle/engineering	3	3.5
			Methods/techniques	1	1.17
			Programming languages	1	1.17
				5	**5.9**
	Two time units reported	**Minutes and hours**	Software life-cycle/engineering	3	3.5
				3	**3.5**
		Hours and days	Software life-cycle/engineering	1	1.17
				1	**1.17**
	Time unit not reported		Methods/techniques	1	1.17
				1	**1.17**
				45	**52.9**
Time as measure of outcome	***One time unit reported***	**Seconds**	Software life-cycle/engineering	1	1.17
			Methods/techniques	1	1.17
			Database/warehouse/mart organization	1	1.17
				3	**3.5**
		Minutes	Software life-cycle/engineering	5	5.9
			Methods/techniques	7	8.2
			Personnel issues	1	1.17
				13	**15.3**
		Hours	Software life-cycle/engineering	6	7.1
			Methods/techniques	3	3.5
			Project/product management	4	4.7
				13	**15.3**
	Two time units reported	**Minutes and hours**	Software life-cycle/engineering	1	1.17
			Methods/techniques	1	1.17
				2	**2.4**
				31	**36.5**
Time as time limit only	***One time unit reported***	**Minutes**	Software life-cycle/engineering	1	1.17
			Methodologies	1	1.17
			Project/product management	1	1.17
				3	**3.5**
		Hours	Software life-cycle/engineering	3	3.5
			Project/product management	1	1.17
			Tools	1	1.17
				5	**5.9**
		Weeks	Methodologies	1	1.17
				1	**1.17**
				9	**10.6**
Total				**85**	**100.0**

Table 4.9 shows that the following five time units were identified: seconds, minutes, hours, days and weeks. The table reveals that all in all most controlled experiments in the three experiment categories were found to specify their time with only one type of time unit. However, there were six controlled experiments (7.1%) that used two time units ("Two time units reported"), i.e. "Minutes and hours" and "Hours and days". This was found in four controlled experiments (4.7%) in the experiment category "Time as dependent variable", and in two controlled experiments (2.4%) in the experiment category "Time as measure of outcome".

In the following sections, 0-4.4.3, the results in each of the experiment categories are presented in more depth. In section 4.4.4, the results are discussed.

4.4.1 Time as dependent variable

In the experiment category "Time as dependent variable" in Table 4.9, 44 out of 45 controlled experiments specified the time units. In one controlled experiment (1.17%), the time unit used was not found. The time unit was neither declared in connection with the description of the time measure, nor in the table where the results of the time measures were listed.

Table 4.9 further shows that the most prominent time unit category was minutes, found in 27 controlled experiments (31.8%). Amongst these controlled experiments, the two dominating topic categories were "Software life-cycle/engineering" (10 controlled experiments; 11.8%) and "Methods/techniques" (13 controlled experiments; 15.3%).

The duration of all the controlled experiments in "Time as dependent variable" dealing with the topic "Software life-cycle/engineering" ranged from approximately one hour to 390 hours, with a median of six hours. For the topic category "Methods/techniques" which also comprised of a high proportion of the controlled experiments in this experiment category, the duration ranged from about 30 minutes to 24 hours, with a median of four hours.

4.4.2 Time as measure of outcome

In the experiment category "Time as measure of outcome", all the 31 controlled experiments specified their time measures in time units. In this experiment category, minutes and hours were most frequently used, i.e. in 13 controlled experiments (15.3%) for both time units.

Regarding the topic categories in this experiment category, Table 4.9 shows that the majority of the controlled experiments within the two dominating topics "Software life-cycle/engineering" and "Methods/techniques", measured time in minutes and hours.

An interesting finding in this experiment category was that all four controlled experiments in the topic category "Project/product management" measured their experimental activities with the same time unit, i.e. hours. Although the numbers are quite small, this may indicate a trend amongst controlled experiments dealing with "Project/product management".

The duration of the controlled experiments within the topic category "Software life-cycle/engineering", ranged from about two hours to 88 hours, with a median of eight hours.

For the other prominent topic, "Methods/techniques", the duration of the controlled experiments ranged from 30 minutes to 55 hours, the median being three hours.

4.4.3 Time as time limit only

As Table 4.9 shows and as section 4.2.3 gave indications of, all the controlled experiments in the experiment category "Time as time limit only" explicitly stated the time units they used. The time units minutes, hours and weeks were found. A majority of the controlled experiments (five controlled experiments; 5.9%), reported the time limits in hours. Three controlled experiments (3.5%) reported minutes, while one controlled experiment (1.17%) reported weeks as the time unit.

When it comes to the topic categories, perhaps the most interesting finding was that a majority of the controlled experiments that addressed the topic "Software life-cycle/engineering" (three controlled experiments; 3.5%) stated their time limits in hours.

The duration of the controlled experiments in this experiment category was reported in section 4.2.3.

4.4.4 Discussion

The results above show that the 85 controlled experiments that reported time were not homogeneous in terms of which time units they used. The time units seconds, minutes, hours, days and weeks were all applied in the controlled experiments. However, if we were to imply an overall trend in the use of time units, the results show that time is most often measured in minutes. 48 (56.5%) out of the 85 investigated controlled experiments specified their time measures in this time unit. In comparison, the second largest time unit, hours, were found in 29 controlled experiments (34.1%). There may be several reasons for the diverse use of time units. One cause that probably is valid for the majority of the controlled experiments is that the time units were chosen to suit the length and nature of the experimental activities.

One problem that can be identified from the results described above is that the controlled experiments that stated their time measures in hours could provide inaccurate results. This may be the case if the conductors of the controlled experiments round the time measures up or down to the nearest hour. Therefore, we would recommend that experimenters as far as possible state their time measures in minutes instead of hours. Nevertheless, the large proportion of controlled experiments using minutes should indicate that this is not a general problem.

An interesting finding from the investigation of the time units was that one controlled experiment in "Time as dependent variable" did not explicitly specify the kind of time unit it used. Obviously, this lack of information is grave due to its importance to the understanding of the time measures. And since the time measures in this controlled experiment was stated as dependent variable, the missing time unit reduces the clarity of the experimental results as well.

4.5 Time as validity threat

This section presents the last aspect of time that we analysed. We show how time was addressed as threat to the validity of the controlled experiments. It is of high significance to identify factors that can represent validity threats, since they may limit the validation of the research results [9]. In this context, we have chosen to focus on the following two types of validity threats; threats to internal validity and threats to external validity. Validity threats can also be separated into threats to conclusion validity and threats to construct validity.

When one talks about the internal validity of a controlled experiment, one addresses internal aspects of the controlled experiment, or more specific, the causal relationship between the treatment and the outcome, as described in section 1.1. If a factor is regarded as a threat to the internal validity, it means that the factor influences the relationship between the treatment and the outcome in such a way that it may seem like the treatment causes the outcome, although it really does not. [9]

When one talks about the external validity of a controlled experiment, one considers generalization aspects of the controlled experiment, i.e. how the results of the experiment can be generalised outside the scope of the controlled experiment's study. If a factor is viewed as a threat to the external validity, it means that the factor impacts the controlled experiment's ability to generalize its results. [9]

In section 4.5.1, we briefly present the frequency of addressing time as validity threat amongst the controlled experiments. In section 4.5.2, we describe the findings in more depth. In section 4.5.3, we discuss the findings.

4.5.1 Frequency of addressing time as validity threat

Table 4.10 – Categorisation of controlled experiments: Time as validity threat

Experiment category	Topic category	No.	%
Time as dependent variable	Software life-cycle/engineering	2	22.2
	Methods/techniques	2	22.2
	Tools	1	11.1
	Measurement/metrics	1	11.1
		6	**66.7**
Time as measure of outcome	Software life-cycle/engineering	1	11.1
		1	**11.1**
Time as time limit only	Software life-cycle/engineering	1	11.1
	Tools	1	11.1
		2	**22.2**
Total		**9**	**100.0**

The first focus of the investigation of time as validity threat was to determine whether the controlled experiments actually discussed time as threat to the validity. Table 4.10 summarises the results of this study.

As the table shows, only nine out of the 85 controlled experiments that reported time were identified as discussing time as a possible validity threat. A majority of these nine controlled experiments (six controlled experiments; 66.7%) were found amongst the controlled experiments that reported time as dependent variable. The other two experiment categories comprised of only one (11.1%) and two controlled experiments (22.2%) respectively. "Software life-cycle/engineering" was the most common topic amongst the nine controlled experiments (four controlled experiments; 44.4%).

4.5.2 Descriptions of time as validity threat

Although the nine controlled experiments above could be considered to discuss time as validity threat, there were significant variations in their perspectives regarding time as validity threat. Table 4.11 presents a details the findings in the nine controlled experiments.

Table 4.11 – Categorisation of controlled experiments: Time as validity threat - Quotes

Experiment category	Type of validity threat	Art. no	Quote from article	Reported in chapter
Time as dependent variable	***Internal validity***	**1**	(Quote 1) "Missing time data occurred as a result of a subset of subjects not following experimental instructions. The concern here is that this data loss was not random and, thus, had an influence on the results." (Quote 2) "Increase the task time. The time allocated for answering questions and performing impact analysis should be increased because many subjects stated the reason they did not complete all the tasks was due to time constraints." (Quote 3) "Improve time data collection procedures. The data collection procedures for collecting precise time data should be improved upon due to our experience with subjects not fully obeying experimental instructions."	4.2 Internal Validity 4.4 Addressing Threats to Validity 4.4 Addressing Threats to Validity
		2	"Precision in the time values. The subjects were responsible for recording the start and finish times of each test. We think that this method is more effective than having a supervisor who records the time of each subject. However, we are aware that the subject could introduce some imprecision."	3.5.2. Internal validity
		3	"Time limit was too short. Analyzing human performance in programming activities, Weinberg and Schulman concluded that unreasonably short deadlines would result in erroneous programs, and warned experimenters against mixing the results of subjects who have easily finished with those of subjects who were pressed for time (Weinberg, 1974). We do not know if this happened to the original experiment. In our case, if we discard all the data from subjects pressed for the deadline there will not be enough observations to analyze. "	4. Discussion

		4	"Hence to minimise this threat to validity, the original experiments strict time limit was abandoned and the subjects were given as much time as they required."	2.1.3. Threats to Internal Validity
	External validity	**5**	"The primary threat to the external validity of this experiment is that subjects' response times were measured in a laboratory environment rather than in a typical work environment."	5.3. Threats to Validity
		6	"Termination criteria were partially specified by the quantitative quality models in terms of duration, but the subject was expected to decide when to stop working. The use of a hard time limit was seen as a threat to external validity and therefore avoided."	5.2 Conducting the experiment
Time as measure of outcome	***Internal validity***	**7**	"The hours of effort recorded by the subjects themselves are potentially inaccurate; however, closely monitoring the time spent by each subject would violate the intent to carry on the experiment under working conditions."	2. Design
Time as time limit only	***Internal validity***	**8**	"Third, sufficient time needs to be allotted for the subjects to perform the system designs. Some of the subjects in this study felt rushed in completing their designs in the allotted hour... Time-pressure is likely to distort subjects' thinking, causing them to rush ahead to show some semblance of a completed design, but skipping steps in the methodology guide or taking the steps in improper order."	5. Discussion
	External validity	**9**	"The programs used may not be representative of the length and complexity of those found in an industrial setting. The programs used were chosen for their length, allowing them to be inspected within the time available. However, the amount of time given to inspect each program was representative of industrial practice quoted in popular inspection literature."	2.6. Threats to Validity/ Threats to External Validity

In Table 4.11, we have organised the articles according to which experiment category the controlled experiments that they described, belonged to, and which type of validity threat we considered time to be addressed as. The table contains quotes from the nine articles, which, in our opinion, represent the sections where time is viewed as validity threat. The chapters in which the quotes were found are also stated in the table, in order to indicate the context of the quote. In addition, an article number ("Art.no") is added to every article that is quoted in the table, so that the articles can be identified in the descriptions beneath.

In the following, the findings in each of the three experiment categories are described.

4.5.2.1 Time as dependent variable

Table 4.11 shows that we identified the articles in the experiment category "Time as dependent variable" as addressing time both in relation to internal and external validity. We interpreted that four articles discussed time as threat to internal validity, and that two articles discussed time as threat to external validity.

Our opinion is that the first and the third quote of the first article issue the same aspect. As we see it, they both imply that defective recording of time data of the subjects could represent a threat to the internal validity of the controlled experiment. They both stress that it is important that the subjects obey the experimental instructions. The reason why we view the statements

in these quotes as threats to internal validity is that the first quote was found in a chapter explicitly referred to as internal validity. In addition, the quotes concern an internal aspect of the controlled experiment.

When it comes to the second quote of the first article, we regard this as addressing a strict time limit in the execution of the experimental assignments as a threat to the internal validity. The quote states that there should be allocated more time. We regarded the statement as a threat to the internal validity, since it deals with an internal aspect of the controlled experiment. Common for all the quotes in the first article, is that they were found in explicitly stated validity-chapters.

Our interpretation of the quote of the second article, which was expressed in a validity chapter dealing with internal validity, is that the chosen method for recording time data, i.e. the subjects recorded the time, could represent a threat to the internal validity of the controlled experiment. This was due to inaccuracy of the subjects' recordings. Nevertheless, this method of time recording was regarded as more effective than having a supervisor recording the time, and was therefore preferred.

In the third article, the quote was found in a discussion-chapter, and not in an explicitly stated validity-chapter. However, we understood the quote as implying that time could be regarded as a threat to the internal validity. More precise, our interpretation is that the quote says that strict time limits could affect the outcomes of the controlled experiments.

We also viewed the quote of the fourth article as addressing time as a threat to the internal validity. Here, the article quite explicitly states that a strict time limit could represent a threat to the internal validity of the controlled experiment.

As Table 4.11 shows, we have categorised the quotes in the fifth and the sixth article as dealing with time as threats to external validity. Both of them explicitly say that they assess a threat to the external validity. On the other hand, while the sixth article views a hard time limit as a threat to the external validity, the fifth article does not directly state time as a threat to the external validity. However, we have included the quote of the fifth article, because the measuring of time is a part of the validity threat. As we see it, the statement recommends that subjects' response times should be measured in a work setting rather than in a laboratory setting, to reduce the threat to the external validity. As Table 4.11 shows, the quote in the sixth article was found in the chapter "5.2 Conducting the experiment", which cannot be regarded as a typical validity-chapter.

4.5.2.2 Time as measure of outcome

In the experiment category "Time as measure of outcome", we only found one article that we regarded as addressing time as validity threat, i.e. the seventh article. However, the quote of the seventh article does not explicitly describe time as a validity threat. As we read it, though, it says that time recording performed by subjects could constitute a threat to the validity, due to the potential inaccuracy of the time recordings. But, this time recording method was chosen anyway, because time recording executed by experimenters (i.e. the conductors of the experiment) would have constituted an even larger threat to the experiment under the working conditions. We characterised this quote as describing a threat to the internal validity, since it

concerns internal aspects of the controlled experiment. As Table 4.11 shows, the quote in the seventh article was found in a non-typical validity-chapter, i.e. “2. Design”.

4.5.2.3 Time as time limit only

As Table 4.11 shows, we identified two articles in the experiment category “Time as time limit only” that discussed time as threat to the validity of the controlled experiments. Our interpretation of the quote of the eight article, is that it says that having strict time limits in the experimental activities could cause defective results. Furthermore, it was stated in the article (not in the quote, but in the context of it) that the matter addressed in the quote could be seen as a limitation from which the study suffered. In our opinion, this surely indicates that the quote issues time as a threat to the internal validity. It is worth noticing that the quote was found in non-typical validity-chapter, i.e. “5. Discussion”.

As we interpret the quote of the ninth article, it explains why the allocated time did not represent a threat to the validity of the controlled experiment. We have characterised the quote of this article as addressing external validity. This is partly because it was found in the chapter “2.6. Threats to Validity/Threats to External Validity”. In addition, the quote states it implicitly, by saying that the time the subjects were given was appropriate, since it followed the general practice in allocating time for inspections.

4.5.3 Discussion

As the overall findings in section 4.5.1 shows, only nine out of 85 controlled experiments discussed time as validity threat. In light of the extensive use of time in controlled SE-experiments overall (see section 4.1) and time’s role as dependent variable (see section 4.2), we find it strange, and also worrying, that time does not appear more frequently in connection with validity-issues. Surely, the validity threats of time described in section 4.5.2 are not unique for only some few controlled experiments. Time limits and time recording represent aspects of time which, as we see it, are valid for most controlled experiments that measure time. Our opinion is that the low frequency of articles addressing time as validity threat must indicate that the SE-researchers documenting the controlled experiments (the authors of the articles) are not enough aware of the potential validity threats that aspects of time represent. Due to time’s role as dependent variable, and thereby, its effect on the conclusions regarding the investigation of the independent variables, we recommend that the SE-researchers documenting the controlled experiments pay more attention to the validity threats that time can constitute. Clearly, this would be important for the validation of the research results that the controlled experiments test.

Our findings in the section above regarding the controlled experiments that reported time as validity threat, show that time recording and time limits are the aspects of time that most frequently are considered as validity threats of controlled experiments. Time limit is discussed as validity threat in six out of 11 quotes in Table 4.11, i.e. in six articles, and time recording is discussed as validity threat in four out of 11 quotes, i.e. in three articles. The last quote views the environment of the measuring of time as a validity threat.

Most of the articles that addressed time limit as validity threat stated a too strict time limit as a threat. They argued that subjects should be given enough time to perform the experimental

tasks, so that the results would not be affected by insufficient time. It is interesting to notice that the strict time limits were treated both as threat to internal validity and as threat to external validity.

When it comes to the articles that issued time recording as validity threat, they were more diverse. However, a similarity is that they all implied that the inaccuracy of the time recording performed by subjects could affect the results of the controlled experiments, and thereby, constitute a threat to the validity. Another, more general similarity is that they all considered time recording as threat to the internal validity.

An interesting finding from the section above is that far from all the quotes were found in explicitly stated validity chapters. Seven out of 11 quotes were found in chapters with titles containing the word validity, while four out of 11 quotes were found in chapters with titles not containing the word validity. Three of the four quotes that did not contain the word validity in the title of the chapter, did, however, not explicitly say that it addressed a threat to the validity. Nevertheless, we interpreted the quotes as clearly dealing with validity-issues. The last of the four quotes (the quote in the sixth article), on the other hand, explicitly stated that the strict time limit was a threat to the validity. This should indicate that validity issues are not always presented in dedicated validity-chapters. In our view, this practice may be negative for the SE-community, since it is important that researchers who want to learn from a controlled experiment, have a complete understanding of the threats to the validity of the experiment. If threats are stated elsewhere in the articles, it may not be noticed by the readers. Therefore, we recommend that all issues concerning the validity of the controlled experiments are presented in chapters that are titled with the word validity.

5 Validity

This section discusses potential threats to the validity of the research of this master thesis.

5.1 Selection of articles

The systematic survey in this master thesis studied 113 controlled SE-experiments from 103 articles. As stated in section 3.2, the identification of these articles was done by others. Sjøberg et al. [6] addressed some potential threats to the validity of the process of selecting articles. Naturally, the potential validity threats that are issued in this section were applicable to the research of this master thesis.

First of all, a potential validity threat is the selected journals and conferences, where the articles were found [6]. Although the 12 journals and conferences were regarded as leaders in SE in general and empirical SE in particular, the selection process should also have included grey literature, such as theses, technical reports, working papers etc. The parts of this literature that described controlled SE-experiments would have provided additional data and allowed the researchers to conclude more generally.

Secondly, Sjøberg et al. [6] stated that the process of selecting articles may have been suffering from bias. This potential validity threat was explained by the large and heterogeneous collection of articles that were surveyed. In addition, there were no keyword standard to differentiate between methods in empirical SE, and that could be applied to extract controlled experiments consistently. These aspects made the selection process difficult, and thus, the researchers may not have managed to find all relevant articles.

5.2 Systematic survey

This section presents possible validity threats to the research of this master thesis specifically.

5.2.1 Potential bias in the data extraction process

The fact that one person only conducted the systematic survey, constituted an important potential validity threat. When one considers the amount of articles in the light of the duration of the analysis (20 working days) in this systematic survey, it is obvious that it was not possible to read the entire content of the articles. This increased the possibility of inaccurate or incorrect data extraction, for example in terms of missing relevant data, gathering defective data, or misunderstanding data. Clearly, the weakness of conducting a survey of this scale alone is that it could be biased, and therefore, probably would not provide an entirely correct picture of the research domain. Ideally, the systematic survey should have been conducted by at least two persons. The two persons could have performed the survey individually, and then they could have compared the results. In cases where there were inconsistencies in the extracted data, they could have reinvestigated them. In that way, the reliability of the extracted data would most likely have increased.

In regards to the specific results from the systematic survey, the overall frequency of reporting time described in section 4.1 was probably not affected as much as the other aspects by a possible bias. Logically, this is because it was the most general aspect and, thus perhaps, the easiest to examine. Amongst the other aspects that were dealt with in the sections 4.2-4.5, it is difficult to determine how much they might have been influenced by a possible bias. However, it is a fact that some findings in relation to these aspects were based on quite subjective considerations. Especially, findings in section 4.5, which described time as validity threat, could be characterised in this way. A reason for this might be that data relevant for this aspect had to be resolved from a context, i.e. time measures assessed in the light of validity considerations. Furthermore, these data were not always found in the same types of sections of the articles. In comparison, data concerning the aspects time units and time recording could be regarded as more independent of the context, and more precise and explicit. And data connected to the use of time described in section 4.2, were to a large extent found in the same sections of the articles, i.e. the experimental design etc.

5.2.2 Means of reducing the potential bias

Here, some means for reducing the potential validity threat that bias constituted to the data extraction process in the systematic survey are mentioned.

First of all, as stated in section 3.3, before performing the systematic survey, the aspects that were going to be investigated were more thoroughly defined. In this way, we had more or less a clear opinion of what kind of data we were going to search for. Most likely, this increased the possibility of extracting relevant data, and reduced the probability of researcher bias.

Secondly, also mentioned in section 3.3, the systematic survey was initiated by reading the entire content of five randomly chosen articles from the 103 articles and trying to gather relevant data from these. The purpose was to get an insight into the structure of the articles, and thereby, recognize the most relevant sections of the articles. However, a problem was that far from all articles had the same structure.

Thirdly, in the systematic survey we used the search function of Adobe Acrobat to search for important keywords in the articles. This helped us to uncover relevant information.

Last, but not least, the process of identifying data was simplified by the support of earlier research on parts of some of the aspects. As mentioned in section 1.1 and 3.3, all dependent variables in the 113 controlled experiments had been surveyed and summed up by a researcher at SIMULA RL. By crosschecking these findings with the corresponding findings in the systematic survey of this master thesis, the potential correctness of these data obviously increased. Another part of the aspects that had been investigated earlier was the topic categorisation of the articles. As stated in section 4.1, these topics were applied in Sjøberg et al. [6].

6 Conclusion

This section summarises and concludes the research of the master thesis, and gives suggestions for future work. Section 6.1 repeats the objective of the research. Section 6.2 sums up the results. Section 6.3 draws the conclusions. Section 6.4 makes proposals for future work.

The results are summarised quite extensively here, since we have not summarised them particularly earlier.

6.1 Objective of research

The objective of the research of the master thesis was to quantitatively analyse how controlled experiments in the field of software engineering reported time. Earlier study had shown that time often occurred as dependent variable in controlled SE-experiments. Due to the fact that the dependent variables provide the outcomes of the controlled experiments, and thereby, contributes to the conclusions regarding the research questions (i.e. the hypothesis of the causal relationship), it was important to identify how time was used. Hopefully, this would increase the knowledge of the existing use of time in controlled experiments, and contribute to the development of a state-of-the-art practice in running experiments. An advancement of a state-of-the-art practice within empirical software engineering research is vital for the private and public IT-industries, in order for them to develop better IT-systems more effectively, i.e. with fewer resources.

The master thesis was a part of the project "Research Methods and Support Tools for Conducting Empirical Research in Software Engineering" at SIMULA RL [5]. It was a short thesis that lasted from the 22nd of August 2005 to the 19th of December 2005, and counted for 30 points in the master degree.

6.2 Results

The investigation of time in the controlled experiments was conducted by a systematic survey. 113 controlled experiments described in 103 scientific articles published in the period 1993-2002 were surveyed. The survey of this master thesis can be viewed as an extension of Sjøberg et al. [6]. The survey focused on identifying the following aspects of time in the controlled experiments: 1) the overall frequency of reporting time, 2) the use of time, 3) the recording of time, 4) the specification of time in terms of time units, and 5) time as validity threat. In our opinion, these aspects constituted the most important features of time in controlled experiments, and would provide a nearby complete picture of the handling of time.

It must be remarked that the analysis only explored time in the sense of being a direct measure. This meant that time was not considered as a part of indirect measures, i.e. in cases where it was reported in calculations with other measures.

6.2.1 Overall frequency of reporting time

The reason for investigating the overall frequency of reporting time was to identify the controlled experiments that explicitly reported time, and present the overall relevance of the research focus of the master thesis.

The systematic survey showed that a large majority of the controlled experiments reported time, i.e. 85 controlled experiments (75.2%) out of the 113 controlled experiments in all. Only 28 controlled experiments (24.8%) did not mention time at all. The large proportion of controlled experiments that reported time indicated that time is frequently used in SE-experimentation. The 85 controlled experiments that reported time were further investigated in relation to the other four aspects.

6.2.2 Use of time

The use of time in the controlled experiments was studied in order to understand how time was actually referred to and used. The results show that time was referred to and used in different ways in the 85 controlled experiments. We divided the use of time into three categories (experiment categories); time as dependent variable, time as measure of outcome, and time as time limit only.

Time as dependent variable was the largest category comprising of 45 controlled experiments (52.9%) out of the 85 controlled experiments. These controlled experiments had in common that time was explicitly reported as dependent variable in the experimental design, and thereby, time was used as a measure of the subjects' solving of the experimental tasks, i.e. activities concerning the independent variables. Although these controlled experiments were grouped in the same experiment category, there were differences in how they actually named the time measures and used time as measures. We grouped the time measures into 13 time measure categories, which reflected the terms the controlled experiments used. In most of the controlled experiments, the names referred to the specific activities that time measured. Even though the time measures had various names, most of the controlled experiments stated explicitly that the overall purpose of the time measures was to resolve the efficiency or effort of the experimental activities.

The 31 controlled experiments (36.5%) that reported time as measure of outcome also used time as a measure of the experimental activities. However, these controlled experiments differed from those that reported time as dependent variable, in the sense that they did not explicitly refer to time as dependent variable. Due to this fact, it was more difficult to identify these time measures than the time measures described above. The time measures were to large extent found in other sections than in the ones describing the experimental design. The same naming practices as for the first experiments were also found amongst the controlled experiments here. We grouped the identified time measures into 16 time measure categories. Amongst these controlled experiments too, the purpose of the time measures was to determine the efficiency or effort of the experimental activities. An interesting finding amongst the controlled experiments in this experiment category was that a large majority did not use the terms dependent or independent variables. This could explain why time was not explicitly reported as dependent variable.

The nine controlled experiments (10.6%) that only reported time as time limits used time just to limit the execution of the experimental activities. This meant that time was not used as a measure of the subjects' performance of the experimental activities. It must be remarked that the controlled experiments in the other two experiment categories, also reported on time limits. Except for one controlled experiment that lasted 21 weeks, all the controlled experiments operated with time limits that ranged from one hour to approximately eight hours, with a median of two hours.

6.2.3 Time recording

The intention of investigating the time recording in the controlled experiments was to identify who/what recorded the time during the experimental activities, and to resolve possible trends in that regard.

We found three different ways of recording time. These were by computerized tools, by subjects and by experimenters. We analysed the frequency of the time recording means in the controlled experiments according to the three experiment categories. In addition, we also counted the numbers of controlled experiments that did not explicitly report how time was recorded. Overall, subjects (24 controlled experiments; 28.3%) and tools (23 controlled experiments; 27%) most often recorded the time of the experimental tasks. Only two controlled experiments (2.4%) were found to report experimenters as the ones who recorded the time. However, the largest group of controlled experiments (36 controlled experiments; 42.4%), were the ones that did not state their time recording means.

The time recording tools were most frequently found in the controlled experiments in the experiment category "Time as dependent variable" (17 controlled experiments; 19.9%). The experiment categories "Time as dependent variable" and "Time as measure of outcome" were quite alike in regards to the frequency of reporting subjects as the ones who recorded the time. None of the controlled experiments in "Time as time limit only", reported on any time recording means. In fact, this is not surprising, since the execution of the experimental activities in these controlled experiments were not measured by time, but only restricted by a time limit.

We identified 11 types of time recording tools amongst the 23 controlled experiments that reported time recording tools. The most commonly used type of tool was computer log files. These were found in seven (30.4%) out of the 23 controlled experiments.

From our investigation of time recording, we identified two possible trends. The first trend is that there has been a possible decrease in the use of time recording tools in controlled experiments. Our results show that only two controlled experiments published in the two last years, 2001 and 2002, reported time recording tools, while as many as 12 controlled experiments published in articles in 1996 and 1997 reported time recording tools. The reasons for this trend are unclear. Perhaps, the SE-researchers documenting the controlled experiments became less accurate in terms of reporting time recording tools, or perhaps, the SE-community found it too expensive or too complicated to use time recording tools in controlled experiments.

The other possible trend we identified is that there has been an increase in the last years of the surveyed articles in the number of controlled experiments that do not describe their time

recording. As many as nine out of 22 controlled experiments published in 2001 and 2002, were found not to explicitly report how they recorded their time measures. In comparison, only four out of 21 controlled experiments published in 1996 and 1997, did not report any time recording means. The reasons for this trend are also unclear.

6.2.4 Time units

The fourth aspect of time that we analysed was time units. The reason for investigating this aspect was to identify and evaluate the time units that time was specified in. This would provide a further understanding of the existing practices in using time in controlled experiments.

Our results showed that the 85 controlled experiments that reported time differed in regards to which time units they specified the time in. We found five time units. These were seconds, minutes, hours, days and weeks. The investigation of the overall frequency of each time unit showed that a majority of the controlled experiments specified the time in minutes, i.e. 48 controlled experiments (56.5%). In comparison, the second largest time unit, hours, was found in 29 controlled experiments (34.1%). For each of the three experiment categories, minutes were used by a dominating proportion of the controlled experiments in "Time as dependent variable". In "Time as measure of outcome", minutes and hours were the prominent time units and were used in the same number of controlled experiments. In "Time as time limit only", hours was the most common time unit.

An interesting finding from the analysis of the time units was that one controlled experiment in "Time as dependent variable" did not explicitly specify the kind of time unit it used. Obviously, this lack of information is grave due to its importance to the understanding of the time measures. And since the time measures in this controlled experiment was stated as dependent variable, the missing time unit reduces the clarity of the experimental results as well.

6.2.5 Time as validity threat

The fifth and last aspect of time that we studied was time as validity threat. The purpose of analysing this aspect was to resolve how time could be seen as a limit to the validation of the research results.

Our investigation showed that only nine out of the 85 controlled experiments that reported time, discussed time as validity threat. In the articles that described these controlled experiments, we identified 11 quotes that addressed time in relation to validity aspects. Only two out of the 11 quotes (the quotes in the fifth and the sixth article in Table 4.11) explicitly stated that time was a validity threat. Our identification of the other nine quotes was based on interpretations of the statements. The recognition of six of these nine quotes was also influenced by the fact that they were stated in chapters dealing with validity-issues, i.e. chapters with titles containing the word validity.

We found that time was considered as threats both to internal and external validity. Time recording and time limit were the two dominating aspects of time that were addressed as validity threats. They were addressed in ten out of the 11 quotes. This should indicate that

time recording and time limit are the aspects of time that SE-researchers regard as the potentially most limiting to the validation of the research results.

Although, we found some controlled experiments that addressed time as validity threats, we were surprised by the low frequency of this aspect overall. When one considers the extensive use of time in controlled experiments as the findings in section 4.1 and 4.2 are proof of, one could expect that time also would be central in validity-issues. We do not believe that the low frequency is a result of a dominating view amongst the SE-experimenters that aspects of time do not constitute real threats to the validity of controlled experiments. The findings amongst the controlled experiments that addressed time as validity threat show that time really can represent a threat. Rather, our opinion is that there is a lack of awareness in the SE-experimentation community of the potential threats that aspects of time can constitute.

6.3 Conclusions

In our opinion, the most interesting findings from the research were related to the controlled experiments that reported and used time as measure of the subjects' executions of the experimental activities, i.e. the experiment categories "Time as dependent variable" and "Time as measure of outcome". Our survey shows that time was used as a measure in a large majority (76 controlled experiments; 89.4%) of the 85 controlled experiments that were found to report time. These 76 controlled experiments differed, however, in regards to how they reported their time measures. 45 (59.2%) out of the 76 controlled experiments explicitly reported time as dependent variable, while 31 controlled experiments (40.8%) were not found to state time as dependent variable, although, in our opinion, they obviously used time in that sense. Based on these findings, our conclusion is that SE-researchers documenting controlled experiments should be more consistent by explicitly reporting time as dependent variable, when time is used in this sense. This must be seen in the light of the important role dependent variables play in controlled experiments. In addition, the significance of our conclusion can be illuminated by the following statement of Sjøberg et al. [6, p.31]: "A more uniform way of reporting experiments will help to improve the review of articles, replication of experiments, meta-analysis and theory building".

From the findings regarding time recording, our main conclusion is that SE-researchers documenting controlled experiments should describe how the recording of the time has been executed. We found that as many as 27 controlled experiments (35.5%), which constituted the largest proportion of the 76 controlled experiments that measured time, did not report their time recording. Describing the time recording is important as practical information for possible replications, and for the validity of the controlled experiment. The validity aspects of time which were dealt with in section 4.5, clearly showed that time recording can constitute a validity threat.

Our conclusion in connection with time units is that SE-researchers documenting controlled experiments should specify which time unit time is measured in, and preferably specify the time measures in minutes. It is reassuring that all controlled experiments except for one, were found to specify the time in time units. Specifying the time units is vital, since they clarify the time measures. Certainly, time measures that are not specified by time unit lead to imprecise outcomes. Our recommendation of specifying the time measures in minutes is due to the granularity that this time measure represents compared to, for example, hours. However, it

may be that it is more appropriate to use hours, if the granularity of the time data is not so important for the outcome of the controlled experiment.

In regards to the last aspect of time that we analysed, our conclusion is that SE-researchers documenting controlled experiments should address time in relation to validity threats to a much larger degree. In our investigation, we only found nine controlled experiments that discussed time as validity threat to the experiment. As the findings within these nine controlled experiments are proof of, time can constitute a threat to the validity. We recommend that SE-researchers should explicitly address time in relation to validity in chapters that are titled with the word validity. Obviously, it is important to address potential validity threats, since they may limit the validation of the research results. Hence, other researchers who want to learn, and perhaps replicate, the results from a controlled experiment should be informed of potential threats to the validity of the experiment.

It is our hope that the results, discussions, conclusions and recommendations of this master thesis would provide useful insight to SE-researchers, so that they can improve the design of their controlled experiments. This is vital in order to advance a state-of-the art practice in empirical software engineering research.

6.4 Future work

One proposal for future work is to analyse other dependent variables. As we stated in section 1.1, the survey of Jo E. Hannay at SIMULA RL showed that correctness was often used as dependent variable in controlled SE-experiments. Therefore, it could be interesting to investigate this dependent variable.

A proposal for future work that can elaborate on the results of this research, is to analyse time in indirect measures. For example, it could be interesting to find out what kind of direct measures time is most often combined with in the indirect measures.

A third proposal for future work, is to examine trends identified in this research in more recently published controlled SE-experiments, i.e. published after 2002. For example, it could be of interest to find out whether the use of tools in the time recording has increased in the most recent years.

7 References

[1] R.L. Glass, I. Vessey, and V. Ramesh. “Research in software engineering: an analysis of the literature”. Journal of Information and Software Technology, 44(8), 2002.

[2] N. Juristo and A.M. Moreno. “Basics of software engineering experimentation”. Kluwer Academic Publishers, 2001.

[3] B. Kitchenham. “Procedures for undertaking systematic reviews”. Preliminary draft, May 2004.

[4] W.R. Shadish, T.D. Cook and D.T. Campbell. “Experimental and Quasi-Experimental Designs for Generalised Causal Inference”. Houghton Mifflin Company, 2002

[5] Simula.no. [URL] http://www.simula.no/departments/engineering/projects/project.2005-04-08.5720533279, visited 23/9-05.

[6] D. Sjøberg, J.E. Hannay, O. Hansen, V.By, A. Karahasanovic, N.-K. Liborg, and A.C. Rekdal. ”A Survey of Controlled Experiments in Software Engineering”, 2004.

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

[8] C. Wohlin and M. Höst. “Editorial Special section: Controlled Experiments in Software Engineering”. Elsevier Science B.V. 2001.

[9] C. Wohlin, P. Rundeson, M. Höst, M.C. Ohlsson, B. Regnell and A. Wesslén. “Experimentation in Software Engineering: An Introduction”. Kluwer Academic Publishers, 1999.

[10] M.V. Zelkowitz and D. Wallace. “Experimental validation in software engineering”. Journal of Information and Software Technology, 39:735-743, 1997.

[11] A. Zendler. “A preliminary software engineering theory as investigated by published experiments”. Empirical Software Engineering, 6(2), 2001.