Information and Software Technology 51 (2009) 71–82

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

A systematic review of quasi-experiments in software engineering

Vigdis By Kampenes [a,b,*], Tore Dybå [a,c], Jo E. Hannay [a,b], Dag I. K. Sjøberg [a,b]

[a] Department of Software Engineering, Simula Research Laboratory, P.O. Box 134, NO-1325 Lysaker, Norway
[b] Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, NO-0316 Oslo, Norway
[c] SINTEF ICT, NO-7465 Trondheim, Norway

ARTICLE INFO

Article history:
Received 20 June 2007
Received in revised form 15 April 2008
Accepted 23 April 2008
Available online 30 April 2008

Keywords:
Quasi-experiments
Randomization
Field experiments
Empirical software engineering
Selection bias
Effect size

ABSTRACT

Background: Experiments in which study units are assigned to experimental groups nonrandomly are called quasi-experiments. They allow investigations of cause–effect relations in settings in which randomization is inappropriate, impractical, or too costly.
Problem outline: The procedure by which the nonrandom assignments are made might result in selection bias and other related internal validity problems. Selection bias is a systematic (not happening by chance) pre-experimental difference between the groups that could influence the results. By detecting the cause of the selection bias, and designing and analyzing the experiments accordingly, the effect of the bias may be reduced or eliminated.
Research method: To investigate how quasi-experiments are performed in software engineering (SE), we conducted a systematic review of the experiments published in nine major SE journals and three conference proceedings in the decade 1993–2002.
Results: Among the 113 experiments detected, 35% were quasi-experiments. In addition to field experiments, we found several applications for quasi-experiments in SE. However, there seems to be little awareness of the precise nature of quasi-experiments and the potential for selection bias in them. The term "quasi-experiment" was used in only 10% of the articles reporting quasi-experiments; only half of the quasi-experiments measured a pretest score to control for selection bias, and only 8% reported a threat of selection bias. On average, larger effect sizes were seen in randomized than in quasi-experiments, which might be due to selection bias in the quasi-experiments.
Conclusion: We conclude that quasi-experimentation is useful in many settings in SE, but their design and analysis must be improved (in ways described in this paper), to ensure that inferences made from this kind of experiment are valid.

Contents

1. Introduction . . . 72
2. Background . . . 72
   2.1. Methods of randomization . . . 73
   2.2. Selection bias, the problem with quasi-experimentation . . . 73
   2.3. Design of quasi-experiments . . . 73
   2.4. Analysis of quasi-experiments . . . 74
3. Research method . . . 75
   3.1. Identification of experiments . . . 75
   3.2. Information extracted . . . 75
4. Results . . . 76
   4.1. Extent of quasi-experiments . . . 76
   4.2. Design of quasi-experiments . . . 76
      4.2.1. The use of pretest scores . . . 76
      4.2.2. The assignment procedures . . . 76
      4.2.3. The field experiments . . . 77
      4.2.4. The use of teams . . . 77

* Corresponding author. Address: Department of Software Engineering, Simula Research Laboratory, P.O. Box 134, NO-1325 Lysaker, Norway. Tel.: +4767828317; fax: +4767828201.
E-mail addresses: vigdis@simula.no (V.B. Kampenes), tore.dyba@sintef.no (T. Dybå), johannay@simula.no (J.E. Hannay), dagsj@simula.no (D.I. K. Sjøberg).

0950-5849/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2008.04.006

4.3. Analysis of quasi-experiments . . . 78
5. Discussion . . . 79
5.1. Extent of quasi-experimentation . . . 79
5.2. Results from quasi-experiments compared with randomized experiments . . . 80
5.3. Indicators of subject performance . . . 80
5.4. Quality of reporting . . . 80
5.5. Ways to improve quasi-experimental designs in SE . . . 80
5.5.1. Nonequivalent experimental group designs . . . 81
5.5.2. Haphazard assignment . . . 81
5.5.3. Some randomization . . . 81
5.5.4. Within-subject design in which all participants apply the treatments in the same order . . . 81
5.6. Limitations of this review . . . 81
6. Conclusion . . . 81
Acknowledgements . . . 82
References . . . 82

1. Introduction

In an experiment, an intervention is introduced deliberately to observe its effects. This is the control that essentially allows the observation of treatment–outcome relations in experiments. Internal validity pertains to the validity of inferring causal relationships from these observations, that is, "whether observed co-variation between *A* (the presumed treatment) and *B* (the presumed outcome) reflects a causal relationship from *A* to *B* as those variables were manipulated or measured" [42]. A challenge in this respect is that changes in *B* may have causes other than the manipulation of *A*. One technique to help avoid such alternative causes is *randomization*, that is, the random assignment of study units (e.g., people) to experimental groups, including *blocked or stratified randomization*, which seeks to balance the experimental groups according to the characteristics of the participants.

However, randomization is not always desirable or possible. For example, in software engineering (SE), the costs of teaching professionals all the treatment conditions (different technologies) so that they can apply them in a meaningful way may be prohibitive. Moreover, when the levels of participants' skill constitute treatment conditions, or if different departments of companies constitute experimental groups, randomization cannot be used. Also, randomization might be unethical. For example, if a new technology is compared with old technology, randomly assigning students to either of these technologies can be unethical, because the value of the experience and knowledge obtained through the experiment can differ for the two groups of students. Such experiments are best performed as quasi-experiments, including professionals already familiar with the technologies. Note that effort must be made to ensure that the two groups have participants with fairly similar skills in the technology they apply.

Laitenberger and Rombach [26] claim that quasi-experiments (in which study units are assigned to experimental groups nonrandomly) represent a promising approach to increasing the amount of empirical studies in the SE industry, and Kitchenham [24] suggests that researchers in SE need to become more familiar with the variety of quasi-experimental designs, because they offer opportunities to improve the rigour of large-scale industrial studies.

Different nonrandom assignment procedures produce different potential alternative causes for observed treatment effects. Hence, in order to support internal validity in quasi-experiments, these potential alternative causes must be identified and ruled out. This is done in the design and analysis of the experiment, for example, by measuring a pretest score and adjusting for initial group differences in the statistical analysis. According to Shadish [40], the theory of quasi-experimentation [4,5,8] provides (1) alternative experimental designs for studying outcomes when a randomized experiment is not possible, (2) practical advice for implementing quasi-experimental designs, and (3) a conceptual framework for evaluating such research (the validity typology). The theory was developed for research in social science and has also been recognized in other fields of research, such as medical informatics [15], environmental research [38], and economics [31].

Even though the theory of quasi-experiments asserts that quasi-experimentation can yield plausible inferences about causal relationships [40], it seems that in many disciplines there is little awareness of the fact that proper inferences from quasi-experiments require methods different from those used for randomized experiments. Shadish et al. [42] claim that the most frequently used quasi-experimental designs typically lead to causal conclusions that are ambiguous, and empirical results from research in medical science and psychology indicate that randomized experiments and quasi-experiments provide different results [6,7,16,41]. The purpose of this article is to report the state of practice in SE on these matters. This is done by a systematic review of the 113 experiments reported in the decade from 1993–2002 in 12 leading journals and conference proceedings in SE. This review is a separately performed part of a review program that assesses the state of practice of experimentation in empirical SE and it uses the same article base as previously published reviews in this program: the topics investigated in SE experiments and information reported about the experiments [47], the level of statistical power in the experiments [11], the reporting of effect size [22], and the use of explanatory theory in the experiments [14]. In the present review, we investigate the extent of quasi-experimentation in SE, the types of quasi-experiments that are performed, how the quasi-experiments are designed and analyzed, how threats to validity are reported, and whether different results are reported for quasi-experiments and randomized experiments.

The remainder of this article is organized as follows. Section 2 presents the concepts used in this investigation. Section 3 describes the research method applied. 4 reports the results of this review. Section 5 discusses the findings and limitations of this review. Section 6 concludes.

2. Background

In this article, we use the vocabulary of experiments defined by Shadish et al. [42], Table 1. Quasi-experiments are similar to randomized experiments, apart from the fact that they lack a random assignment of study units to experimental groups (randomization).[1] In a between-subject design, there is exactly one experimen-

[1] The *random assignment* of study units to treatment conditions should not be confused with the *random selection* of study units from the study population to form the study sample, which is also referred to as *random sampling*.

Table 1
Vocabulary of experiments, from [42]

Experiment:	A study in which an intervention is deliberately introduced to observe its effects
Randomized experiment:	An experiment in which units are assigned to receive the treatment or an alternative condition by a random process, such as the toss of a coin or a table of random numbers
Quasi-experiment:	An experiment in which units are not assigned to conditions randomly

tal group for each treatment condition, and the assignment procedure then assigns each subject to exactly one treatment. In a within-subject design, the experimental units are exposed to multiple treatments, possibly in different orders, and in this case, the assignment procedure assigns each subject to one of these multiple treatment sequences. We use the following operational definition of a controlled experiment defined by Sjøberg et al. [47]:

> *A controlled experiment in software engineering is a randomized or quasi-experiment, in which individuals or teams (the study units) conduct one or more software engineering tasks for the sake of comparing different populations, processes, methods, techniques, languages or tools (the treatments).*

For simplicity, whenever we use the term "experiment" in the following, we use it in the above-mentioned sense of "controlled experiment". Moreover, the notion *to apply a treatment* will be used, even if the participant's level of SE skill also can constitute a treatment.

2.1. Methods of randomization

Several types of method for random assignment are described by Shadish et al. [42]. The two types most relevant for this study are simple random assignment (also called complete randomization) and random assignments from blocks (matches) or strata, which represent a restriction on the randomization.

In simple randomization, the participants are divided into each experimental group by a random procedure, that is; the probability of being assigned to a given group is the same for all the participants. Simple randomization does not guarantee equal experimental groups in a single experiment, but because differences are only created by chance, the various participant characteristics will be divided equally among the treatment conditions in the long run, over several experiments. In order to avoid large differences occurring by chance in a single experiment, blocking or stratifying can be used, in which study units with similar scores on the variables of interest are divided into blocks or strata and then assigned randomly to experimental groups from each block or stratum. When blocking, the participants are divided into pairs when there are two treatment conditions, into groups of three if there are three conditions, etc. Then, the study units in the pairs or groups are divided randomly to the different treatments. When stratifying, the participants are divided into strata that are larger than the number of treatment conditions, for example, one may place the 10 persons with the greatest number of years of programming experience in one stratum, and the 10 persons with the fewest number of years experience in another stratum. Then the study units in each stratum are divided randomly to treatments in such a way that equally many from each stratum are assigned to the different treatments. The use of blocks and strata in statistical analysis is described in most statistical textbooks. Determining the optimum number of blocks for a given research setting is discussed by Feldt [12] and Myers [33].

Randomization methods span from flipping a coin to using a random number computer generator. The latter procedure is recommended in guidelines for statistical methods in psychology [52], because it enables the supply of a random number seed or a starting number that other researchers can use to check the methods later.

2.2. Selection bias, the problem with quasi-experimentation

Selection bias is a threat to internal validity. It is defined by Shadish et al. [42] to be "*systematic differences over conditions in respondent characteristics that could also cause the observed effect*". When a selection is biased, treatment effects are confounded with differences in the study population. Selection bias is presumed to be pervasive in quasi-experiments. Hence, the assignment procedures used in quasi-experiments may lead to pre-experimental differences that in turn may constitute alternative causes for the observed effect. There may also be interactions between selection bias and other threats to internal validity. For example, the participants in one quasi-experimental group might drop out from the experiment (attrition) more often than participants from another experimental group, not because of the treatment, but because they have characteristics that participants in the other group do not have.

Different types of nonrandom assignment procedures might induce different types of causes for selection bias. For example, when projects are compared within a company, there is a chance that participants within projects are more alike than between projects, e.g., in terms of some types of skills that influence the performance in the experiment. Moreover, if the participants select experimental groups themselves, people with similar backgrounds might select the same group. Such differences between experimental groups might cause other differences of importance for the experimental outcome as well.

When the nonrandom assignment procedure has no known bias, it is called *haphazard assignment*. This might be a good approximation to randomization if, for example, participants are assigned to experimental groups from a sorted list on an alternating basis. However, when haphazard assignment is possible, randomization is often possible as well.

2.3. Design of quasi-experiments

Experimental designs are built from design elements, which can be categorized into four types: assignment methods, measurements, comparison groups, and scheduling of treatments. Corrin et al. [9] and Shadish et al. [42] show how quasi-experimental designs can be strengthened by adding thoughtfully chosen design elements in order to reduce the number and plausibility of internal validity threats. Among these, only two elements were observed used in the reviewed experiments: *pretest scores* and *within-subject designs*. We will here describe these two types and to further exemplify how to use design elements to improve quasi-experiments, we will also describe the use of *nonequivalent dependent variables* and *several experimental groups* (see Table 2).

A *pretest measure* is either taken from a real pretest, i.e., from a task identical to the experimental task, but without any treatment, or it is a measure that is assumed to be correlated with the dependent variable, for example, a similar task (calibration task or training task) [3], exam score, or years of experience. The two latter examples are indicators of the performance of human subjects, which include skill, abilities, knowledge, experience, etc. A challenge is to define which of these characteristics are most relevant in the given experimental setting and to find good

Table 2
Techniques for handling threats to selection bias

Techniques	Examples
Pretest scores for controlling for pre-experimental differences between experimental groups	Results from pre-treatment tasks or measures of indicators of subject performance, such as exam scores or years of experience
A *nonequivalent dependent variable* for falsifying the hypothesis of alternative explanations for observed effect or lack of effect	Time used to perform a task if the technology used can be assumed not to influence performance time
Several experimental groups for some or each treatment condition in order to allow comparison of effect of different types of groups	Each treatment condition is applied in two companies
*Within-subject design*for enabling each subject to be its own control	Cross-over design: two programming languages are compared and half the participants apply first one language and then the other. The order of language is reversed in the other group

operationalizations of those indicators. Pretest scores are used when analyzing the final results to check, or adjust for, pre-experimental differences between the experimental groups. In haphazard assignment, a pretest score can also be used in the assignment procedure (similar to blocked or stratified randomization) to prevent initial differences between the experimental groups.

The *nonequivalent dependent variable* is an additional dependent variable that is expected not to be influenced by the treatments and is used to falsify the hypothesis of alternative explanations for treatment effects or lack of effect. For example, when the outcome is measured in terms of answers to a questionnaire, the nonequivalent dependent variables are questions, the answers to which are assumed not to be influenced by the treatment, but are related to the participants' performance. If the answers from the outcome differ among the experimental groups, whereas the answers from the nonequivalent dependent variables do *not* differ among the groups, the belief that there are no other explanations for the results than the effect of the treatment is strengthened. If both the outcome and the nonequivalent variables differ among the experimental groups, there is an indication that treatment effects might be confounded with group effect. See [45] for an example of use of this kind of nonequivalent dependent variable.

Applying *several experimental groups* allows control of how the quasi-experimental groups influence the results. If the same result is observed for several experimental groups using the same treatment, it confirms the belief that the result is due to treatment and not group characteristics. This is a kind of replication within the single experiment.

The *within-subject design* is a method for compensating for initial experimental group differences, because each subject or team serves as its own control. A challenge with within-subject designs is the possibility that the effect of applying a treatment in one period might influence the use of another treatment in the following period. Such a period effect is confounded with treatment effect in a within-subject design if all the participants apply all the treatments in the same order. However, a common within-subject design is the crossover design (also called counter-balanced design), where different sequences of treatments are applied for different groups of participants. The crossover design caters for period effects, but inferential problems still arise if there is a nonnegligible interaction between period and treatment. For example, when there is a difference in learning effects between the treatments. In fact, the most basic $2 * 2$ crossover design is particularly vulnerable to this inferential problem, because it does not allow a proper estimation of the presence of the period-by-treatment interaction effect [20]. Kitchenham et al. [25] demonstrate the hazards in using a crossover design in cases where there is a period by treatment interaction. They recommend using crossover designs only if the researchers are sure, before undertaking the experiment, that there is no period by treatment interaction or that the period by treatment interaction is smaller than the treatment effect.

Strategies for ruling out threats to selection bias are also presented by Reichardt [37]. These strategies mainly involve hypothesis formulations and constructions of comparison groups and are called relabelling, substitution, and elaboration:

- *Relabelling* means that the researcher rephrases the research question or hypothesis of the treatment effect to include the joint effect from treatment and the effect of the selection differences among the groups. The relabelling method can always be applied, but is probably the least desirable method to use because the hypothesis of joint effect is often not as interesting to investigate as the treatment effect alone. An example is a quasi-experiment that aims at comparing software engineering method *A* and *B*, and which includes students from two classes that are being taught the respective methods. The comparison of the methods will be influenced by the background of the students and the quality of the teaching. The original research question can be rephrased to include these additional effects: What is the joint effect of the method, students' curricula profile and quality of teaching?
- *Substitution* implies that the comparison is substituted by another comparison or by a pair of other comparisons to control for possible threats. For example, instead of making one comparison in which the selection threat is difficult to rule out, a pair of comparisons is made, in which one is constructed in such a way that the threat is expected to have a positive effect, and the other one in such a way that the threat is expected to have a negative effect. If the results of both comparisons are in the same direction, then the researcher can conclude that the threat has been taken into account. Another example: if the intended use of a quasi-experiment appears to lead to a too large selection bias to be justifiable, a randomized experiment should be chosen instead, even if this would be more difficult to implement.
- *Elaboration*can be described as the "opposite" of substitution. The researcher retains the original comparison for which the selection threats are difficult to rule out and adds other comparisons, for example, by measuring a nonequivalent dependent variable or using several comparison groups, as described in Table 2. The additional comparisons allow the researcher to disentangle the relative contributions of the treatment and of the threat to validity as explanations for the results in the original comparison. Elaborations can take several forms. One is to show that the size of the selection effect is zero. For example, if a threat to the validity of a comparison of two methods is a potential difference in a particular type of skill in the two experimental groups, a comparison between the experimental groups that tests this particular skill is constructed. If this comparison shows no difference, then the threat is ruled out.

2.4. Analysis of quasi-experiments

Cook and Campbell [8] give the following general advice when analysing quasi-experiments: (1) plan the design carefully, so as to have available as much information that is required for the analysis as possible, (2) use multiple and open-minded analyses, and (3)

use an explicit appraisal of the validity of the findings and the plausibility of alternative explanations.

An open-minded analysis means to be prepared to not necessarily use standard procedures for analysis. An example is an investigation of two methods for software cost estimation accuracy [13]. Nineteen projects were used and each project self selected which estimation method to apply. The researchers observed that project characteristics (based on pretests scores) seemed to overrule the effect of the estimation method. Hence, they analysed the projects within blocks of similar projects. Note that researchers need to be careful about performing post hoc analysis in order to identify subsets of the data for further analyses, because this can degenerate into fishing for results. The possible threats to selection bias should, as far as possible, be identified before the data gathering begins and lead to planned strategies for dealing with the threats.

A pretest score may be applied in the analysis of continuous outcomes either (i) in an analysis of pretest–posttest differences (gain-score), (ii) by creating blocks or strata (retrospectively) within each experimental group on the basis of the pretest scores and including the blocking variable in the analysis (ANOVA with blocking or stratifying), or (iii) by applying the pretest as a covariate in the analysis (ANCOVA) [8]. These methods are described and compared by Cook and Campbell [8]. Among other things, a convincing illustration of how the use of a simple ANOVA yields an incorrect inference compared with using ANCOVA when the experimental groups differ at pretest. An example of the use of ANCOVA is reported by Arisholm et al. [3]. In that study, a calibration task was used to measure pretest scores (applied as a covariate in an ANCOVA), which affected the overall conclusion. Further improvement to an ANCOVA by making a reliability adjustment is suggested by Trochim [50].

Scepticism regarding the use of traditional statistical methods, such as ANCOVA, to adjust for selection bias in experimentation of the effect of social intervention is discussed by Lipsey and Cordray [29]. The principal problem is the sensitivity of the results to the violation of model assumptions for such methods, especially the requirement that all relevant variables be specified. Lipsey and Cordray recommend qualitative methods as part of experimental evaluations, as well as incorporating additional variables into experimental and quasi-experimental designs. They state that such variables can be fully integrated into analyses of change by applying new statistical techniques (see [29] for details).

The use of Bayesian statistics is suggested by Novich [34]. He argues that statistical analyses involve much more than textbook tests of hypotheses and suggests applying Bayesian statistics because this method allows background information to be incorporated into the analysis. However, according to Rubin [39], sensitivity to inference of the assignment mechanism in nonrandomized studies is the dominant issue, and this cannot be avoided simply by changing the modes of inference to Bayesian methods.

3. Research method

This section describes how the experiments and tests reviewed in this article were identified and how the data was gathered.

3.1. Identification of experiments

The 103 papers on experiments (of a total of 5453 papers), identified by Sjøberg et al. [47], are assessed in this review. Table 3 shows the actual journals and conference proceedings, which were chosen because they were considered to be representative of empirical SE research. The 103 articles reported 113 experiments. The process for selecting articles was determined from predefined criteria, as suggested by Kitchenham [23]; see [47] for details. The list of articles is presented in [21].

Table 3
Distribution of articles describing controlled experiments in the period January 1993–December 2002

Journal/Conference proceeding[a]	Number	%
Journal of Systems and Software (JSS)	24	23.3
Empirical Software Engineering (EMSE)	22	21.4
IEEE Transactions on Software Engineering (TSE)	17	16.5
International Conference on Software Engineering (ICSE)	12	11.7
IEEE International Symposium on Software Metrics (METRICS)	10	9.7
Information and Software Technology (IST)	8	7.8
IEEE Software	4	3.9
IEEE International Symposium on Empirical Software Engineering (ISESE)	3	2.9
Software Maintenance and Evolution (SME)	2	1.9
ACM Transactions on Software Engineering (TOSEM)	1	1.0
Software: Practice and Experience (SP& E)	–	–
IEEE Computer	–	–
Total	103	100

[a] The conference *Empirical Assessment & Evaluation in Software Engineering* (EASE) is partially included, in that 10 selected articles from EASE appear in special issues of JSS, EMSE, and IST.

3.2. Information extracted

Each of the 113 experiments was categorized as *randomized experiment*, *quasi-experiment* or *unknown* with respect to the assignment procedure. Since one experiment could comprise several tests for which some were exposed to randomization and some were not, we based our categorization on the primary tests when these could be identified. In total, 429 primary tests were identified in 92 experiments in a multi-review process; see [11] for details. We defined the primary tests to be what the experiments were designed to evaluate, as indicated in the descriptions of the hypotheses or research questions. If no hypothesis or research question was stated, we classified as *primary* those tests that were described to address the main incentive of the investigation. *Secondary tests* comprised all other tests.

The assignment procedure was not always described clearly in the articles. An experiment was categorized as randomized if it was stated explicitly that randomization was used for all the primary tests. An experiment was categorized as a quasi-experiment when a nonrandom procedure was reported explicitly for at least one primary test and when the experimental design or the experimental conduct was such that randomization was obviously impossible for at least one primary test. In other cases, the experiment was categorized as unknown. An e-mail request was sent to the authors of the 27 experiments with an unknown assignment procedure. Answers were received for 20 experiments, for which eight apparently employed randomization and are categorized as such in this review.

In 14 of the experiments, no statistical testing was performed. In seven experiments, it was impossible to track which result answered which hypothesis or research question. For these experiments, no primary tests were identified and hence, the assignment procedure was determined from the description of assignment to the experimental groups. When teams were used as the study unit, we regarded the assignment procedure to be the assignment of teams to experimental groups. We regarded the forming of the teams as being part of the sampling procedure.

In addition to the categorization of each experiment as *randomized experiment*, *quasi-experiment* or *unknown* with respect to the assignment procedure, the following attributes were registered per experiment:

- study unit
- assignment method for randomized experiments and assignment procedure for quasi-experiments

- whether techniques for ruling out threats to selection bias was used for at least one primary test
- whether the quasi-experiments with a within-subject design addressed the potential challenges with period effect and period by treatment interaction effect
- whether internal validity was addressed for at least one primary test
- whether threats to selection were reported for at least one primary test
- whether professionals were used as study unit
- whether commercial applications were used
- standardized mean difference effect size for each primary test

Regarding the last five bulleted points, data on internal validity, threats to selection, the use of professionals, and the use of commercial applications were gathered by Sjøberg et al. [47] and effect size was estimated by By Kampenes et al. [22]. This data is presented separately for quasi-experiments and randomized experiments in this article.

Although attributes for data collection should ideally be determined prior to a review [23], our experience is that the determination of which attributes to use and their appropriate wording often needs revision during data collection. We therefore conducted a dual-reviewer (by the first and third authors) pilot on approximately 30% of the articles in order to stabilize (1) the comprehension of description of study unit and experimental design and (2) the categorization of each experiment as *randomized experiment*, *quasi-experiment* or *unknown.*

4. Results

This section presents the extent of randomization observed in the reviewed experiments and how the quasi-experiments were designed and analyzed compared with randomized experiments.

4.1. Extent of quasi-experiments

Of the 113 surveyed SE experiments, 40 (35%) were quasi-experiments (Table 4), although the term "quasi-experiment" was used for only four experiments. There were 66 (58%) randomized experiments. For seven experiments, randomization or non-randomization was neither explicitly stated nor obvious from the experimental design and clarifications were not obtained from correspondence by email. Examples of phrases from these seven articles are: "subjects were divided into two groups" and "subjects were assigned to groups A and B so that both had subjects of equal ability". For seven experiments, randomization was performed for some of the tests or to some of the experimental groups, but not completely. We categorized these as quasi-experiments. Only three experiments described the randomization method applied: drawing a letter from a hat, drawing a number from a hat, and drawing lots.

4.2. Design of quasi-experiments

The only techniques for handling threats to selection bias that we observed in the review experiments were the use of pre-test scores and within-subject designs. In this section, we present the design of quasi-experiments in terms of the extent of use of pretest scores, which assignment procedures that were used (including the extent of within-subject designs), the extent of field experiments, and the use of teams as the study unit.

4.2.1. The use of pretest scores

Only 45% of the quasi-experiments applied a pretest measure (Table 4). This was slightly more than for the randomized experiments. The majority of the pretest measures were applied in the assignment procedure (in 13 of 18 quasi-experiments and in 18 of 26 randomized experiments (blocked or stratified randomization)). The pretest scores were mainly skill indicators, such as exam scores, years of experience, or number of lines of code written. However, for three experiments, a pre-treatment task was performed and a real pretest score measured. Two of these experiments collected data through a questionnaire that was completed by the participants both before and after the treatment was applied. For one experiment, which investigated the effect of using design patterns, SE maintenance tasks were performed both before and after the participants attended a course in design patterns.

4.2.2. The assignment procedures

We found four main types of nonrandom assignment procedures. The number and characteristics of these types are shown in Table 5.

1. *Assignment to nonequivalent experimental groups.* There were four types of nonequivalent group designs:
   (a) Five experiments were designed to investigate the effect of indicators of subject performance, such as experience and skill. The experimental groups were formed to be unequal regarding these indicators. The groups were also nonequivalent with respect to other types of experience or skill, due to the nonrandom assignment procedure. Subjects were assigned on the basis of either questionnaire results or the sampling of subjects from different populations.
   (b) For one of the experiments, subjects were assigned to experimental groups by including subjects with specific knowledge of the technology (treatment) used.
   (c) Three experiments included subjects from different classes, projects, or universities.
   (d) Six experiments assigned participants to experimental groups on the basis of their availability.
2. *Haphazard assignment.* Four experiments applied a pretest-based formula or procedure in the assignment, which was not formally random but seemed to be a good approximation; for example, assignment on an alternating basis from a ranked list of examination scores. For eight experiments, a more judgmental approach was used to assign participants to experimental

Table 4
The extent of randomization and use of pre-test

Type of experiment	Total number of experiments		Use of pretest scores							
			Total		In assignment		In descriptive analysis		In statistical analysis	
	N	%	*N*	%[a]	*N*	%[a]	*N*	%[a]	*N*	%[a]
Quasi-experiments	40	35.4	18	45.0	13[b]	32.5	3	7.5	2	5.0
Randomized experiments	66	58.4	26	39.4	18	27.3	8	12.1	3	4.5
Unknown	7	6.2	3	42.9	3	42.9	0	0	0	0
Total	113	100	47	41.6	34	30.1	11	9.7	5	4.4

[a] Percentage of the total number of experiments for that particular type of experiment.
[b] In addition to the twelve experiments using a pretest based assignment, one experiment, categorized as some randomization, used blocked randomization.

Table 5
Quasi-experiments detected in this review (number of experiments)

(1) Nonequivalent experimental groups (15)
(a) Investigation of skill, experience, etc. as treatment (5)
- Assignment, for already included participants, based on answers to a questionnaire (2) (C++ experience, Database knowledge)
- Inclusion of subjects from different skill populations (3) (Students versus professional, Programming knowledge, Personal software process knowledge)

(b) Assignment based on knowledge of the technology (1)
- Subjects with knowledge of formal methods versus those without such knowledge were used in a comparison of formal methods versus no formal analysis

(c) Experimental groups created from similar groups (classes or projects) at different times (3)
- Student classes from two succeeding years were used as experimental groups (2)
- Development courses at a company from two succeeding years were used as experimental groups (1)

(d) A natural assemblage of participants into experimental groups (6)
- Two sections of a student class were used as experimental groups(2)
- Availability and schedule played a role in the assignment of subjects to experimental groups (4)

(2) Haphazard assignment (12)
(a) Formula-based (4)
Assignment method:
- On an alternating basis from a ranked list of previous marks (2)
- An algorithm was used on a ranked list of previous marks (2)

(b) Assignment based on the researcher's subjective judgement (8)
The judgement was based on:
- Knowledge of the subjects' skills (1)
- Background information collected from the subjects (2)
- Combination of experiences with the subjects' skills and background information (3)
- Grade point average (2)

(3) Some randomization (7)
(a) Randomization and nonequivalent group design (4)
- Experimental groups created partly from different physical locations (1) In a three-group experiment, one experimental group was selected from one university, while the two others were selected from a different university and assigned randomly to two groups
- Assignment based partly on knowledge of the technology (1) In a three-group experiment, one experimental group was formed by subjects who already understood the component before assignment, while other subjects were assigned randomly to the two other groups in a study of reusable components
- Randomization and skill assessment in a factorial design (2)

(b) Randomization for individuals, but not for teams, both being study units (1)
(c) Randomization for three experimental groups (1). A fourth group was created by using the participants from one of the other groups
(d) Randomization for two experimental groups (1). Some primary tests compared the pre- and post-treatment scores within the groups, i.e. a nonrandomized comparison

(4) Within-subject experiments in which all participants applied the treatment conditions in the same order (6)
(a) In an inspection experiment, first the usual technique was applied; then the participants underwent training in a new technique followed by applying the new technique in the experiment (3)
(b) In an assessment of the effectiveness of inspection team meetings, individual results were compared with team results, individual inspection being performed first by all participants (1)
(c) All participants first performed a paper-based inspection, followed by using a web tool (1)
(d) All participants applied estimation methods in the same order (1)

groups, based on pretest scores and previous knowledge about the participants. For 11 of the 12 experiments that used haphazard assignment, the assignment procedure was not described clearly in the article but information was obtained through mail communication.

3. *Some randomization.* For seven of the experiments, randomization was performed for some, but not all, of the experimental groups or the primary tests. Hence, a nonrandom assignment procedure was used as well.
4. *Within-subject experiments in which all participants apply the treatment conditions in the same order.* For six experiments, all the participants were assigned to the same experimental groups, applying both technologies in the same order.

Assignment to *nonequivalent experimental groups*, *haphazard assignment* and *some randomization* were applied for both between-subject designs and cross-over designs for quasi-experiments, see Fig. 1.

Within-subject design is regarded as one way of reducing selection bias when applying a nonrandom assignment procedure. Still, the extent of within-subject designs was smaller for the quasi-experiments than for the randomized ones (35% versus 51%). Four quasi-experiments used a mix of between- and within-subject design for the primary variables. Among the crossover experiments, all the randomized experiments and all, but one, quasi-experiment compared two treatments. One of the quasi-experiments had a $3 * 3$ crossover design. The randomized experiments with *other within-subject designs*compared more than two treatment conditions and scheduled the treatments in such a way that not all possible sequences were applied. An example of such a design is a (3 treatment * 2 period) within-subject design, where one group of participants used the same treatment in both periods, and the two other groups of participants switched treatments in the second period. Note that the analysis of such nonstandard designs is not straightforward. Hence, such designs have validity challenges that are not solved by randomization.

4.2.3. The field experiments

The percentage of experiments applying professionals as the study unit was roughly equal for quasi-experiments and randomized experiments (20% versus 18%; see Table 6). Commercial applications were used in 13% of the experiments, more in randomized experiments. However, the professionals worked with commercial applications in five of the quasi-experiments (13%) and in four of the randomized experiments (6%). Hence, on the basis of type of study unit and application, a greater industrial focus was seen for quasi-experiments than for randomized experiments. In addition, the quasi-experiments had slightly larger sample sizes than the randomized experiments; see Table 6.

4.2.4. The use of teams

SE tasks are often performed in teams, and the team was the study unit in 26% of the experiments, more often in quasi-experiments (40%) than in randomized experiments (17%); see Table 6.

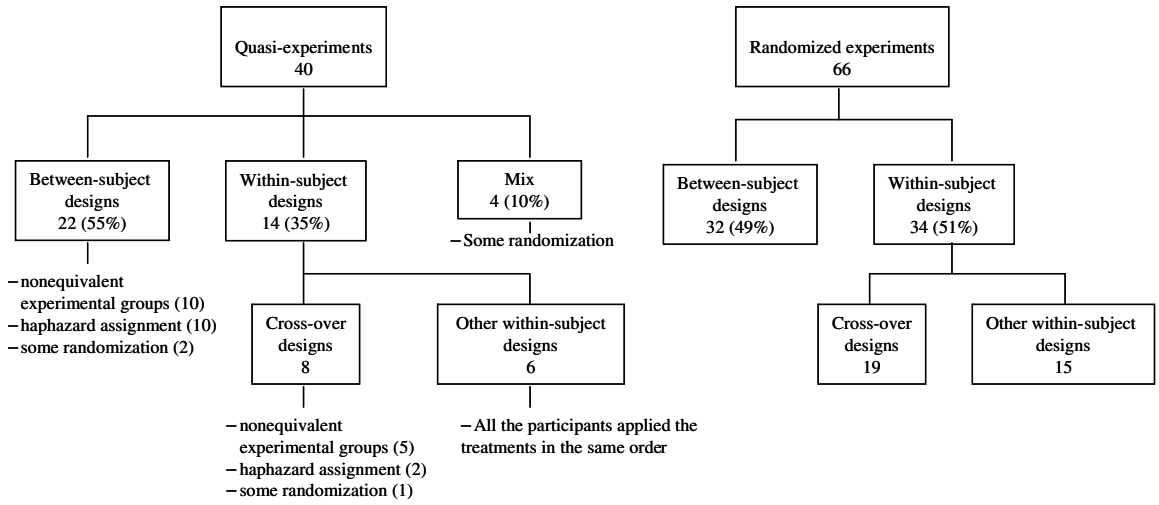

Fig. 1. Experimental designs detected in this review.

Table 6
Number of randomized and quasi-experiments in the reviewed articles, by type of study unit

Type of experiment	Total	Median sample size[c]	Professionals as study unit[a]		Commercial applications[b]		Teams as study unit	
			N	%	*N*	%	*N*	%
Quasi-experiments	40	42.0	8	20.0	5	12.5	16	40.0
Randomized experiments	66	34.5	12	18.2	10	15.2	11	16.7
Unknown	7	13.5	1	14.3	0	0	2	28.6
Total	113	36.0	21	18.6	15	13.3	29	25.7

[a] Students only were used in 82 experiments and a mix of subjects in nine.
[b] Other types of applications in the experiments were constructed applications (81), student applications (5), unclear (9) and other (3).
[c] Based on the comparison with the largest number of data-points per experiment for the 92 experiments in which this was reported.

For eight of the 16 quasi-experiments with teams, the teams were reported as having been formed as follows: by random assignment (4), by random assignment within experimental groups (1), by the participants themselves (2), or on the basis of the researcher's judgment for creating equal teams based on the participants' C++ marks (1). For eight of the 16 cases, the method was not reported. In all the eleven randomized experiments with teams, the teams were formed by assigning individuals by a random process.

A pretest score was used for 36 of the 84 (43%) experiments using individuals and for 11 of the 29 (38%) experiments using teams. For all these experiments, the pretest was a measure of the individual skill level, not of the overall team level.

One experiment reported that cost and time were constraints that hindered the use of teams, even if teams would have been a more realistic study unit than individuals for that particular experiment.

4.3. Analysis of quasi-experiments

Only two of the 40 quasi-experiments (5%) applied a pretest score in the *statistical* analysis of results in order to adjust for pre-experimental differences in the participants' characteristics and only three (8%) of the quasi-experiments compared pretest scores in a *descriptive* analysis, see Table 4. In the randomized experiments, adjusting for pre-experimental differences happening by chance, 3 (5%) experiments applied a pretest score in the *statistical* analysis and 8 (12%) experiments applied such a score in the *descriptive* analysis.

The sparse use of pretest scores is one indication that researchers are, in general, unaware of the potential selection bias in quasi-experiments and how the problem can be handled in the analysis of the results. Another indication of this is that internal validity issues were discussed to a lesser extent for quasi-experiments than for randomized experiments (60% versus 70%); see Table 7, i.e., it is addressed less where it is needed more. Moreover, in most cases when internal validity was addressed, no threat was claimed to be present. The presence of at least one threat was reported to an equal extent for quasi- and randomized experiments. Threats to selection bias were reported for only three of the quasi-experiments. There seems to be some confusion regarding the term *selection bias*, because among the randomized experiments, 11% reported threats to selection bias, probably referring to differences that occurred by chance. In addition, it seems as though some experimenters referred to selection bias when they meant lack of sampling representativeness.

The effect of the assignment procedure is reduced in within-subject designs, because the participants apply several treatment conditions. However, in within-subject designs a potential period effect might confound or interact with treatment. For six of the reviewed quasi-experiments with within-subject design, treatment was completely confounded with period, i.e. the effect of treatment could not be separated from the effect of order of treatment. One of these experiments argued that the interpretation of results had to take this aspect into account. Another experiment argued that there would be no learning effect of importance. Among the eight quasi-experiments that used a crossover design, five of the experiments addressed the issue of the potential period-by-treatment interaction. Two experiments tested and found an interaction effect and three experiments argued that a potential interaction effect could be ignored.

We attempted to measure whether selection bias influenced the results from the quasi-experiments in this review. There was sufficient information for the effect size to be estimated for 284

Table 7
Threats to internal validity, as reported in the surveyed experiments

Type of experiment	Total	Internal validity awareness		At least one internal validity threat present		Threats to selection bias present	
		N	%	*N*	%	*N*	%
Quasi-experiments	40	24	60.0	10	25.0	3	7.5
Randomized experiments	66	46	69.7	16	24.4	7	10.6
Unknown	7	1	14.3	0	0	0	0
Total	113	71	62.8	26	23.0	10	8.8

Note. N is the number of experiments.

Table 8
Experimental results in terms of standardized mean difference effect size

Assignment Procedure	Experimental design	Effect size results from the primary tests				Number of experiments
		Mean	Median	Std	Number of tests	
Nonrandom	Between-subject design	0.53	0.39	0.50	31	11
	Cross-over design	0.83	0.81	0.50	19	6
	Same order of treatments	0.51	0.38	0.51	26	6
	Total nonrandom assignment	0.61	0.50	0.52	76	23
Random	Between-subject	0.83	0.69	0.69	104	24
	Cross-over design	0.99	0.63	0.91	31	12
	Other within-subject designs	0.87	0.77	0.69	61	8
	Total random assignment	0.86	0.68	0.73	196	44
Unknown		1.25	1.32	0.85	12	3
Overall		0.81	0.60	0.69	284	70*

* Some experiments had primary tests in several assignment categories. A total of 64 unique experiments were represented in this table.

Table 9
Proportion of quasi-experiments

Study	Inclusion criteria	No of experiments	Quasi-experiments	
			N	%
Meta-analysis of psychology studies [43]	◦Published reports in *Psychological Abstracts* 1975–1979 ◦At least three comparison groups ◦Between-subject design ◦Information for effect size estimation available	143	–	10
Review of methods in clinical trials [10]	◦Comparative clinical trials published in one of four medical journals in July–December 1979	67	–	16
Review of controlled clinical trials within surgery [32]	◦Published controlled clinical trials in six medical journals in 1983 ◦Minimum total sample size: 10 (five for cross-over studies)	96	15	16
Review of controlled clinical trials of acute myocardial infarction [6]	◦Studies published in 1946–1981reporting on a comparison of a treatment to a control	145	43	30
Review of controlled clinical trials within medicine [7]	◦Published controlled clinical trials in a sample of medical journals in 1980 ◦Minimum total sample size: 10 (five for cross-over studies)	114	49	43
Review of experiments in criminology [51]	◦All available comparative studies within seven areas of criminal justice	204	158	77
Meta-analysis of experiments within school-based prevention of problem behavioural [54]	◦All available reported comparisons from published in journals (80%), other publications (10%) and unpublished reports (10%) ◦165 studies included, the results reported on comparison level, not study level	216	174	81
This study	◦Controlled experiments within SE published in nine journals and three conference proceedings in 1993–2002	113	40	35

primary tests in 64 experiments; see [22] for details. None of these experiments adjusted the results by pretest scores to control for selection bias. Overall, the randomized experiments had higher average and median effect sizes than had the quasi-experiments; see Table 8. However, the result was ambiguous across types of design; the quasi-experimental cross-over designs had effect size values in the same range as the randomized experiments.

5. Discussion

5.1. Extent of quasi-experimentation

Compared with the extent of quasi-experiments observed in other research areas (range 10–81%), SE places itself in the middle (35%), see Table 9. Fewer quasi-experiments than randomized ones are conducted in research on medical science and psychology, whereas in experimental criminology, more quasi-experiments than randomized ones are conducted.[2]

Guidelines and textbooks on research in medical science and psychology typically favour randomized experiments for cause–effect investigations, because of their potential to control for bias [2,19,36,52]. This might explain the relatively large extent of randomization in these areas of research. In addition, especially in medical research, randomization is made possible by patients easily enrolling themselves to randomization procedures at hospitals, health care centres and medical doctors.

In contrast, sparse use of randomized experiments is reported in criminology. Many kinds of intervention pertaining to criminal justice do not lend themselves readily to randomized designs [27], because practical, ethical, financial and scientific factors play a role

[2] For simplicity, we use the terms "quasi-experiments" and "randomized experiments" even if these terms are not always used in other research areas for comparative studies (trials) that use nonrandom and random assignment procedures.

[44]. Hence, it seems that experiments in criminology have mostly been performed in field settings, where randomization is not feasible.

In SE, even if 35% of the experiments were quasi-experiments, only 13% (five) of them were field experiments in the sense that the subjects were professionals working with commercial systems. So, most of the quasi-experimentation in SE consists of research other than field experiments, even though the running of field experiments is regarded as the main incentive for running quasi-experiments in SE [24,26]. The sparse use of field experiments may be explained by practical constraints, such as costs for the industry, and methodological challenges, such as the level of experimental control that can be achieved in a practical setting [26]. Whereas these constraints seem to lead to a large amount of quasi-experiments being conducted in criminology, the same constraints seem to lead SE researchers to use students as subjects and run randomized experiments rather than quasi-experiments.

In addition to its use in field experiments, we observed the use of quasi-experimental design in the following: investigations of how subject-performance indicators influence the results; comparisons of students from different classes, years, universities, or with treatment-specific knowledge; investigations that make assignments on the basis of the participant's availability; investigations of both teams and individuals for which randomization for both are difficult; within-subject designs for which all participants apply all treatments once and in the same order; and quasi-experiments using haphazard assignment. Except for haphazard assignment, these quasi-experiments represent settings for which randomization is not feasible, but where participants are available and the investigation of cause–effect relationships is possible through a quasi-experimental design. For experiments that use haphazard assignment, blocked or stratified randomization would probably have been possible instead. The use of blocked or stratified randomization for these experiments would have reduced the extent of quasi-experiments from 39% to 23%.

5.2. Results from quasi-experiments compared with randomized experiments

We found that, on average, effect sizes were larger for the randomized experiments than for the quasi-experiments. This might indicate that selection bias in the quasi-experiments influenced the results. There is probably no single explanation for the observed direction of difference. Selection bias in one nonrandomized comparison might be offset by an opposite bias in another such comparison. Hence, it might act more as a random error than a systematic bias that is due to a cause. This will reduce the confidence in the findings, but effect sizes will be consistently neither over- nor underestimated [53]. Moreover, other types of biases might have influenced the results, for example, the potential period effects or the period-by-treatment interaction effects in the within-subject designs. Also, the nonstandard, within-subject designs observed among the randomized experiments might have resulted in biases that influenced the results for this group of experiments. The small number of quasi-experiments in our review also gives us reason to view with caution the observed differences in effect sizes from randomized experiments and quasi-experiments. Nevertheless, we should take note of the results, because the hypothesis that selection bias might influence the results from quasi-experiments has a theoretical foundation and is also empirically supported in other research fields. Meta-analyses in psychology, medical research, cognitive behavioural research and criminology found treatment differences partly in favour of randomized experiments [43,48,54], partly in favour of quasi-experiments [6,7,32,51], and some found no difference [28,35]. In these investigations, the observed differences were all explained by the potential bias in the quasi-experiments.

The theory of quasi-experimentation suggests how to control for selection bias. Researchers have attempted to assess these suggested precautions empirically. Researchers in psychology have found that by avoiding self-selection of experimental groups as the assignment method and/or adjusting for pre-experimental differences, selection bias could be eliminated completely [1], or at least to some extent [16,17,30,41] by using a pretest score. We did not have sufficient data to evaluate properly any techniques for handling selection bias. Our aim with presenting the comparison in results between the randomized experiments and the quasi-experiments was to illustrate that biases might influence the results. Further investigations are required to find out how and when, in general, these biases occur. However, this review shows that biases can be reduced in ESE experiments by using the current knowledge.

5.3. Indicators of subject performance

Pretest scores are useful for controlling and adjusting for undesirable pre-experimental differences between experimental groups. Among the 49 experiments that measured a pretest score, subject-performance indicators (measured as exam score, years of experience, and number of lines of code written) were used in all but three experiments. This shows that subject-performance indicators are much more commonly used as pretest scores than measures from real pretest tasks.

Nevertheless, over half of the quasi-experiments did not apply a pretest score to control for selection bias. We believe that even if this is partly due to lack of awareness of its importance, it is also partly due to the fact that a relevant subject-performance indicator score is often difficult to measure. Hence, we conclude that the SE community needs to conduct more research on how to measure different concepts such as skill, ability, knowledge, experience, motivation, etc. and how these concepts interact with different types of technologies [46]. In our review, all the investigations of subject-performance indicators were quasi-experimental. We believe that including participants with certain skills in a quasi-experiment is often more relevant than teaching some kind of knowledge as part of a randomized experiment.

5.4. Quality of reporting

There was incomplete reporting of several of the variables that were investigated in this review: type and rationale for assignment procedure, randomization method, threats to internal validity, and information used for effect size estimation. In our experience, this makes it difficult both to understand and evaluate experiments and to conduct systematic reviews and meta-analyses. For 1/4 of the experiments, the assignment procedure was not described in the articles. Only three of the randomized experiments reported the randomization method. Sparse reporting of the method is also found in medical research; in four studies on clinical trials, the randomization method was reported in, respectively, 0.8%, 4%, 19% and 51% [10,18,35,49].

Even though some of the articles in our review provided excellent descriptions of experimental design issues, in general, justification for the choice of assignment method was lacking. Moreover, internal validity was addressed in only 55% of the experiments and there was sufficient descriptive information for effect size to be estimated for only 64 of the 92 experiments that reported significance testing; see [22] for details.

5.5. Ways to improve quasi-experimental designs in SE

We detected four main types of quasi-experiment. We will here suggest how these experimental designs could be strengthened by using the design elements described in Section 2.

5.5.1. Nonequivalent experimental group designs

The main question to ask when the experimental groups are nonequivalent is: which factors could cause these groups to differ before treatment is administered? The answer depends on the assignment procedure. We observed four types of assignment procedures for nonequivalent group designs; see Table 5 (1a–d).

(a) When investigating skill, the experimental groups differ deliberately regarding this skill. In addition, the groups might differ with respect to other relevant types of skills or with respect to other factors that differ between the populations for which the participants are sampled. The ways of controlling this are to (1) use pretest measures, for example examination score from a common course that concern types of skills other than treatment skill, (2) nonequivalent dependent variables that are assumed not to be influenced by the treatment skill, and (3) several comparison groups that differ with regards to other factors that may influence the results. If possible, we will recommend including participants from different populations because this enables a balanced design. The alternative, which we do not recommend, is to divide already included participants into skill groups on the basis of, for example, a questionnaire.
(b) The same recommendations as above apply for quasi-experiments that include subjects with knowledge of the technology under investigation, i.e., participants with different knowledge in the different experimental groups. The experimental groups might differ with respect to skills other than knowledge of the particular technology. This potential difference must be controlled.
(c–d) When the experimental groups are formed from different student classes, projects or universities, and when participants are included in experimental groups distant in time, or based on availability, the potential factors that could cause the groups to differ are to be found in the characteristics of the groups from which the participants are sampled. Do the students from the different courses have the same curriculum history? Do the project participants have the same amount of experience? What is the reason for their availability at certain time points? Mainly pretest measures and nonequivalent dependent variables are used to control for differences between the experimental groups. However, within-subject design and several comparison groups are also useful if the experimental constraints allow it.

5.5.2. Haphazard assignment

Haphazard assignment might be a good approximation to randomization, especially when the assignment procedure is formula based, which is the case for two of the reviewed experiments. However, little is known about the consequences of haphazard assignment, whereas the statistical consequences of randomization procedures have been well researched [42]. In addition, haphazard assignment that is based on the researcher's subjective judgment, which was seen in eight of the experiments, is difficult to report and recheck. The haphazard assignment procedures observed in the reviewed experiments all used a pretest score in the assignment. In general, we recommend using blocked randomization for such experiments.

5.5.3. Some randomization

For seven of the experiments, the design was partly randomized and partly quasi-experimental. Our recommendation for such experiments is to make this mix of design explicit in the article and control threats to selection bias in the quasi-experimental part of the experiment. Ways of controlling threats to selection bias depend on the actual nonrandom assignment method; see Section 5.5.1.

5.5.4. Within-subject design in which all participants apply the treatments in the same order

When the treatments are applied only once, this is a weak quasi-experimental design, because it does not allow proper control of how learning effects may influence the second technology. Still, it was used in six of the reviewed experiments. One explanation given was that the assumed larger learning effect from one of the technologies prevented a cross-over design and that there were too few participants available to achieve sufficient power in a between-subject design. We recommend avoiding such designs and rather using a between-subject design that is analyzed by confidence intervals and effect size measures, thus avoiding the power problem.

5.6. Limitations of this review

Limitations regarding the selection of articles and tests are described in, respectively, [47] and [11]. An additional threat regarding the set of selected articles is that there is a risk that the findings are obsolete; the articles selected are from 5 to 14 years old.

Another threat to this review is possible inaccuracy in data extraction. The data was extracted by one person (the first author). However, we conducted a dual-reviewer pilot (by the first and third authors) on approximately 30% of the articles in order to stabilize such attributes as study unit, experimental design and the categorization of randomized experiment and quasi-experiments, prior to the full review. Moreover, data for the attributes that were perceived to be potential sources of inaccuracy were also checked by the third author. No disagreements were found.

Effect sizes were not calculated for all the tests, due to the lack of sufficient information reported in the articles. In addition, there were few experiments in each quasi-experimental group. These are limitations to the comparison of effect size values between quasi-experiments and randomized experiments. Another limitation to this comparison is that the experiments differ in respects other than the assignment procedure, for example, methodological quality, topic of investigation, and type of outcome measured.

6. Conclusion

The purpose of this systematic review of the literature was to investigate the extent of randomization and quasi-experimentation in SE, how the quasi-experiments were designed and analyzed, how threats to validity were reported, and whether different results were reported for quasi-experiments and randomized experiments.

One third of all the experiments investigated were quasi-experiments. Of these, four main types were observed: (1) nonequivalent experimental group designs, (2) experiments using haphazard assignments, (3) experiments using some random and some nonrandom methods of assignment, and (4) experiments in which all participants were assigned to the same experimental groups in a within-subject design.

Reports of threats to selection bias were conspicuous by their absence. Pretest scores were measured in nearly half of the quasi-experiments and cross-over designs were used in eight quasi-experiments. Still, for nearly half the quasi-experiments, no effort to handle selection bias was reported. Overall, the randomized experiments had higher average and median effect sizes than had the quasi-experiments. This result is based on few quasi-experiments, but is in line with quasi-experimental theory and findings in other fields of research: quasi-experiments might lead to results other than those of randomized experiments unless they are well designed and analyzed to control for selection bias.

To conclude, there seems to be little awareness of how to design and analyze quasi-experiments in SE to obtain valid inferences, for

example, by carefully controlling for selection bias. Nevertheless, several of the reviewed quasi-experiments were very well performed and reported, and contributed to the recommendations given in this article on how to improve the general conducting of quasi-experiments. We hope that this article will contribute to an increased understanding of when quasi-experiments in SE are useful and an increased awareness of how to design and analyse such experiments.

Acknowledgements

We thank the Research Council of Norway for financing this work through the INCO project; the editor, Barbara Kitchenham, and the anonymous reviewers for valuable comments; Gunnar Bergersen for useful discussions; and Chris Wright for proofreading.

References

[1] L.S. Aiken, S.G. West, D.E. Schwalm, J.L. Carroll, S. Hsiung, Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation, Evaluation Review 22 (2) (1998) 207–244.
[2] D.G. Altman, K.F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P.C. Gøtzsche, T. Lang, The revised CONSORT statement for reporting randomized trials: explanation and elaboration, Annals of Internal Medicine 134 (8) (2001) 663–694.
[3] E. Arisholm, H. Gallis, T. Dybå, D.I.K. Sjøberg, Evaluating pair programming with respect to system complexity and programmers expertise, IEEE Transactions on Software Engineering 33 (2) (2007) 65–86.
[4] D.T. Campbell, Factors relevant to the validity of experiments in social settings, Psychological Bulletin 54 (1957) 297–312.
[5] D.T. Campbell, J.C. Stanley, Experimental and Quasi-Experimental Designs for Research, Houghton Mifflin Company, Boston, 1963.
[6] T.C. Chalmers, P. Celano, H.S. Sacks, H. Smith, Bias in treatment assignment in controlled clinical trials, The New England Journal of Medicine (1983).
[7] G.A. Colditz, J.N. Miller, F. Mosteller, How study design affects outcomes in comparisons of therapy. I: Medical, Statistics in Medicine 8 (1989) 441–454.
[8] T.D. Cook, D.T. Campbell, Quasi-Experimentation: Design & Analysis Issues for Field Settings, Houghton Mifflin Co., Boston, 1979.
[9] W.J. Corrin, T.D. Cook, Design elements of quasi-experiments, Advances in Educational Productivity 7 (1998) 35–37.
[10] R. DerSimonian, L.J. Charette, B. McPeek, F. Mosteller, Reporting on methods in clinical trials, The New England Journal of Medicine 306 (22) (1982) 1332–1337.
[11] T. Dybå, V.B. Kampenes, D.I.K. Sjøberg, A systematic review of statistical power in software engineering experiments, Information and Software Technology 48 (8) (2006) 745–755.
[12] L.S. Feldt, A comparison of the precision of three experimental designs employing a concomitant variable, Psychometrika 23 (1958) 335–353.
[13] S. Grimstad, M. Jørgensen, A framework for the analysis of software cost estimation accuracy, in: The International Symposium on Empirical Software Engineering (ISESE), Rio de Janeiro, Brazil, September 21–22, ACM Press, 2006, pp. 58–65.
[14] J.E. Hannay, D.I.K. Sjøberg, T. Dybå, A systematic review of theory use in software engineering experiments, IEEE Transactions on Software Engineering 33 (2) (2007) 87–107.
[15] A.D. Harris, J.C. McGregor, E.N. Perencevich, J.P. Furuno, J. Zhu, D.E. Peterson, J. Finkelstein, The use and interpretation of quasi-experimental studies in medical informatics, Journal of the American Medical Informatics Association 13 (2006) 16–23.
[16] D.T. Heinsman, Effect Sizes in Meta-Analysis: Does Random Assignment Make a Difference? Doctoral thesis, Memphis State University, 1993.
[17] D.T. Heinsmann, W.R. Shadish, Assignment methods in experimentation: when do nonrandomized experiments approximate answers from randomized experiments?, Psychological Methods 1 (2) (1996) 154–169
[18] M. Hotopf, G. Lewis, C. Normand, Putting trials on trial-costs and consequences of small trials in depression: a systematic review of methodology, Journal of Epidemiology and Community Health 51 (1997) 354–358.
[19] ICH, Statistical Principles for Clinical Trials, 1998 (cited 2007 13 June). Available from: <http://www.ich.org/cache/compo/276-254-1.html/>.
[20] B. Jones, M.G. Kenward, Design and Analysis of Cross-Over Trials, Chapman & Hall/CRC, 2003.
[21] V.B. Kampenes, Quality of Design, Analysis and Reporting of Software Engineering Experiments, A Systematic Review, Doctoral Thesis, University of Oslo, 2007.
[22] V.B. Kampenes, T. Dybå, J.E. Hannay, D.I.K. Sjøberg, A systematic review of effect size in software engineering experiments, Information and Software Technology 49 (11–12) (2007) 1073–1086.
[23] B. Kitchenham, Procedures for Performing Systematic Reviews, Keele University, UK, Technical Report TR/SE-0401 and National ICT Australia, Technical Report 0400011T.1, 2004.
[24] B. Kitchenham, Empirical paradigm – the role of experiments, in: V.R. Basili, et al. (Eds.), Empirical Software Engineering Issues: Critical Assessment and Future Directions, Proceedings from Int. Workshop, Dagstuhl Castle, June 26–30, 2006, Lecture Notes in Compute Science 4336, Springer, 2007, pp. 25–32.
[25] B. Kitchenham, J. Fry, S. Linkman, The case against cross-over designs in software engineering, in: Eleventh Annual International Workshop on Software Technology and Engineering Practice (STEP'04), 2004.
[26] O. Laitenberger D. Rombach, (Quasi-)experimental studies in industrial setting, in: N. Juristo, A.M. Moreno, (Eds.), Series on Software Engineering and Knowledge Engineering, vol. 12, Lecture Notes on Empirical Software Engineering, 2003, World Scientific, Singapore, 167–227.
[27] M.W. Lipsey, Improving the evaluation of anticrime programs: there's work to be done, Journal of Experimental Criminology 2 (2006) 517–527.
[28] M.W. Lipsey, D.B. Wilson, The efficacy of psychological, educational and behavioral treatment, American Psychologist 48 (12) (1993) 1181–1209.
[29] M.W. Lipsey, D.S. Cordray, Evaluation methods for social intervention, Annual Review of Psychology 51 (2000) 345–375.
[30] J.R. McKay, A.I. Alterman, A.T. McLellan, C.R. Boardman, F.D. Mulvaney, C.P. O'Brien, Random versus nonrandom assignment in the evaluation of treatment for cocaine abusers, Journal of Consulting and Clinical Psychology 6 (4) (1998) 697–701.
[31] B.D. Meyer, Natural and Quasi-experiments in Economics, Technical Working Paper No. 170, National Bureau of Economic Research, Cambridge, MA, 1994.
[32] J.N. Miller, G.A. Colditz, F. Mosteller, How study design affects outcomes in comparisons of therapy. II: Surgical, Statistics in Medicine 8 (1989) 455–466.
[33] J.L. Myers, Fundamentals of Experimental Design, Allyn and Bacon, Boston, 1972.
[34] M.R. Novick, Data analysis in the absence of randomization, in: R.F. Boruch, P.M. Wortman, D.S. Cordray (Eds.), Reanalyzing Program Evaluations: Policies and Practices for Secondary Analysis of Social and Educational Programs, Jossey-Bass, San Francisco, 1981.
[35] K. Ottenbacher, Impact of random assignment on study outcome: an empirical examination, Controlled Clinical Trials 13 (1992) 50–61.
[36] S.J. Pocock, Clinical Trials. A Practical Approach, John Wiley & Sons Ltd., 1983.
[37] C.S. Reichardt, A typology of strategies for ruling out threats to validity, in: L. Bickman (Ed.), Research Design, Donald Campbell's Legacy, Sage Publications, Inc., 2000, pp. 89–115.
[38] L.L. Roos, Quasi-experiments and environmental policy, Policy Science 6 (1975) 249–265.
[39] D.B. Rubin, Practical Implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism, Biometrics 47 (1991) 1213–1234.
[40] W.R. Shadish, The empirical program of quasi-experimentation, in: L. Bickman (Ed.), Reserach Design: Donald Campbell's Legacy, Sage, Thousand Oaks, CA, 2000, pp. 13–35.
[41] W.R. Shadish, K. Ragsdale, Random versus nonrandom assignment in controlled experiments: do you get the same answer?, Journal of Consulting and Clinical Psychology 64 (6) (1996) 1290–1305
[42] W.R. Shadish, T.D. Cook, D.T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference, Houghton, Mifflin, Boston, 2002.
[43] D.A. Shapiro, D. Shapiro, Meta-analysis of comparative therapy outcome studies: a replication and refinement, Psychological Bulletin 92 (3) (1982) 581–604.
[44] J.P. Shepherd, Explaining feast and famine in randomized field trials, Evaluation Review 27 (3) (2003) 290–315.
[45] D.I. Simester, J.R. Hauser, B. Wernerfelt, R.T. Rust, Implementing quality improvement programs designed to enhance customer satisfaction: quasi-experiments in the United States and Spain, Journal of Marketing Research (2000) 102–112.
[46] D.I.K. Sjøberg, T. Dybå, M. Jørgensen, The future of empirical methods in software engineering research, in: L. Briand, A. Wolf (Eds.), Future of Software Engineering IEEE Computer Society, 2007, pp. 358–378.
[47] D.I.K. Sjøberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N.-K. Liborg, A.C. Rekdal, A survey of controlled experiments in software engineering, IEEE Transactions on Software Engineering 31 (9) (2005) 733–753.
[48] M.L. Smith, G.V. Glass, T.I. Miller, The Benefits of Psychotherapy, The Johns Hopkins University Press, USA, 1980.
[49] B. Thornley, C. Adams, Content and quality of 2000 controlled trials in schizophrenia over 50 years, British Medical Journal 317 (1998) 1181–1184.
[50] W.M.K. Trochim, The Research Methods Knowledge Base, Atomic Dog Publishing, 2001.
[51] D. Weisburd, C.M. Lum, A. Petrosino, Does research design affect study outcomes in criminal justice?, Annals of the American Academy of Political and Social Science 578 (2001) 50–70
[52] L. Wilkinson and the Task Force on Statistical Inference, Statistical methods in psychology journals: guidelines and explanations, American Psychologist 54 (8) (1999) 594–604.
[53] D.B. Wilson, M.W. Lipsey, The role of method in treatment effectiveness research: evidence from meta-analysis, Psychological Methods 6 (4) (2001) 413–429.
[54] D.B. Wilson, D.C. Gottfredson, S.S. Najaka, School-based prevention of problem behaviors: a meta-analysis, Journal of Quantitative Criminology 17 (3) (2001) 247–272.