Software Industry Experiments: A Systematic Literature Review

Oscar Dieste
Universidad Politécnica de Madrid
Boadilla del Monte, Spain
odieste@fi.upm.es

Natalia Juristo
Universidad Politécnica de Madrid
Boadilla del Monte, Spain
natalia@fi.upm.es

Mauro Danilo Martínez
Escuela Politécnica del Ejército
Sangolquí, Ecuador
mdmartinez@espe.edu.ec

University of Oulu
Oulu, Finland
natalia.juristo@oulu.fi

Abstract—**Background: There is no specialized survey of experiments conducted in the software industry. Goal: Identify the major features of software industry experiments, such as time distribution, independent and dependent variables, subject types, design types and challenges. Method: Systematic literature review, taking the form of a scoping study. Results: We have identified 10 experiments and five quasi-experiments up to July 2012. Most were run as of 2003. The main features of these studies are that they test technologies related to quality and management and analyse outcomes related to effectiveness and effort. Most experiments have a factorial design. The major challenges faced by experimenters are to minimize the cost of running the experiment for the company and to schedule the experiment so as not to interfere with production processes. Conclusion: Companies appear to be disinclined to run experiments because they are not perceived to have direct benefits. We believe that researchers staging a field experiment in a company should adopt a business-aligned stance and plan an experiment that clearly benefits managers and professionals.**

I. Introduction

Software engineering (SE) experiments are becoming increasingly common in academia. The experimental subjects of these studies are usually students with little or no professional experience. Additionally, the experimental setting and materials tend to be artificial or only partially related to real projects [1]. All this raises concern about whether the results can be generalized. For example, Höst et al. contend that there are only slight differences between students and professionals [2]. Contrariwise, Dybå et al. observe substantial differences between students and professionals, and also find that the setting and the materials have a marked influence on the experimental results [3]. Runeson's findings are similar [4].

Experiments run in industry are often considered to produce more generalizable results than their counterparts run in academia [5]. However, the SE community has taken little notice of field experiments in industry. For example, no specialized survey reporting information about how many experiments have been run in industry, what factors they tested, what response variables they used, etc. has so far been conducted.

At first glance, this information may not appear to be of much consequence. However, the experiments that have been run must have been of some interest to industry, otherwise they would never have been conducted in the first place. An understanding of their features is likely to be useful for preparing proposals better adapted to real-world businesses and, consequently, further maturing the experimental paradigm applied to SE research. Note that no experimental discipline can generate generalizable knowledge unless it runs both laboratory and field experiments. Consider medicine, for example. Laboratory experiments using animals are not the be all and end all; experiments run in hospitals on real patients (clinical trials) are just as necessary.

This paper reports the preliminary results of a systematic literature review, namely a scoping study [6, 7]. The paper is structured as follows. Section II describes the related work. Section III states the research questions and describes the study research methodology. The review protocol is attached as Appendix A. Section IV describes how we conducted the study. Section V outlines the answers to the stated research questions, which are discussed in Section VI. Finally, Section VII lists the validity threats to this study.

II. Related Work

Scoping studies by Sjøberg et al. [1] and Kampenes et al. [8] are currently the primary source of the information about experiments run in industry. Both studies examine experiments and quasi-experiments published from 1993 to 2002.

Sjøberg et al. classify experiments by *location*, which has two possible values: *office environment* or *laboratory/classroom*. Experiments run in an *office environment* can, for all practical purposes, be equated to industrial experiments. However, the 103 experiments that Sjøberg et al. identified are not well reported, which prevents them from being reliably classified. In actual fact, only one of the identified 103 experiments is clearly reported to have been run in an *office environment*.

An alternative way of inferring which experiments were conducted in industry is to consider the type of experimental subjects used. Experiments using professionals are more likely, albeit not certain, to have been run in industry. In this respect, Sjøberg et al. identify 27 studies (22 experiments and five quasi-experiments) with professional participants. Of these 27 studies, 17 do not explicitly state the type of *location* in which they were conducted, and another seven were reported to have been run in a *laboratory/classroom location*. The experimental subjects of the only experiment run in an *office environment* were, as you would expect, professionals. The inference is then that probably no more than a total of about three experiments were run in industry during the study period (1993-2002).

Experiments conducted with professionals have distinctive features. Sjøberg et al. have determined that the number of subjects and the total workload tend to be considerably less than for experiments conducted with students. Kampenes et al. suggest that this can be put down to the cost factor. Interestingly, Sjøberg et al. state that there are no differences between the two subject types (professionals and students) with respect to the duration of the experiments. These can realistically be assumed to be typical features of experiments run in industry.

Finally, Sjøberg et al. report that seven out of the 27 experiments that used professionals provided information on their background. Subjects are catalogued as developers, practitioners, analysts, professionals, etc. These are generally nonspecific names and do not provide information about the exact activities that the professionals perform in their routine work. From seven to 13 experiments with professionals also provide information on experience (the number of experiments varies depending on the type of reported experience).

Therefore, the information available about experiments in industry has been gathered in a roundabout way (i.e. experiments with professionals are used to infer which experiments may have been run in industry). Besides it is confined to a rather vague idea of the total number of experiments, plus some methodological attributes, such as sample size, duration and subject types.

III. Research Questions and Methodology

We have run a scoping study which has surveyed experiments conducted in industry (i.e. experiments actually run at companies, not experiments with professional participants) up until July 2012. The stated research questions were:

RQ1. How many experiments have been run in industry and what is the observed experiment time distribution?

RQ2. What independent variables (/technologies) do they study?

RQ3. What dependent variables do they study?

RQ4. What types of experimental designs do they use?

RQ5. How many and what categories of subjects participate in industry experiments?

RQ6. What challenges does experimentation in industry raise?

The research methodology that we used was a systematic literature review (SLR) compliant with Kitchenham et al.'s guidelines [7]. The review protocol is attached as Appendix A.

IV. Review Execution

The search strategy replicates the structure of the review question. For each PICOC term [7] used (see Annex A), we have defined search substrings (denoted by numbers for usability) by connecting several keywords by conjunction (logical connector OR). Table I shows these substrings and the associated keywords.

The keywords used to run searches of SE experiments with respect to Intervention were sourced from [9]. The keywords for Context and Population were obtained by identifying synonyms, as suggested in [7]. The final string was built by connecting the Intervention, Context and Population terms (using the logical connector AND).

TABLE I. Search Strategy

Search substring (keywords linked by OR)	Keywords	PICOC term
1	Software	Population
2	Experiment Empirical Empirical study Empirical evaluation Experimentation Experimental comparison Experimental analysis Experimental evidence Experimental setting Empirical data	Intervention
3	Industry / Industries	Context
4	Company / Companies	
5	Business / Businesses	
6	Enterprise / Enterprises	

A total of five searches were run on the SCOPUS database. Table II shows the results. The first four searches were planned in the review protocol, whereas the last was run after reading the titles and abstracts of the papers identified by the first four searches. Altogether the searches returned a total of 658 articles (including duplicates).

TABLE II. Results of Search and Selection of Primary Studies

	Search strings[a]	Papers		
		Identified	*Pre-selected*	*Selected*
Planned	1 AND 2 AND 3	117	16	6
	1 AND 2 AND 4	129	11	4
	1 AND 2 AND 5	188	1	0
	1 AND 2 AND 6	64	4	0
Unplanned	1 AND 2 AND "Industrial"	160	16	8
	Total		39	15

a. The search string codes are taken from Table I

We used the inclusion criteria (applied to the title and abstract) to pre-select 39 papers. The exclusion criteria (applied to the full text) led to the rejection of 23 papers. Additionally, we detected one duplicate paper. As a result, 15 papers were finally selected. There are some overlaps across searches that explain why the totals and subtotals do not add up.

The 15 primary studies are shown in Appendix B. Ten of the studies report experiments and five, quasi-experiments. For each study, we extracted the data specified in the review protocol, which are available at [10]. Finally, we studied and synthesized the extracted data in response to the stated research questions, which were plotted and/or tabulated. The results are reported in the next section.

b.

V. Results

A. (RQ1) How many experiments have been run in industry and what is the observed experiment time distribution?

As shown in Figure 1, two studies were conducted as early as 1997, but not many experiments were run in industry until 2003. The rate of experiments run in industry rose to approximately two per year over the 2003-2008 period. As of 2009, there was a sharp drop in the number of experiments conducted.

Our hypothesis to explain this decrease is that companies are reluctant to engage in non-productive activities in the current scenario of economic recession. We rule out a loss of interest on the part of researchers in running experiments in industry as a possible explanation because, as members of the SE research community, we have perceived clear signs that this is not the case; rather the contrary.

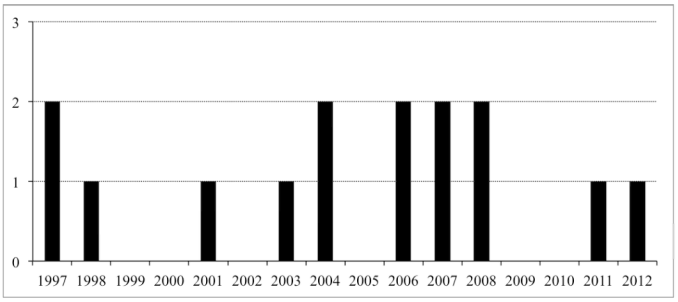

Fig. 1. Experiments run by year of publication

B. (RQ2) What independent variables (/technologies) do industrial experiments study?

The independent variables used in the identified experiments and quasi-experiments are available at [10]. The number of independent variables is more or less equivalent to the number of studies, and no patterns were observed with regard to the specific technology tested. However, patterns are clearer if the perspective is raised to the level of development activity or paradigm tested in the experiments, as shown in Table III.

In this respect, Inspection is the most often tested technology (31% of studies), followed by Estimation Techniques (19%). Grouped by areas rather than by technologies, Quality (= Inspection + Testing) would take the top slot with 47% of primary studies (7 out of 15), followed by Management (= Estimation + Agile), which would be ranked second with 25%, and third place would go to Object-Orientation (= Object-oriented development + UML models) with 13%. Altogether, these three areas account for 81% of all the experiments (and quasi-experiments) in industry.

TABLE III. Tested technologies

Area	*Technologies*	*Studies*	*Total*
Quality	Inspection	E3, E4, E7, E10, QE4	5
	Testing	E2, E6	2
Management	Estimation	QE5, QE3	2
	Agile	E1, E5	2
Object-Orientation	Object-oriented development	E8	1
	UML models	E9	1
	Others	QE1, QE2	2

The percentage of studies addressing the Quality area is only slightly higher than the figure given by Sjøberg et al. [1] for all SE experiments (rated at 35%). However, the percentage of studies identified by this scoping study that fall into the Management area is substantially higher than was reported by Sjøberg et al. (where the percentage was 7.1%). This merely goes to strengthen the impression that experiments run in companies tend to be aligned with business goals/processes, which do not always tie in with researchers' interests.

TABLE IV. Techniques used

Technologies	*Techniques*	*Studies*	*Year of publication*
Inspection	Perspective-based reading	E4	1997
	Perspective-based reading Checklist-based reading	QE4	2001
	-	E3, E7, E10	1997, 2003, 2008
Testing	Test-driven development	E2, E6	2004, 2006
Estimation	-	QE5, QE3	1998, 2012
Agile	Pair design	E1	2007
	Pair programming	E5	2007
Object-oriented development	UML	E8	2004
UML models	UML	E9	2006
Others	-	QE1, QE2	2008, 2011

a. Years are given in increasing order and do not correspond to the column headed Studies

Finally, we tried to identify the SE techniques used in the experiments and examine whether there is any sort of pattern in their use. Note that *technique* is not necessarily equivalent to the *independent variable*, as the technique often only represents the way in which the experimental task is performed. This is the case of study E4, for example, where the experimental task is to review a requirements specification using *perspective-*

based reading, but one of the independent variables is system domain.

Table IV shows the results. We use "-" to denote experiments that use techniques that are not representative of the state of the art in SE. We observe what appear to be two patterns, which should be considered with due caution given the limited data available. First, just over 50% of studies use widespread techniques. Second, techniques are not tested in industry for a good many years after they are invented. This appears to suggest that companies have a preference for mature technologies. The four studies on Test-driven development and Pair design/programming are an exception; they were run soon after the techniques were invented (within 1-4 years compared with over 10 years for UML). These technologies raised great expectations, which might be plausible explanation for this phenomenon.

b.

C. (RQ3) What dependent variables do industrial experiments study?

As shown in Table V, there are three main response variables: effectiveness (used in 60% of studies), effort (33%) and quality (27%). Note that, as an experiment can study more than one response variable, the sum of the above percentages is greater than 100%. The three variables refer to key business aspects, and so their majority use is by no means surprising.

Additionally, each response variable can be measured using different metrics. Thus, for example, effort can be measured using the *time* metric (which is reasonable when the respective task is performed by only one person as in experiment E1) or by means of the *person/hours* metric, as applies in experiment E5. As shown in Table 5, the *number of defects* and *time* are, predictably, the most commonly used metrics.

TABLE V. RESPONSE VARIABLES USED

Response variable	*Most common metrics*	*Studies*	*Total*
Effectiveness	Number of defects (9 cases)	E3, E4, E7, E8, E9, E10, QE1, QE4, QE5	9
Effort	Time (5 cases)	E1, E3, E5, E7, QE2	5
Quality	-	E1, E2, E5, E6	4
Others	-	E2, E6, QE3	3

D. (RQ4) What types of designs do experiments use?

As Table VI shows, most experiments used factorial designs [11] (60% of studies), testing one, two or, more often, three factors simultaneously. There is one case of a fractional factorial design [12] (7%). As a whole, factorial designs account for 67% of all studies. Some counter-balanced (13%) and unbalanced (6%) cross-over designs [13] were also identified.

The proportion of factorial and cross-over experiments is surprising. In a factorial experiment, each subject applies one and only one of the experimental treatments just once. Now, one weakness of SE experiments is that the number of experimental subjects is typically small. Sjøberg et al. [1] report that the average number of subjects in SE experiments is 48.6; when subjects are professionals, the average is 20.0. Less than 50 subjects are generally considered small samples [14]. Consequently, factorial designs run the risk of not achieving the statistical power necessary to be able to detect significant differences.

TABLE VI. EXPERIMENTAL DESIGNS

Design		*Studies*	*Total*
Factorial	Full	E2, E3, E4, E5, E8, E9, E10, E7, QE1	9
	Fractional	QE2	1
Cross-over	Counterbalanced	E1, E6	2
	Unbalanced	QE4	1
Correlational study		QE3, QE5	2

Cross-over designs offer the possibility of multiplying the experiment sample size in terms of the number of experimental sessions executed, the only trade-off being the risk of a carry-over effect [15]. An increased sample size raises statistical power. On top of this, as there are few opportunities for running experiments in industry, one would expect most researchers to opt for cross-over designs, which is contrary to the observed pattern.

The most plausible explanation is that companies are reluctant to apply cross-over designs and prefer factorial designs. Experiments with factorial designs tend to be run in a single session and are, therefore, faster and less expensive, as well as easier to plan and execute (e.g. they need only one rather than two or three free slots in the schedule of the participating professionals). Workload minimization is a feature that Sjøberg et al. [1] also point out as being characteristic of experiments with professionals.

Finally, there are two quasi-experiments (QE3, QE5) where all subjects perform all the experimental tasks. Quasi-experiments necessarily apply correlation or dichotomization during the analysis phase.

E. (RQ5) How many and what categories of subjects participate in experiments?

Although all the primary studies were conducted at industrial sites, the subjects used in four cases (E8, E9, E10 and QE1) are both students and professionals. This is because these studies compare novice (student) against expert (professional performance). However, as the goal of this research is to gain a better understanding of experiments conducted in industry using practitioners, we will focus exclusively on the type of professionals participating in the experiments. This typology is shown in Table VII.

Generally, we, like Sjøberg et al. [1], find that the names used to refer to the professionals are very vague. The most common term is Professionals (27% of cases), followed by Software developers (20%), Engineers/software engineers (13%) and Developers (13%). Interestingly, Practitioner, which is a fairly common term in the literature, is only used once.

Most of the reviewed papers report professional experience poorly. Only two studies (E4, E9) specify experience in years. Sjøberg et al. [1] already detected this weakness. The other studies either fail to reference experience at all or class

professionals according to common terms like "Junior", "Senior", without giving a precise definition of their values or bounds. There are even cases where experience is rated with respect to the subject's academic qualifications (e.g.; "holds a BSc" or "holds an MSc"), which is evidently a different (albeit interesting) variable, namely, background.

TABLE VII. SUBJECT CATEGORIES AND NUMBER

Subject type	*Studies*	*Total number of subjects*	*Average*
Professionals	E5, E8, E10, QE1	382	96.5
Software developers	E4, QE3, QE4	445	148.3
Engineers/software engineers	E1, QE2	26	13
Developers	E3, E7	21	11.5
Programmers	E2	24	-
Practitioners	E9	44	-
Employees	E6	28	-
Others	QE5	68	-
Total		1,038	69.2

The total number of professionals that participated as experimental subjects in the 15 primary studies amounts to 1,038, where the mean value per study is 69.2. This figure is quite a lot higher than the 20 subjects per study reported by Sjøberg et al. [1]. However, we believe that 69.2 is a misleading figure. There are three experiments (E5, E8, E10 and QE3) that have unusually high sample sizes (197, 99, 73 and 374, respectively). If we were to exclude these three studies from the average, the resulting value would be [1,038 – (197 + 99 + 73 + 374)] ÷ (15 – 4) = 26.8, which is more in line with Sjøberg et al.'s findings.

There does not appear to be any relationship whatsoever between the number of subjects used in a study and the study's other features. With respect to design type, for instance, cross-over experiments do not use noticeably fewer subjects than other designs, especially factorial designs.

F. (RQ6) What challenges does experimentation in industry raise?

Researchers refer to several obstacles to experimenting in industry. The most important concerns are the time and cost demands that running an experiment places on the host company and the participating professionals. Often time problems may even threaten the validity of the experiment by either enforcing the use of substandard designs (due to time shortages) or small samples (due to the unavailability of subjects). This shows that, like laboratory experiments, experiments in industry are not a win-win situation either. The external validity of laboratory experiments would appear to be untrustworthy (although this belief has yet to be confirmed experimentally), but their internal validity is potentially sound. On the other hand, experiments in industry have the potential to achieve high external validity, whereas their internal validity is questionable.

The biggest drawback appears to be time. In this respect, study E9 indicates that *"because of professionals' time constraints they performed only one experiment run"*. Similarly, study E3 points out that *"the experiment might be assumed as time-consuming for the project, causing delay and hence being rejected"*. With respect to cost, study E4 states that *"in many organizations it is hard to motivate experimental studies because organizations are concerned about financial issues"*.

In view of the time and cost problems, study E1 arrives at quite a reasonable explanation for why there are so few experiments in industry: *"it is difficult to find professionals for empirical studies, whereas students are more accessible, easier to organize, and cheaper"*.

Other problems, apart from time and cost, also stand in the way to running experiments in industry. Study E3 particularly mentions aspects related to experiment workload and planning. With respect to workload, study E3 suggests that the workload spent on the experiment should be minimized. Specifically, it states *"Good planning and preparation was necessary to minimize the effort spent"*. With respect to planning, E3 refers to it being hard to establish a definite schedule. In this respect, it reports that *"the time schedule for the experiment had to be coordinated with the internal [...] plan. In fact, the experiment was delayed for almost one month"*, and goes on to highlight the fact that *"the industrial reality at [...] is very hectic, and pre-planning of all details was not feasible"*.

Finally, the most surprising hurdle is a sociological concern and refers to the *academicism*, which experiments are in many milieu assumed to have, as opposed to the *reality* of industry. Study E4 reads: *"We realized that the term "experiment" itself was demotivating because it focuses much more on the academic than the industrial benefit. Thus we used the term training instead. Perhaps we must show the value experimental studies have to motivate for them, e.g., that the results can be used as an input in the companies' experience factory"*.

VI. DISCUSSION

From the review that we have conducted, we conclude that the situation of software industry experiments is unsure. This uncertainty is patent from both the shortage of studies and the obstacles to running experiments at companies.

Although the joint interpretation of data as disparate as the information reported in Section V is necessarily subjective, the conclusion that we have reached is that the window of opportunity for running experiments in industry is very narrow and is linked to three factors:

- **Interference of experiments in production processes:** experiments should not be presented or allowed to be conceptualized as extra work. Whenever possible, the company should be able to use the experiment result directly (e.g. an inspection experiment run using software specifications of the ongoing project). Otherwise, the experiment should be designed to at least represent the practical part of a training course.

- **Alignment with business goals:** the experiment should be run on a topic that is directly useful to the company. Verification and validation or estimation technologies are potentially good candidates. Additionally, outcomes related to effectiveness, efficiency or quality appear to be more interesting to industry. Negotiation with the company in pursuit of a win-win situation is the best possible alternative, but business goals take priority, and researchers should fit in with this constraint.
- **Human resource optimization:** experiments should take up as little of professionals' time as possible. In particular, experiments with multiple sessions like cross-over designs are less suitable than single-session experiments with factorial designs.
- **Schedule flexibility:** experiments cannot be planned to a strict schedule, and execution times have to be flexible. Consequently, researchers should carefully consider whether cross-over experiments are a good option, as it might not be possible to hold temporally adjacent sessions as required by the experimental protocol.

VII. Validity Threats

The main validity threat to this research is the use of only one bibliographic database: SCOPUS. This could result in the identified primary studies representing only a subset of all studies, probably limited to reputed publications (journals and top-ranking conferences). This would bias the results and, consequently, lead to mistaken findings. However, the threat of bias is very unlikely to materialize. On the one hand, SCOPUS indexes publications from other databases like IEEE, ACM, Springer and Elsevier. Therefore, coverage is wide, and the identified primary studies were actually published by several of the above publishers. On the other hand, the SE researcher community sets such store by the very few studies conducted in industry that they are unlikely to be published in low-ranking media. This maximizes the likelihood of their being located in SCOPUS. A secondary validity threat is related to the keywords used for study selection. We have used pre-packaged search strings aimed at locating experiments, but those strings were tested mainly using experiments from academia [9]. It is possible that experiments in industry are referred to using different terms such as *field experiment*, to cite an example. We will explore those synonyms in future research.

VIII. Conclusions

This paper reports the preliminary results of a scoping study exploring the features of experiments run in industry. We have located a total of 15 studies (10 experiments and five quasi-experiments). For this set of primary studies, we have gathered data about time distribution, independent and dependent variables, types of professional subjects, experimental designs used and the pros and cons of running experiments in industry.

The results highlight that experimenting in industry is generally perceived by SE community to be problematic. Few experiments have been run, and their number has dropped sharply over that last four years (probably due to the economic recession). Additionally, companies are disinclined to run experiments because they are not perceived as having any direct benefits.

Finally, we believe that researchers promoting an experimental study in a company should adopt a business-aligned viewpoint and plan an experiment that clearly benefits the respective managers and professionals.

Acknowledgment

This research has been partially supported by the grants TIN-2011-23216 (Spanish Ministry of Economy and Competitiveness), FiDiPro (Finnish Funding Agency for Technology and Innovation) and CA2011 (SENESCYT Ecuador).

References

[1] D. Sjøberg, J. Hannay, O. Hansen, V. Kampenes, A. Karahasanovic, N. Liborg and A. Rekdal, "A survey of controlled experiments in software engineering," *IEEE Transactions on Software Engineering,* vol. 31, pp. 733-753, 2005.

[2] M. Höst, B. Regnell and C. Wohlin, "Using Students as Subjects - A Comparative Study of Students and Professionals in Lead-Time Impact Assessment," *Empirical Software Engineering,* vol. 5, pp. 201-214, 11/01, 2000.

[3] T. Dybå, D. I. K. Sjøberg and D. S. Cruzes, "What works for whom, where, when, and why?: on the role of context in empirical software engineering," in *5th ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2012)*, Lund, Sweden, 2012, pp. 19-28.

[4] P. Runeson, "Using students as experiment subjects – an analysis on graduate and freshmen student data," in *7Th International Conference on Empirical Assessment & Evaluation in Software Engineering (EASE 2003),* Keele, UK, 2003, pp. 95-102.

[5] L. G. Votta, "By the way, has anyone studied any real programmers, yet?" in *9th International Software Process Workshop*, Airlie, Virginia, USA, 1994, pp. 93-95.

[6] M. Petticrew and H. Roberts, Systematic Reviews in the Social Sciences: A Practical Guide. Wiley-Blackwell, 2005.

[7] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering. version 2.3. Keele University. 2007.

[8] V. B. Kampenes, T. Dybå, J. E. Hannay and D. I. K. Sjøberg, "A systematic review of quasi-experiments in software engineering," *Information and Software Technology,* vol. 51, pp. 71-82, 1, 2009.

[9] O. Dieste, A. Griman and N. Juristo, "Developing search strategies for detecting relevant experiments," *Empirical Software Engineering,* vol. 14, pp. 513-539, 2009.

[10] M. D. Martínez. La ingeniería del software experimental en la industria: Una revisión sistemática. 2013.

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

[12] G. E. P. Box, J. S. Hunter and W. G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery.* Wiley-Interscience, 2005.

[13] B. W. J. Brown, "The Crossover Experiment for Clinical Trials," *Biometrics,* vol. 36, pp. 69-79, 1980.

[14] F. Richy, O. Ethgen, O. Bruyere, F. Deceulaer and J. Reginster, "From Sample Size to Effect-Size: Small Study Effect Investigation (SSEi)," *The Internet Journal of Epidemiology,* vol. 1, 2004.

[15] B. Kitchenham, "The case against cross-over designs in software engineering," in 11th *International Workshop on Software Technology and Engineering Practice (STEP 2003)*, Amsterdam, The Netherlands, 2003, pp. 65-67.

APPENDIX A. REVIEW PROTOCOL

We created the review protocol according to Kitchenham et al. guidelines [7]. The protocol is composed of five parts that we summarize below: review question, search strategy, selection process, criteria for including and excluding a study and data extraction. The inclusion/exclusion criteria provide a full definition of the target population of primary studies, for which reason we omit an explicit quality assessment process.

A. Review Question

We used PICOC to define the research question:

- **Population:** Studies conducted in SE.
- **Intervention:** Experimental or quasi-experimental methodology.
- **Context:** Industry.

We omitted the *Comparison* and *Outcome* terms, as this research aimed to conduct a scoping study to determine the features of experiments run in the software industry rather than to study the outcome of a treatment.

B. Search Strategy

The strategies used to construct the search string from the research question were as follows:

- Identify related words and synonyms for PICOC terms. We followed Dieste et al.'s recommendations [9] for the term *Intervention*, in particular.
- Use the logical operator OR to link synonyms.
- Use the logical operator AND to link different PICOC terms.

Table I lists the search strings used, which it would be redundant to reproduce here. The searches will be based on the SCOPUS database. We selected this database because it indexes publications by the major publishers, like IEEE, ACM, Springer and Elsevier. Publications must be written in English.

C. Selection Process

The selection of primary studies will be divided into two phases.

- In the first phase, we will apply the inclusion criteria to select a preliminary set of papers. To do this, we will check the title, abstract and keywords.
- In the second phase, we will apply the inclusion/exclusion criteria to the full text of the preliminary selection of papers, paying special attention to the introduction and methodology. This will result in the final selection of primary studies.

D. Inclusion/Exclusion Criteria

The inclusion criteria defined for this scoping study are designed to select papers that report an empirical SE study conducted in a company:

- The paper must address the SE area.
- The paper must report an empirical study.
- The paper must be contextualized in industry.

The exclusion criteria are designed to reject any studies that are not experiments or quasi-experiments run in a company using professionals:

- The study design must be experimental or quasi-experimental.
- The experimental subjects must be professionals.
- The professional subjects must have participated in an experiment or quasi-experiment at an industrial site.

E. Data Extraction

For each study we will gather: study title, publication year, authors, number and type of subjects, dependent and independent variables, design type and reported problems.

APPENDIX B. PRIMARY STUDIES

A. Experiments

Code	*Reference*
E1	Canfora, G., Cimitile, A., Garcia, F., Piattini, M., Visaggio, C.A. **Evaluating performances of pair designing in industry** (2007) *Journal of Systems and Software*, 80 (8), pp. 1317-1327.
E2	George, B., Williams, L. **A structured experiment of test-driven development** (2004) *Information and Software Technology*, 46 (5 SPEC. ISS.), pp. 337-342.
E3	Conradi, R., Mohagheghi, P., Arif, T., Hegde, L.C., Bunde, G.A., Pedersen, A. **Object-oriented reading techniques for inspection of UML models - An industrial experiment** (2003) *Lecture Notes in Computer Science*, 2743, pp. 483-500.
E4	Laitenberger, O., DeBaud, J.-M. **Perspective-based reading of code documents at Robert Bosch GmbH** (1997) *Information and Software Technology*, 39 (11), pp. 781-791.
E5	Arisholm, E., Gallis, H., Dybå, T., Sjøberg, D.I.K. **Evaluating pair programming with respect to system complexity and programmer expertise** (2007) *IEEE Transactions on Software Engineering*, 33 (2), pp. 65-86.
E6	Canfora, G., Cimitile, A., Garcia, F., Piattini, M., Visaggio, C.A. **Productivity of test driven development: A controlled experiment with professionals** (2006) *Lecture Notes in Computer Science*, 4034 LNCS, pp. 383-388.
E7	Porter, A.A., Siy, H.P., Toman, C.A., Votta, L.G. **An experiment to assess the cost-benefits of code inspections in large scale software development** (1997) *IEEE Transactions on Software Engineering*, 23 (6), pp. 329-346.
E8	Arisholm, E., Sjøberg, D.I.K. **Evaluating the effect of a delegated versus centralized control style on the maintainability of object-oriented software** (2004) *IEEE Transactions on Software Engineering*, 30 (8), pp. 521-534.
E9	Lange, C.F.J., Chaudron, M.R.V. **Effects of defects in UML models - An experimental investigation** (2006) *Proceedings - International Conference on Software Engineering*, 2006, pp. 401-410.
E10	Carver, J.C., Nagappan, N., Page, A. **The impact of educational background on the effectiveness of requirements inspections: An empirical study** (2008) *IEEE Transactions on Software Engineering*, 34 (6), pp. 800-812.

B. Quasi-experiments

Code	*Reference*
QE1	Bishop, B., McDaid, K. **Expert and novice end-user spreadsheet debugging: A comparative study of performance and behaviour** (2011) *Journal of Organizational and End User Computing*, 23 (2), pp. 57-80.
QE2	Sinnema, M., Deelstra, S. **Industrial validation of COVAMOF** (2008) *Journal of Systems and Software*, 81 (4), pp. 584-600.
QE3	Jørgensen, M., Grimstad, S. **Software development estimation biases: The role of interdependence** (2012) *IEEE Transactions on Software Engineering*, 38 (3), pp. 677-693.
QE4	Laitenberger, O., Emam, K.E., Harbich, T.G. **An internally replicated quasi-experimental comparison of checklist and perspective-based reading of code documents** (2001) *IEEE Transactions on Software Engineering*, 27 (5), pp. 387-421.
QE5	Stensrud, Erik, Myrtveit, Ingunn **Human performance estimating with analogy and regression models: An empirical validation** (1998) *International Software Metrics Symposium, Proceedings*, pp. 205-213.