Crossover Designs in Software Engineering Experiments: Benefits and Perils

Sira Vegas, Cecilia Apa, and Natalia Juristo

Abstract—In experiments with crossover design subjects apply more than one treatment. Crossover designs are widespread in software engineering experimentation: they require fewer subjects and control the variability among subjects. However, some researchers disapprove of crossover designs. The main criticisms are: the carryover threat and its troublesome analysis. Carryover is the persistence of the effect of one treatment when another treatment is applied later. It may invalidate the results of an experiment. Additionally, crossover designs are often not properly designed and/or analysed, limiting the validity of the results. In this paper, we aim to make SE researchers aware of the perils of crossover experiments and provide risk avoidance good practices. We study how another discipline (medicine) runs crossover experiments. We review the SE literature and discuss which good practices tend not to be adhered to, giving advice on how they should be applied in SE experiments. We illustrate the concepts discussed analysing a crossover experiment that we have run. We conclude that crossover experiments can yield valid results, provided they are properly designed and analysed, and that, if correctly addressed, carryover is no worse than other validity threats.

Index Terms—Experimental software engineering, controlled experiment, data analysis, crossover design, carryover

1 Introduction

Experimentation is a key issue in science and engineering. It is now very common practice to conduct laboratory experiments in software engineering (SE). However, designing an experiment and analysing the gathered data is a challenging error-prone activity. Shepperd et al. [36] analysed the results of 42 papers reporting studies that compare methods for predicting fault-proneness. They found that the research group accounted for 30 percent of the differences between studies, whereas the main topic of research accounted for only 1.3 percent. They concluded that *it matters more who does the work than what is done*. Experimentation is quite a recent practice in SE (compared with other much more mature experimental disciplines). We all need to learn more, and much more effort and research is needed to adapt the experimental paradigm to SE.

Experimental design aims to ensure that the effects observed in the response variable are due to the treatments. To do this, designs try to control other possible sources of variation that are liable to influence the response variable. There is no such thing as a perfect experiment, but there are better or worse experimental designs depending on the degree of control achieved [21].

Crossover is a particular type of design where each experimental subject applies all treatments, but different subjects apply treatments in a different order. Additionally, it counterbalances some of the effects caused by applying the treatments in a particular order, such as practice or fatigue.

Crossover designs are commonly used in experimental disciplines that work with human beings and animals [25], like psychology, pharmaceutical science, animal science, and healthcare (especially medicine). They are also popular in SE, as stated in [38] and shown by the literature survey that we have conducted (see Section 2).

However, crossover designs are complex and have been criticized and/or discouraged in both SE [23] and other disciplines [11], [12], [14], [15], [20], [34] for two main reasons: they are liable to the carryover threat and are hard to design and analyse. Even the US Food and Drug Administration (FDA) has for years discouraged [7], [16] the use of crossover studies on these grounds.[1]

Looking at a literature survey that we have performed, it appears from the recently reported approaches that the SE community has not yet fully grasped the multiple and complex issues involved in the design and analysis of crossover experiments. Such a faulty understanding could render results unreliable. Therefore, it is important to improve the state of the practice of designing and analysing crossover experiments among SE researchers.

For this research, we have studied writings from the field of experimentation concerning the design and analysis of crossover experiments, as well as literature on a specific

- *S. Vegas is with the Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte 28660, Madrid, Spain. E-mail: svegas@fi.upm.es.*
- *C. Apa is with the Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Julio Herrera y Reissig 565, Montevideo 11300, Uruguay. E-mail: ceapa@fing.edu.uy.*
- *N. Juristo is with the Escuela Técnica Superior de Ingenieros Informáticos, Universidad Politécnica de Madrid, Campus de Montegancedo s/n, Boadilla del Monte 28660, Madrid, Spain and the Department of Processing Information, University of Oulu, Finland. E-mail: natalia@fi.upm.es.*

Manuscript received 1 Apr. 2014; revised 12 May 2015; accepted 27 July 2015. Date of publication 11 Aug. 2015; date of current version 19 Feb. 2016. Recommended for acceptance by T. Menzies.
For information on obtaining reprints of this article, please send e-mail to: reprints.@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TSE.2015.2467378

1. The FDA is responsible for protecting the US public health by assuring the safety, efficacy and security of human and veterinary drugs, biological products, medical devices, our nation's food supply, cosmetics, and products that emit radiation. As such, it provides guidance in different topics, which reflect the FDA's thinking on the topic. One of the topics is clinical trials.

TABLE 1
Advantages and Disadvantages of Design Types Regarding Subject Allocation

Design	Advantages	Disadvantages
Independent measures	- Accounts for order effect	- Requires more participants - Does not account for differences between participants
Repeated measures	- Requires fewer participants - Accounts for differences between participants	- Does not account for order effect
Matched-pairs	- Reduces differences between participants - Accounts for order effect	- Requires more participants - Pair formation is very time-consuming - People are not always comparable

discipline, namely medicine.[2] We have identified the good practices driving this type of experiments, which we have adapted to SE. We have conducted a literature survey of crossover experiments in SE, identifying the most widespread shortcomings, and have drafted some good practices designed to raise the quality of this type of experiments.

The main contribution of this paper is twofold. First, it provides a picture of the state of the practice of crossover experiments. Second it provides mechanisms for SE researchers to run valid and reliable crossover experiments.

We have organized the article as follows. Section 2 discusses the foundations (generic principles) of crossover experiments. Section 3 shows the results of the literature survey that we have performed on crossover experiments in SE. Sections 4 and 5 explore good practices concerning crossover experiment analysis and design, respectively, according to the following schema. First, we adapt the generic principle to SE. Then, we review some real SE experiments that did not apply the good practice and highlight the dangers of not having adhered to the practice. Section 6 provides practical advice by summarizing the suggested good practices for SE researchers running crossover experiments. Section 7 illustrates an application example reporting a crossover experiment that we have conducted and discusses the differences in the results depending on the proper use of this type of design. Finally, Section 8 outlines the conclusions of our research.

2 Background

An important issue in experimental design is the assignment of experimental units to treatments to form experimental groups. When planning a SE experiment where the experimental units are subjects, there are three ways of assigning subjects to the different levels of the main factor (treatments) [28]. This leads to three types of designs:

- *Independent measures (or parallel or between-subjects) designs*. Each subject is assigned to only one treatment.
- *Repeated measures (or within-subjects) designs*. Each subject is assigned to all treatments.
- *Matched-pairs designs*. Pairs of subjects that can be considered twins are formed. Each subject in a pair is assigned to a different treatment.

2. Of all the disciplines that carry out crossover experiments, medicine appears to be the most similar to SE. To be precise, we ruled out any disciplines that experiment with psychology, as it often cannot randomly assign treatments to experimental subjects, and disciplines that experiment with animals.

Each design type has the advantages and disadvantages that are summarized in Table 1.

A crossover design is a particular case of a within-subjects or repeated measures design [25], devised to deal with the order effect—which is especially harmful in the case of experiments with subjects. In a repeated measures design, there are k observations on each subject. These k observations correspond to the same subject observed under each treatment. When groups of subjects apply treatments in different orders, we get a crossover design. A crossover design establishes different treatment application sequences, and subjects are assigned to these sequences. Table 2 shows an example of a crossover design for two treatments.

Note that while all crossover designs are repeated measures designs, not all repeated measures designs are crossover designs. The experiment published by Latorre [26] is an example of a repeated measures SE experiment that is not a crossover design; it reports an experiment that explores test-driven development (TDD). Subjects are asked to implement eight sets of requirements (RS1-RS8), with three requirements per set (R1-R2-R3, that is, easy, moderate and difficult, respectively). All subjects complete each set of requirements in the same order. RS1 and RS2 are developed according to the test-last development practice (control strategy), and TDD is used for RS3-RS8. It is a within-subjects design because repeated measures (of effectiveness and efficiency) are taken on each subject for each set of requirements, but it is not a crossover design because the order of the treatments is the same for all subjects: the test-last approach is used first and the TDD approach is used last.

As Table 1 shows, repeated measures designs, and particularly crossover designs, are useful for addressing two key problems to which SE experiments are commonly prone: small sample sizes and large between-subject variations:

- They *require fewer subjects* than a parallel design, as t (number of treatments) measures are taken from each subject. A parallel (independent measures) design with n subjects produces n observations, whereas a crossover design yields $n \times t$ observations (n subjects x t treatments).
- They *increase the sensitivity of the experiment*. The observation of the same subject exposed to all treatments controls between-subject differences. In crossover experiments, treatment effects from subject i are measured relative to subject i's average response to all treatments. As a result, subjects serve as their own control. This eliminates variability due to

TABLE 2
Example of a Two-Treatment Crossover Design

Sequence	Period	
	Period 1	Period 2
Group I: AB	Treatment A	Treatment B
Group II: BA	Treatment B	Treatment A

differences in average subject responsiveness from experimental error.

In a crossover design, the experimental *groups* are the *sequences*, that is, the order in which subjects apply treatments. The times at which each treatment is applied are known as *periods*. For example, the design in Table 2 has two treatment application sequences: AB and BA. Additionally, it has two periods: period 1, where subjects in sequence AB apply treatment A and subjects in sequence BA apply treatment B; and period 2, where subjects in sequence AB apply treatment B and subjects in sequence BA apply treatment A.

Carryover is an internal validity threat of crossover designs. It occurs when a treatment is administered before the effect of another previously administered treatment has completely receded [6]. Consequently, the treatments administered last might appear to be more effective than those administered first if the first treatment is boosting the effectiveness of the second, or less effective if the first treatment is detracting from the effectiveness of the second. The carryover effect may have a major impact on and even invalidate the final results of an experiment. Carryover effects are specific to disciplines and treatments. It is important to understand the carryover threat in SE experiments if we want to guarantee the validity of the results of experiments using crossover designs.

Apart from the treatment factors and possible blocking variables, other factors that intervene in a crossover design are [34]: period, sequence, carryover and subject. The effects of all the above factors must be studied in order to satisfactorily analyse a crossover design.

Taking into account the theoretical foundations of crossover designs sourced from a general-purpose experimentation book [25] and a book specializing in medical crossover experiments [34], as well as the shortcomings of the state of SE crossover experiment practice identified in our literature survey, we have compiled the following good practices that should drive a crossover experiment in SE:

1. *Define periods*. Decide how many times the subjects are going to repeat the experimental task and study the implications.
2. *Define sequences*. Specify the orders in which treatments are to be applied.
3. *Deal with carryover at design time*. Select the strategy to be used to account for carryover.
4. *Take into account subject variability*. The chosen data analysis technique must be able to account for dependent measures (repeated measures on the same subject).
5. *Deal with carryover at analysis time*. The manner in which carryover is accounted for in the analysis must match the design decision about carryover.
6. *Match analysis with design*. The variables included in the analysis have to be consistent with design decisions.
7. *Beware of effect size*. Depending on the design and/or results of data analysis, effect size may or may not be measured.

Good practices 1 to 3 are related to design, whereas good practices 4 to 7 are related to analysis. As we will see in Sections 4 and 5, each good practice addresses a different validity threat: design good practices (define periods, define sequences and deal with carryover at design time) address internal validity threats; analysis good practices (take into account subject variability, deal with carryover at analysis time, match analysis with design and beware of effect size) address conclusion validity threats. Validity threats cannot be ranked in terms of importance, as there is a trade-off relationship between them that experimenters have to address.

3 Crossover Experiments in SE: State of the Practice

Crossover designs were introduced at the very beginning of SE experimentation [2] and have been in use ever since. We have performed a survey of the literature published in the top SE conferences and journals over the last three years (2012–2014) in search of crossover experiments: IEEE Transactions on Software Engineering (TSE), ACM Transactions on Software Engineering and Methodology (TOSEM), the Empirical Software Engineering Journal (EMSE), the International Conference on Software Engineering (ICSE), the European Software Engineering Conference/Foundations on Software Engineering (ESEC/FSE), and the International Symposium on Empirical Software Engineering and Measurement (ESEM). The literature survey was the result of a manual search of all published papers.

We searched a total of 930 published papers (disregarding journal editorials) for papers with the following characteristics: 1) papers that report controlled experiments or quasi-experiments, 2) papers that compare at least two treatments, 3) papers where subjects are used to apply treatments, and 4) papers where treatment assignment to subjects is randomized (or, in the case of repeated measures designs, subjects apply all treatments or, in the case of crossover designs, subjects are assigned randomly to sequences of treatments). Notice that we omitted other empirical studies (exploratory studies, case studies, surveys, etc.), quasi-experiments that explore only one treatment (without a control group) or where subjects cannot be randomly assigned to treatments (due to treatment idiosyncrasy), experiments where the experimental subjects are not humans and experiments where the performance of human subjects is compared against a tool.

Additionally, we considered papers in which the experiment accounts for just part of the paper (typically the evaluation or validation section) or where the experiment is the focus.

Rows 3 to 6 in Table 3 show the number of identified papers by design type (we have replaced 0s by dashes for readability): independent measures, matched pairs, repeated measures (non-crossover) and crossover, respectively. We found that some papers report more than one experiment. Furthermore, some papers report several

TABLE 3
Results for Papers in Literature Review

Source	TSE			EMSE			TOSEM			ICSE			ESEM			ESEC/FSE			TOTAL
Design Year	2012	2013	2014	2012	2013	2014	2012	2013	2014	2012	2013	2014	2012	2013	2014	2012	2013	2014	
Independent Measures	5	2	1	1	1	3	-	1	-	3	1	5	2	1	2	-	-	1	29
Matched Pairs	-	-	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	-	2
Repeated Measures	-	-	2	-	-	1	-	1	1	1	1	4	-	-	-	-	-	-	11
Crossover	2	2	2	2	-	5	-	1	3	6	1	4	1	2	-	1	1	-	33
Mixed (CO/RM)	1	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	2
Mixed (CO/IM)	-	-	-	-	1	-	1	1	-	-	-	-	-	-	-	-	-	1	4
Mixed (IM/RM)	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	-	-	-	1
Papers experiments with subjects	8	4	5	3	3	10	1	4	5	10	3	14	3	3	2	1	1	2	82

TABLE 4
Results for Experiments in Literature Review

Source	TSE			EMSE			TOSEM			ICSE			ESEM			ESEC/FSE			TOTAL
Design Year	2012	2013	2014	2012	2013	2014	2012	2013	2014	2012	2013	2014	2012	2013	2014	2012	2013	2014	
Independent Measures	6	2	1	1	3	4	1	2	-	3	1	6	3	1	2	-	-	2	38
Matched Pairs	-	-	-	-	-	-	-	-	1	-	-	1	-	-	-	-	-	-	2
Repeated Measures	1	-	3	-	1	2	-	1	1	1	2	4	-	-	-	-	-	-	16
Crossover	3	6	3	4	5	10	1	3	9	7	1	7	2	3	-	1	1	2	68
Experiments with subjects	10	8	7	5	9	16	2	6	11	11	4	18	5	4	2	1	1	4	124

experiments having different designs. These papers appear in rows 7, 8 and 9: mixed crossover and repeated measures papers, mixed crossover and independent measures papers and mixed independent measures and repeated measures papers, respectively. Additionally, row 10 shows the total number of identified papers in which experimental subjects are humans. Table 4 shows the same results but for experiments. Note that, in this case, the mixed categories are not necessary, and only the total number of experiments using humans is reported.

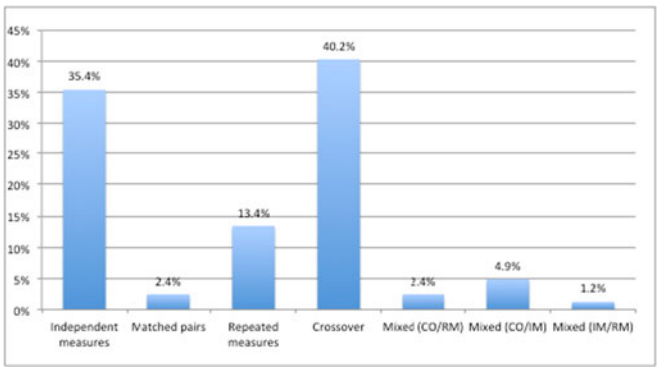

Fig. 1. Percentage of papers for each design type.

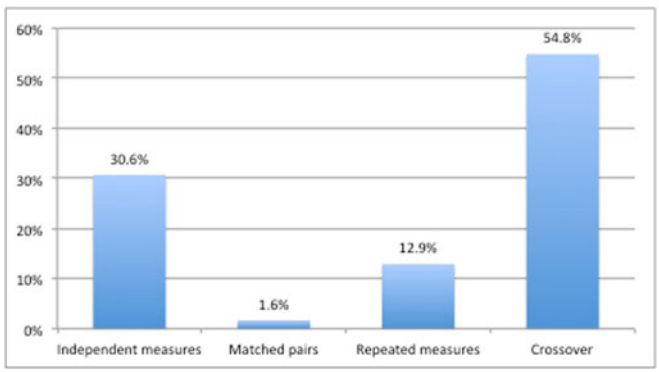

Fig. 2. Percentage of experiments for each design type.

As Figs. 1 and 2 show, crossover designs are the most used, followed by independent measures (between-subjects) designs. They appear in 47.5 and 41.5 percent of the papers, respectively, and are used in 54.8 and 30.6 percent of the experiments, respectively. Repeated measures (non-crossover) and matched-pairs designs are less common, appearing in 17 and 2.4 percent of the papers, respectively, and being used in 12.9 and 1.6 percent of the experiments, respectively.

In some cases (especially conference papers, which are shorter than journal papers), papers omitted particular information regarding experimental design and/or analysis. Note that experiment reporting guidelines [19] require the publication (irrespective of the source being a conference or journal paper) of certain information about the study that is essential for rating the quality of the experiment. Some authors describe experimental design textually without support in tabular form. This is an obstacle to the comprehension of crossover designs, which are more complex than other alternatives.

We have found that there is some confusion with respect to what a crossover experiment is among researchers. To be more precise:

- Only five papers (out of 39) use the "crossover" label to describe the design.[3] The others use names that are

3. One of them has a co-author in common with our paper. From now on, this paper will be disregarded.

TABLE 5
Adherence to Good Practices

Good practice	Number of papers adhering to good practice (out of 38)
Define periods[4]	1
Define sequences	2
Deal with carryover at design time	0
Take into account subject variablity	21
Deal with carryover at analysis time	0
Match analysis with design	0
Beware of effect size[5]	0

[4]*In this case, there are some papers that address it partially (see details in Section 4).*
[5]*This number is out of 13 papers that calculate effect size. The other 25 papers do not report effect size.*

either incorrect such as "factorial" or "fractional factorial" or are correct but not specific enough such as "within-subjects", "counterbalanced", "cross-validation", "cross-experiment" or "block". Note that the use of the wrong name is often not a symptom of experimenters not using the proper terminology but of researchers not being fully aware of the specific type of design that they are using and its implications. Although not all of these terms are incorrect, the use of an imprecise term (like, for example, within-subjects) also results in a deficient analysis.

- There is one experiment in which all sequences apply the treatments in the same order, and only the order in which the experimental object is applied is changed. Note that this is not a crossover experiment, but a **non**-crossover repeated measures design.

Additionally, there is always at least one paper that does not properly adhere to each good practice defined in Section 2. Table 5 shows the number of papers that properly adhere to each good practice defined in Section 2.

We would like to note two issues regarding Table 5:

- The exact consequential effect of experimental errors cannot be assessed, since this would mean that we would have to have access to and reanalyse information about the reported experiments missing in the papers and the raw data.
- When assessing the quality of an experiment based on its reporting, it is sometimes hard to determine whether the fault lies with the reporting or the experiment. Specifically, this applies to good practices concerning design (define periods, define sequences and deal with carryover at design time). For example, if an experiment does not report why one particular number of periods and not another were taken into account, we have no way of knowing whether the experimenters failed to weigh up the consequences of selecting one or other number (experimental error) or just omitted the issue in the report (reporting error). This does not apply to good practices concerning analysis (take into account subject variability, deal with carryover at analysis time, match analysis with design and beware of effect size). For example, if the statistical analysis technique used in an experiment with a within-subjects design does not match this design, it is the experiment and not the reporting that is at fault.

Despite these two issues, we think that the SE community would still benefit from researchers having access to advice on how to deal with crossover experiments.

TABLE 6
Two Treatment One Extra-Period Crossover Design

	Period 1	Period 2	Period 3
Sequence I	Treatment A	Treatment B	Treatment A
Sequence II	Treatment A	Treatment A	Treatment B
Sequence III	Treatment B	Treatment A	Treatment A
Sequence IV	Treatment B	Treatment B	Treatment A
Sequence V	Treatment B	Treatment A	Treatment B
Sequence VI	Treatment A	Treatment B	Treatment B

Subjects apply each treatment once except one treatment applied at least twice.

4 Design Issues

4.1 Define Periods

The AB/BA crossover design shown in Table 2 is a special type of design called *factorial crossover design*. Factorial crossover designs have the same number of periods as treatments, where all subjects apply every treatment under study once and once only. However, a crossover design may generally have n periods, n being greater or smaller than the number of treatments (called extra-period crossover designs and incomplete block crossover designs, respectively). Table 6 shows an example of a two-treatment one extra-period crossover design.

Period should not be confused with experimental session, although the two are equivalent in some experiments. A period is defined by the application of one treatment by one subject to one experimental object. A session is a portion of time spent by a subject on completing (one or more) experimental tasks. For example, a session may be composed of two periods (subjects applying two treatments one after the other). A period could cover several sessions.

The choice of the number of periods generally depends on the duration of the experimental task, its characteristics, available resources and training needs for task performance (in which case training and experiment are usually interspersed). Additionally, subjects may be allowed to take a break between periods (if they take place immediately one after the other) or within the period (if it is very long). Thus, for example, an experiment from the literature survey that we conducted investigating the impact of identifier style on a subject's ability to read phrases defines eight periods (two styles and four identifiers of different length were used) within a single session, as each period lasts just a few minutes [4]. Another experiment investigating program comprehension of domain-specific and general-purpose languages establishes two periods in two sessions [24]. The periods are distributed throughout an academic semester. The literature survey failed to identify any examples of an incomplete block crossover experiment.[6]

6. This could be because only two experiments examining more than two treatments were identified.

When designing a crossover experiment, the impact of the number and distribution of periods should be rated, as this can cause different types of internal validity threats. Let us look at some threats affecting periods using the two experiments discussed above. This should illustrate instances of such threats in SE experimentation:

- *Learning by practice.* This occurs when subject responses improve as they repeatedly perform the task (irrespective of the treatment(s)). Learning by practice can easily give the impression that the treatments administered last are more effective than the treatments administered first (when there really are no differences between treatments) simply because subjects have learned by having performed the task repeatedly to do the job better. In the above example on program comprehension, for example, subject performance might improve simply because they have practised program comprehension activities.
- *History.* The circumstances surrounding the periods may differ. In the above example on program comprehension, for example, subject training occurs just before the respective period. If all the training had been done at the outset, the subjects would have had longer for independent study and practice before applying the treatment in second place.
- *Copying.* This may occur when there is an interval between the periods, and the experiment is concerned with the same experimental objects in each period. Although each subject works on different experimental objects in each period, they may discuss their experimental objects with other subjects at the end of each period. Going back to the experiment on program comprehension, four experimental objects are defined, and subjects work on two in the first period and another two in the second period in order to rule out the possibility of this happening. There are other ways of mitigating this problem in practice: prevent participants from copying or taking experiment material with them, do not let people know that the same experimental objects will be used in all periods and, last but not least, do not create evaluation apprehension.
- *Tiredness/boredom.* This happens when the session (or experiment) is too long. The subject gets tired or bored of the session/experiment. For example, in the case of programming styles, subjects might experience this effect if they were to perform 24 instead of eight tasks.

All the crossover experiments examined in our literature survey establish periods. However, researchers do not appear to be familiar with their real meaning and all of their implications, as:

- None of the papers use the right term to refer to this concept (i.e., period). They use vague terms like "lab", "run", "experiment", "phase", etc. As with the term "crossover" mentioned in Section 3, the use of the wrong name here is not a problem of experimenters not binge acquainted with the terminology but of their failure to appreciate the actual design and its implications, and it results in a deficient analysis.
- None of the papers take into account the possible influence of the period; that is, they do not discuss the characteristics of the experiment that may vary between periods).[7]
- Twenty-one percent of the papers do partly account for period, possibly having a bearing on the results. They consider some of the validity threats associated with the period (typically learning) rather than the period itself. Understanding the implications of the period means examining in detail which changes there are to the experimental context between periods and discussing their implications in terms of validity threats.

Failure to take into account the effect of the period leads to the possible concealment of internal validity threats to the experiment that might influence results.

TABLE 7
Two-Treatment Factorial Crossover Design Where the Experimental Object Is a Two-Level Blocking Variable

	Period 1	Period 2
Sequence I	Treatment A, Object 1	Treatment B, Object 2
Sequence II	Treatment B, Object 1	Treatment A, Object 2
Sequence III	Treatment A, Object 2	Treatment B, Object 1
Sequence IV	Treatment B, Object 2	Treatment A, Object 1

4.2 Define Sequences

Crossover designs may account for all application sequences of the treatments ($t!$, t being the number of treatments), more sequences than $t!$ or just a few sequences. Indeed, it is quite common in SE for the number of sequences to be greater than $t!$. There are two reasons for this:

- *Blocking variables and/or other factors are added to the design*. Many of the experiments examined in our literature survey block by the experimental object (although other factors may be added). In this case, the number of sequences is $t! \times b$ (where b is the number of levels of the blocking variable). Thus, for example, one of the experiments from our literature survey assessing the effectiveness and efficiency of source code obfuscation techniques [5] establishes a two-period crossover design, where the subjects apply a different treatment to a different program in each period and the program is a two-level blocking variable. The proposed design is outlined in Table 7.
- *The treatment-induced learning effect is to be studied*. For example, Dias-Neto and Travassos [9] define a crossover design that accounts for all possible combinations of two treatments. The proposed design is outlined in Table 8.[8] Sequences III and IV are used to understand the treatment-induced learning effect, as response variable improvements observed in the second period using the same treatment must be due to subject learning.

7. For discretion's sake, we have preferred not to reference the criticized papers, as our goal is not to find fault with a particular paper but to illustrate the state of practice.

8. Sequence IV does not use exactly treatment B, but a slightly modified version of it.

TABLE 8
Two-Treatment Factorial Crossover Design with More than t! Sequences

	Period 1	Period 2
Sequence I	Treatment A	Treatment B
Sequence II	Treatment B	Treatment A
Sequence III	Treatment A	Treatment A
Sequence IV	Treatment B	Treatment B

As with periods, the impact of the number and formation of the sequences should be assessed during the design of a crossover experiment, as they are at the root of different threats to internal validity. Let us look at some threats affecting sequences using experiments from our literature survey. This should illustrate instances of such threats in SE experimentation:

- *Optimal sequence*. There is a sequence of treatment application that is conducive to experimental subjects achieving better results, which we refer to as *optimal sequence* in order to suggest that, applied in this sequence, the treatments systematically achieve better results. Thus, for example, an experiment from our literature survey that investigates whether the use of sequence diagrams in conjunction with functional requirements improves requirements understanding defines two periods and sequences [1]. Students have no knowledge of either requirements or sequence diagrams. The sequence in which sequence diagrams are omitted in the first period could benefit experimental subjects (they have to learn one technique, not two).

All the crossover experiments examined in our literature survey specify the established sequences. However, only two of the papers take into account the possible influence of the sequence; that is, all the experiments but one fail to discuss sequence characteristics that could lead subjects to perform better as a result of one or the other. Additionally, there is no justification of the sequences used in the experiments, arguing why the selected sequences are necessary and the omitted sequences are not.

As with the period, failure to take into account the sequence effect leads to the possible concealment of threats to the internal validity of the experiment that might have a bearing on the results.

4.3 Deal with Carryover at Design Time

The term carryover was first used in conjunction with crossover trials in medicine [6], [15], [17], where it is very common practice for patients to be given all treatments and act as their own control. Crossover designs in SE experiments can cause carryover as well. In a SE experiment, if the effect of one treatment carries on after the treatment is withdrawn, then the response to a second treatment may well be partly due to the previous treatment, and carryover occurs. For example, one of the experiments, taken again from our literature survey, assessing the reliability and effort of test first versus test last [27] establishes a two-period crossover design blocked by program. The application of test first may influence subjects to such an extent that the results of applying test last in second place are different from what they would be if the subject had not undergone a first treatment (test first). For example, subjects could learn to think about test cases before programming, and program accordingly without actually running the test cases. It will definitely take time for experiment participants to forget a SE method used as a treatment in an experiment. Therefore, it is no overstatement to say that SE methods tested in experiments may well have some sort of carryover effect on treatments tested later on.

In medicine, carryover is a physical effect (for example, a compound in the blood stream). However, carryover also has a psychological component in SE. For example, one of the experiments from our literature survey assessing effort, code compactness, parallel programming and debugging for the Scala and Java programming languages [31] establishes a two-period crossover design implementing two sets of requirements for the same software system (all subjects work on the same set of requirements in each period). Subjects have experience with Java, but not with Scala. Subjects applying Scala in the first period might get stuck with the problem and have trouble completing the code. This would give them a bad feeling about the experiment. On the other hand, subjects applying Java solve the task without difficulty. In both cases, the experiences during the first period influence how the subjects will go about the second period. Therefore, subjects using Java during the second period could perform worse than subjects using Scala.

Whenever possible, the effects of carryover must be neutralized. In medicine, this is usually done by means of a washout period. A washout period leaves sufficient time for the effect of the treatment to recede completely. Yet, as [23] notes, it is not possible in SE to establish a washout period to neutralize the carryover effect. For example, subjects cannot unlearn a technique that they have learned. They might forget how to use it, but how long would that take? And even if they did fail to recall the technique, would they be unable to remember anything at all or would a trace remain?

Carryover is just one of several possible interactions between treatment and period [34]. Let us look at an example illustrating another type of interaction. Imagine that, during the first period of the Java/Scala experiment discussed above, students participating in the experiment are nervous because they have an examination afterwards. This nerve-wracking situation is very likely to affect participants working with Scala (whose treatment is apparently harder to apply) more than the others. In this case, the existing treatment*period interaction is due not to carryover but to stress. In the particular case of an AB/BA crossover design, however, the treatment*period interaction is intrinsically confounded with carryover and with the sequence effect, and it is impossible to distinguish which of the three is occurring. This is where designs such the ones shown in Tables 6 (extra-period designs) and Table 8 come in, as they are able to single out these effects [25]. A crossover design is balanced for carryover effects when each treatment follows each of the other treatments an equal number of times [25]. Table 9 shows a three-treatment example. As we can see, treatments are shifted in each successive period.

The possibility of carryover should be considered in a crossover design. If there is any hint that carryover may exist, there are three options for dealing with it:

TABLE 9
Example of a Balanced Crossover Design for Three Treatments

	Period 1	Period 2	Period 3
Sequence I	A	B	C
Sequence II	A	C	B
Sequence III	B	A	C
Sequence IV	B	C	A
Sequence V	C	A	B
Sequence VI	C	B	A

- Run an independent measures experiment, which is not susceptible to carryover (but is not as good as a crossover design on the points shown in Table 1).
- Include carryover as a factor in the design and therefore take it into account in the analysis stage.
- Omit carryover as a factor, but thoroughly discuss the validity threat that it poses to the experiment.

Two papers (5 percent) in our literature survey mention carryover. However, they incorrectly associate it with the learning or fatigue effect (note that these effects are not carryover, but period effects). None of them take carryover into consideration as either a factor or a validity threat. This goes to show that the surveyed experimenters are not aware of the danger of carryover. Note that the carryover effect could invalidate experimental results, and it is not a good strategy just to turn a blind eye to this effect.

5 Analysis Issues

5.1 Take into Account Subject Variability

The observations in a crossover experiment are not independent, as we are measuring the same response variable several times on the same person (as many times as there are periods in the design). The chosen data analysis technique must be able to deal with this dependency.

Twenty-nine percent (11) of the papers examined in our literature survey use tests for independent samples: the Mann-Whitney, Student's t-test and one-way ANOVA. Some authors give explanations for using these tests (others do not explain why they use them):

- Some assume that, as the experimental subjects apply different treatments to different programs, data are unpaired. As the same variable is being measured more than once on the same person (even if they are working with different programs), data are dependent, and tests for independent samples are not applicable.
- Others claim that they have unpaired data, but do not explain how they did so. The only way to unpair the data of a crossover experiment is to analyse each and every experimental period as if they were different experiments, considering the second period to be a replication. But, even so, carryover can cause the second-period replication to yield different results. In this case, it would be impossible to identify the reason for that difference. In fact, [15], [17] recommend analysing the first-period data and omitting the data from successive periods unless there is evidence that there is no carryover between the applied treatments.

Using tests for independent samples to analyse the data of a crossover experiment would result in the experiment being exposed to the threat of statistical conclusion validity known as *violated assumptions of statistical tests* [35]. Experimental validity is a multi-layer construction; the different types of threats to validity are cumulative and build upon one another [37]. The bottom layer is (statistical) conclusion validity (is there a relationship between dependent and independent variables?). Next come internal validity (is the relationship causal?) and construct validity (can we generalize to other concepts?). Finally, at the top, we have external validity (can we generalize to other contexts?). Conclusion validity is the groundwork upon which everything else rests; it answers the most elementary questions of the cumulative perspective [33].

5.2 Deal with Carryover at Analysis Time

As discussed in Section 4.3, carryover may or may not be included in the design. This section only applies if it has been decided to include carryover in the crossover design. Carryover should be omitted from the analysis if, during the experimental design, the decision was taken not to take into account carryover and deal with it as a possible validity threat to the experiment.

The two-stage procedure [15], [17] has for many years been the standard recommended analysis to deal with carryover in a crossover experiment. It involves first conducting a statistical test on the data to examine the possibility of carryover having occurred. If it is judged not to have occurred, then the results of the analysis are reliable. If it is judged to have occurred, then a between-subjects test is carried out on the first-period data on the basis that carryover does not affect the values in the first period. The data collected in the other periods are discarded. There are two possible ways of checking whether carryover has occurred:

- Check the treatment*period interaction, that is, check whether there is an interaction between the factor under study and the period variable. If that interaction is not significant, then there is no carryover. However, there are other possible types of treatment*period interaction apart from carryover. Therefore, if there is another type of interaction, the treatment*period interaction might be significant even if there is no carryover.
- A more precise option is to add carryover to the analysis model. But statistical tests for carryover have limitations. An experiment might reveal significant carryover, which could, however, be due to type-I error and not to the existence of carryover. Likewise, the experiment might reveal a non-significant carryover, possibly due to type-II error, despite the fact that carryover really does exist.[9]

9. For example, D'Angelo [8] analysed a series of 324 AB/BA designs for carryover for two response variables at a 10 percent confidence level. Under these circumstances, we would expect (due to type-I error) 10 percent of the studies (32.4) to show significant carryover, even if there is none. D'Angelo found that there was significant carryover in 37 trials for one of the response variables and 34 for the other. These values are compatible with the non-existence of carryover.

Due to the untrustworthiness of these two approaches, some authors [12] suggest that a better approach is to analyse crossover experiments[10] without taking carryover into account. But we believe that it is not a good option for SE. Medicine has already learned what carryover is and its implications. SE is not yet at that stage.

As discussed in Section 4.3, none of the papers from our literature survey have taken into account carryover at design time. Carryover should be dealt with in any of the established manners to avoid threats to the validity of the results.

5.3 Match Analysis with Design

The experimental analysis of the data must be consistent with the proposed experimental design [21]. According to the design good practices defined in Section 4, the analysis of a crossover experiment must include both the period and the sequence, as well as the treatments (and other factors and blocking variables, if any). Additionally, it may or may not include carryover depending on the chosen design.

If the period and the sequence are certain not to influence the results, and there is no carryover[11] either (or it has been decided not to include carryover as factor in the design), the analysis can be simplified as follows:

- For two-treatment and two-period crossover designs that have no blocking variables, either a paired-samples t-test (also called matched-pairs t-test) or, if the sample formed by the difference scores of each subject for the response variable is not normally distributed, a non-parametric Wilcoxon test (also called matched-pairs signed-rank test) should be used.
- For n-treatment and n-period crossover designs, or two-treatment crossover designs with one or more additional between factors or with blocking variables, the repeated measures general linear model[12] should be used. This model assumes that the variances of the differences between all possible treatments are equal (sphericity). If this constraint is not met, the non-parametric Friedman test should be used (but only in the case of crossover designs with more than two treatments, without blocking variables and without additional factors).

Note that this data analysis procedure examines no more than the possible effect of treatments and additional factors on the response variable, because the period, sequence and carryover are certain not to have any bearing and therefore do not have to be taken into account.

If there is no evidence that the period and the sequence are likely to influence the results and/or it has been decided to take into account carryover, the linear mixed model analysis methods can be used. This model is an extension of the general linear model. It is a better method for analysing models with random coefficients (as is the case of subject in SE experiments) and data dependency due to repeated measures (as is the case of treatments in SE crossover designs) [29]. It assumes that the residuals follow a normal distribution with a mean of 0.[13] In the absence of normality, transformation of the response variable data is an option (using, for example, a logarithm, power or exponent).

In our literature survey, we found that 58 percent of the papers analyse the data using a paired-samples t-test or a non-parametric Wilcoxon test without evidence that the period, sequence and carryover can be omitted, whereas 29 percent include other variables in the analysis (for example, learning), but conduct several paired-samples t-tests (or Wilcoxon tests).

If the data from a crossover design are analysed in this way, the possible effects of sequence, period and carryover are ignored [34]. This means that the treatment's effect on the experiment results is being confounded with the effect of the other three variables. This could compromise the validity of the findings.

5.4 Beware of Effect Size

When analysing the data of an experiment, statistical significance (the probability that the observed differences in the response variable are due to treatments [10]) is always reported. However, statistical significance is not enough, since it does not say anything about the magnitude of the difference caused by the treatments. Effect size is used in SE to measure the magnitude of the difference. There are different methods to measure effect size in an experiment. The measures for effect size found in our literature survey are: Cliff's d, Cohen's d and probability of superiority.

All the papers in our literature survey report the statistical significance, and 35 percent report effect sizes for treatments. However, it does not always make sense to measure effect size. The effect size of the treatments should only be measured if the period, the sequence or any blocking variables have no bearing and there is no carryover. If there is carryover (but the period, sequence, blocking variables, other factors or subject have no bearing), the literature contemplates two different options: 1) use the data collected in the first period of the experiment (which were not exposed to carryover) [15], [17] or 2) take into account all the data from the experiment (not just data from the first period) [34], since the omission of the other periods could have an effect on the results if there are not many subjects or there is a large variability between subjects (which is why it was decided to use a crossover design in the first place). None of these papers take this issue into consideration.

6 Summary of Good Practices for Designing and Analysing Crossover Experiments

In order to help researchers to properly design and analyse crossover experiments, we have compiled the good practices discussed in Section 4 and Section 5.

During crossover experiment **design**, the following issues have to be taken into account:

10. Using any of the analysis methods described in Section 5.3.

11. A belief is not good enough. There must be knowledge (for example, experiments have already been run to confirm this point) that neither period, nor sequence, nor carryover influence the response variable.

12. The repeated measures general linear model is a statistical analysis that performs, (among others) repeated measures ANOVA.

13. Although the F test is quite robust against departures from the normal distribution [15]. However, such deviations occur very often in SE, and in many cases such deviations are also quite strong ones.

- With respect to *period*, it is necessary to:
  - Decide how many periods the experiment should have. When reporting the experiment, these points should be specified, justifying the chosen number of periods and explaining the criteria applied to establish this.
  - Consider what changes the subject may have experienced between periods and assess the possible internal validity threats to the experiment caused by such changes. The experiment report must discuss the possible changes and threats that they pose.
- With respect to *sequence*, it is necessary to:
  - Select the sequences for the experiment. When reporting the experiment, these points should be specified, justifying the chosen sequences and explaining the criteria applied to establish them.
  - Consider the possibility of any of the sequences being conducive to the treatment results and assess the possible consequences of this. The experiment report must state these points, explaining the assessed consequences.
- Use tables to illustrate treatment assignment (combined with other factors and/or blocking variables, if any) to periods and sequences because a mere textual description of all the details of the crossover design is hard to understand (as readers will surely have appreciated at some points of this manuscript where we have had to use tables to better present design details).
- With respect to carryover, it is necessary to:
  - Examine whether it is likely to exist.
  - If there is a risk of carryover, as the definition of adequate washout periods for each treatment is not an option in SE, use any of the following three alternatives: 1) opt for a between-subjects design,[14] 2) omit carryover as a factor, or 3) include carryover as a factor in the design. If it is decided to include carryover in the design, the design must be balanced.
  - If options 2 or 3 above are chosen, thoroughly discuss the carryover threat. It must be accepted that the findings will always be conditional upon the assumption that carryover has not seriously distorted the results.
  - Carry out replications with different designs to check results. The only way to find out whether or not there is carryover is to replicate experiments with different designs [13]. If the effects observed in an experiment with a certain design really exist, such effects should also be observable with other experimental designs.

During **analysis**, the following issues have to be taken into account in order to properly analyse a crossover design:

- Do not use techniques that do not deal with data dependency (like Student's t-test, the Mann-Whitney test or the one-way ANOVA).
- Analyse the effects of period and sequence and not just of treatments, unless there is proof that they will not influence the results.
  - If the period, sequence and carryover (and the other experiment factors and/or blocking variables, if any) have no influence, the paired t-test or the Wilcoxon test (for two-period crossover experiments) or the repeated measures general linear model[15] (for experiments with more than two periods or including blocking variables or more between-subjects factors) are possibilities.
  - If there is any hint that the period, sequence or carryover might be influencing the results (or more within-subjects factors have been included), use the linear mixed model (if the requirement of the test is not met, a logarithmic, exponential or power transformation is a possibility).
- With respect to carryover, there are two possible approaches depending on how the experiment has been designed:
  - If carryover has not been included as a factor in the design and is merely considered in the discussion of validity threats, no further action is necessary.
  - If carryover has been included as a factor in the design, add it as a factor to the analysis. However, there is a possibility of type-I or type-II error, and the possible existence of carryover still needs to be discussed as a threat to the validity of the experiment.

  Even if carryover is statistically significant, take into account all the data from the experiment (not just data from the first period). The omission of the other periods could have an effect on the results if there are not many subjects or there is a large variability between subjects (which is the very reason why a crossover design was used in the first place).
- With respect to effect size, calculate this measure only when the main factor of the experiment is the only statistically significant variable.

Table 10 summarizes the contents of this section.

7 Case Study: A Crossover Experiment

Let us examine a real AB/BA crossover experiment with one factor and two treatments that we have run. We present the design and analyse the collected data according to all the approaches explained in Section 5. This application example illustrates for real observations the differences in results when applying the different approaches for analysing crossover experiments.

7.1 Experimental Design

The **goal** of this experiment is to investigate the effectiveness of two test case design techniques: equivalence partitioning (EP) and branch testing (BT). The **null hypothesis** of the experiment is: *There is no difference of effectiveness between equivalence partitioning and branch testing*.

14. Subject to the threat that this could have an effect on the results if there are not many subjects or there is a large variability between subjects.

15. Or the repeated measures ANOVA.

TABLE 10
Issues to Be Taken into Account When Designing and Analysing Crossover Experiments

Stage	Issue	Good Practice
Design	**Period**	-Decide on the number of periods and justify the criteria applied to make the selection -Consider what changes subjects may have experienced between periods to assess internal validity threats
	Sequence	-Select sequences and justify the criteria applied to make the selection -Consider the possibility of any of the sequences being conducive to the treatment results and assess possible consequences
	Carryover	-Examine whether there is likely to be carryover -If there is a risk of carryover, select an option: 1. Choose a between-subjects design 2. Consider carryover as a threat to validity 3. Include carryover as a factor in the design -For options 2 or 3, discuss the carryover threat thoroughly. Accept that findings could be influenced by carryover -Run replications of the experiment with different designs to check results
	General	-Use tables to illustrate treatment assignment to periods and sequences
Analysis	**Subject variability**	-Do not use techniques that do not deal with data dependency (Student's t-test, Mann-Whitney or one-way ANOVA)
	Match analysis with design	-Analyse effects not just of treatments but also of period and sequence, unless there is prior empirical evidence that they will not influence the results -If period, sequence, carryover and other experiment factors and blocking variables have no influence, use the paired t-test, Wilcoxon tests or repeated measures GLM/ANOVA -If period, sequence or other factors might be influencing results, use the mixed linear model
	Carryover	-If carryover has not been included as an experiment factor, no further action is necessary -If carryover has been included as a factor in the design, add it as a factor of the analysis. Do not forget that type-I or type-II errors are possible -If carryover is significant, decide whether you want to use all or only first period experiment data
	Effect size	-Only calculate effect size when the main factor of the experiment is the only statistically significant variable

The experiment has one **response variable**: technique effectiveness. Technique effectiveness is measured as the percentage of faults detected by the set of test cases generated by a subject. Technique is the main (and only) experiment **factor**, with two treatments (or levels): equivalence partitioning [30] and branch testing with 100 percent branch coverage [3].

Treatments are applied by subjects, and people are intrinsically quite different. Due to dissimilarities between people already existing prior to the experiment (competences, abilities, etc.), there may be a relatively large variability between different people applying the same testing technique. For this reason, we choose a crossover **design**, shown in Table 11.

We ask all subjects to apply both techniques. Therefore, there are two **periods** in the experiment (one to apply the first technique, and another to apply the second technique), as shown in Table 11. Each period takes place in a different session, held one week apart from one another. As each session lasts four hours, both periods cannot take place in the same session. We are not interested in the effect of learning on the treatment, and therefore we have not considered adding an extra period.

We consider all possible **sequences** of technique application (BT-EP and EP-BT). Therefore, there are two experimental groups, each applying a different sequence. A priori we do not think that there is a chance of either of the sequences improving the experimental results of the other, as the white-box and black-box testing techniques are based on different principles and use different inputs. White-box techniques use the program, whereas black-box techniques use the specifications. Subjects are given either the program or the specifications in each treatment.

The experiment uses two similar **programs**, written in C (used in other empirical studies about testing techniques like [22] or [32]): nametbl (implementation of symbol table data structure and operations) and ntree (implementation of n-ary tree data structure and operations). They have 172 and 146 LOC, respectively. Each program contains six different **faults**. The programs are applied in the ntree-nametbl order.

The use of a crossover design in this experiment poses the threat of **carryover** from one technique to another. Subjects may carry over something (knowledge, an opinion, an impression, a frame of mind, etc.) from one testing technique to another applied in succession to supplement the strategy of the technique applied later on. At first glance,

TABLE 11
Experimental Design

	Program	**Ntree**		**Nametbl**	
	Period	**Period 1**		**Period 2**	
Technique	**Sequence**	**BT**	**EP**	**BT**	**EP**
Group I: BT-EP		X	-	-	X
Group II: EP-BT		-	X	X	-

TABLE 12
Descriptive Statistics

Technique	Mean	Std. Dev.	St. Error	N
Branch Testing	55.3027	20.82104	4.43906	22
Equivalence Partitioning	46.9700	17.54679	3.74099	22
Total	51.1364	19.48959		44

the differences between BT and EP technique strategies appear to be too great for carryover from BT to EP and vice versa. But we must be sure, opinions are not enough. So carryover has to be analysed to empirically check our suppositions. Therefore, a carryover factor is introduced in the design. As mentioned in Section 4, three variables are unavoidably confounded in AB/BA designs: sequence, carryover, and period*treatment interaction. We will discuss the implications of this for data analysis in Section 7.4.

The **subjects** participating in the empirical study are 22 undergraduate computer engineering students taking the Software Verification and Validation course at the Technical University of Madrid. They are trained in SE and have little or no professional experience in software development.

The **experimental procedure** consists of four sessions. The first two 2-hour sessions are training sessions in which subjects learn how to apply the techniques. The other two 4-hour sessions are experiment execution periods. In each period, subjects apply techniques and run the generated test cases. The procedure is as follows: subjects apply the corresponding technique and design a set of test cases; afterwards, they are given an executable version of the program and they run the generated test cases; finally, subjects identify failures in terms of incorrect outputs.

7.2 Threats to the Validity of the Experimental Design

The design shown in Table 11 addresses the validity threats as follows:

- There is a learning by practice threat. However, this threat can be mitigated by comparing effectiveness in both periods, and any improvement in effectiveness due to subjects repeatedly performing the experimental task can be studied.
- There is a history threat. All training takes place before the execution of the experiment. This means that subject effectiveness could be better in the second period, as subjects have an extra week in which to study.
- There is no copy threat, as the subjects work on a different program in each period. This counteracts the risk of them speaking to each other and discussing the experimental objects.
- Despite there being a total of six four-hour sessions, there is no tiredness/boredom threat because the experiment is part of a V&V course. Also subjects will be motivated because the result of the experimental task is the course grade.
- There is no threat of technique learning, since subjects apply each treatment just once, thereby ruling out the possibility of them applying a particular technique better due to the repeated application of the same technique.

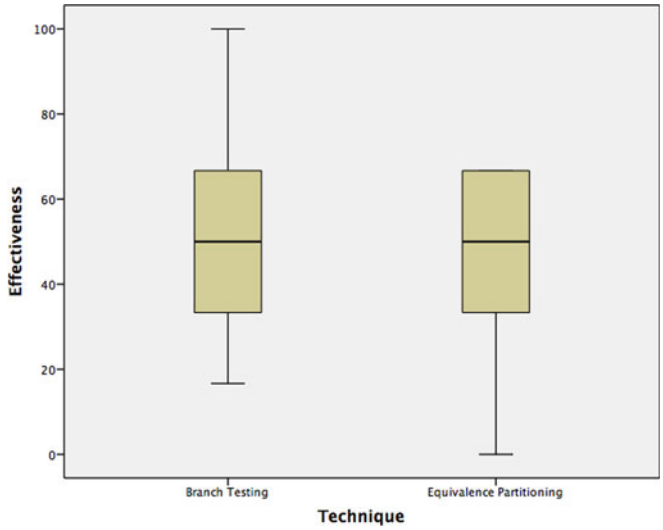

Fig. 3. Boxplot for technique effectiveness.

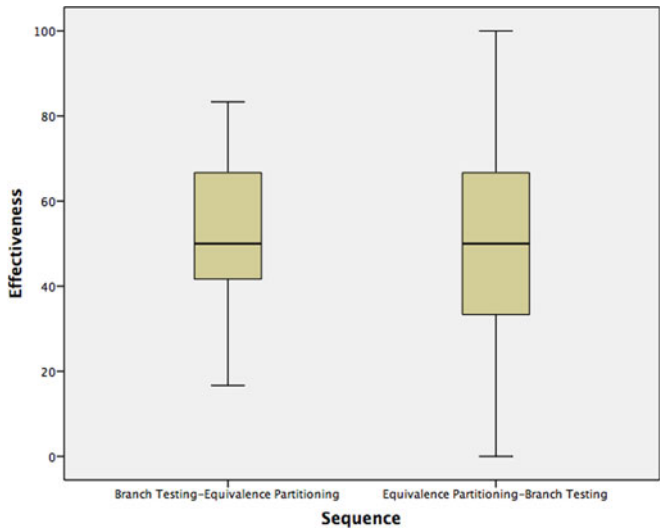

Fig. 4. Boxplot for sequence effectiveness.

- Period and program are confounded. Given our context, this is an unavoidable threat. It is impossible to run both periods in the same session (it would be too long). Additionally, if the two programs were exercised in the same period, subjects would be able to exchange information with each other about programs and faults, threatening the validity of the measured effectiveness.
- There is no threat of object learning because each subject applies the techniques to different programs. Subjects use each program once and do not get the chance to learn how it works for a second application.
- There is, in our opinion, no optimal sequence threat due to the differences between the two treatments (testing techniques) discussed above.
- The findings of the experiment are conditional upon the assumption that carryover has not seriously distorted the results.

7.3 Collected Data

Table 12 shows the descriptive statistics, and Figs. 3, 4, and 5 show boxplots of the effectiveness of each

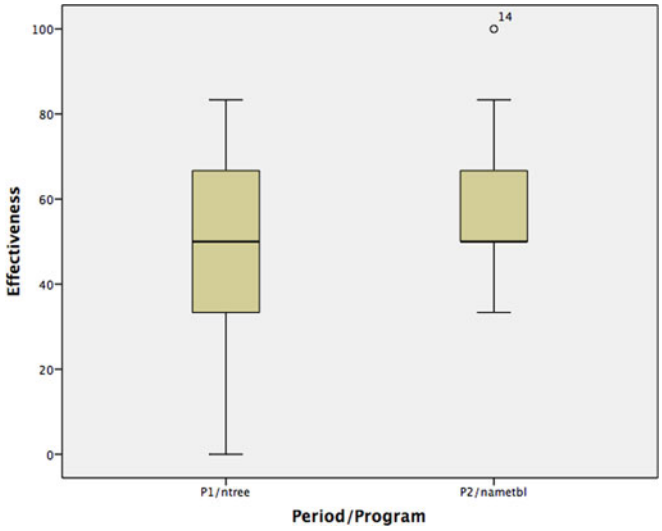

Fig. 5. Boxplot for period/program effectiveness.

technique, sequence and period/program respectively. Note that no outliers have been detected. We have used the Statistical Package for the Social Sciences (SPSS v20) for data analysis. Under normal circumstances, we would analyse this experiment using the linear mixed model. However, as the experiment is being used as case study in order to illustrate the contents of Section 4 and Section 5, we will apply all the analysis methods used in Section 5.

7.4 Data Analysis

7.4.1 Incorrect Data Analysis: Violated Assumptions of Statistical Tests

Let us start by wrongly applying the **independent samples t-test** on the treatments. Remember that this test cannot be applied in order to analyse this experiment because it assumes that data are independent, and a crossover design implies dependency (different measures taken on the same subject). The results are exposed to the validity threat of violated assumptions of statistical tests, which renders the findings unreliable, if not incorrect. However, as it is a relatively common mistake, we will apply this procedure as an anti-pattern.

To find out whether we can run this test, the normality of each treatment sample needs to be checked. The Shapiro-Wilk test shows a significance value of 0.218 (greater than 0.05) for BT and 0.006 (smaller than 0.05) for EP, which means that EP does not conform to a normal distribution.

As the independent-samples t-test cannot be applied, we apply the **Mann-Whitney test**. We find that the null hypothesis cannot be rejected, and therefore **equivalence partitioning is as effective as branch testing**, as shown in Table 13 (sig. 0.267).

TABLE 13
Mann-Whitney U Test

Total N	44
Mann-Whitney U	196.500
Wilcoxon W	449.500
Test Statistic	196.500
Standard Error	40.965
Standardized Test Statistic	–1.111
Asymptotic Sig. (2-sided test)	0.267

TABLE 14
Matched-Pairs Wilcoxon Signed-Rank Test

Total N	22
Test Statistic	39.500
Standard Error	18.951
Standardized Test Statistic	–1.504
Asymptotic Sig. (2-sided test)	0.133

7.4.2 Incorrect Data Analysis: Omitting Period, Sequence and Carryover

Let us now continue wrongly applying the **paired-samples t-test** on the treatments. Remember that this test cannot be applied to analyse this experiment because it is unable to study the influence of critical variables in a crossover experiment. The effect of the treatments is confounded with the effect of period, sequence or carryover in the results. This means that we cannot attribute any observed differences in the response variable (effectiveness) to the treatments (testing techniques), as they may be due to carryover or to an uncontrolled difference between periods or between treatment application orders. However, as it is a relatively common mistake, we will apply this procedure as an anti-pattern.

To find out whether we can run this test, the normality of the sample formed by the difference in the scores of each subject per treatment for the response variable needs to be checked. The Shapiro-Wilk test shows a significance value of 0.023 (smaller than 0.05), which indicates that the differences do not conform to a normal distribution.

As the paired-samples t-test cannot be used, we apply the **non-parametric Wilcoxon test**. We find that the null hypothesis cannot be rejected, and therefore **equivalence partitioning is as effective as branch testing**, as shown in Table 14 (sig. 0.133).

7.4.3 Incorrect Data Analysis: Analysing Period, Sequence and Carryover Separately

Let us now continue (again wrongly) the analysis reported in Section 7.4.2, this time including the additional factors involved in a crossover design—session and sequence[16]—in order to find out whether they are influencing effectiveness. Remember that conducting a separate analysis of each factor is not the correct procedure for analysing a design with more than one factor. However, as it is a relatively common mistake, we will apply this procedure as an anti-pattern.

In order to use the **paired-samples t-test** to analyse the within-subjects factor period, we need to check the sample formed by the difference in the scores of each subject per

16. Note that we do not analyse carryover. As explained in Section 4, carryover is confounded with sequence and with the technique*period interaction in AB/BA designs. This means that any (but no more than one) of these variables can be analysed. We opt to analyse sequence.

TABLE 15
Matched-Pairs Wilcoxon Signed-Rank Test for Session

Total N	22
Test Statistic	35.500
Standard Error	18.951
Standardized Test Statistic	–1.715
Asymptotic Sig. (2-sided test)	0.086

TABLE 16
Mann-Whitney U Test for Sequence

Total N	44
Mann-Whitney U	226.000
Wilcoxon W	436.000
Test Statistic	226.000
Standard Error	40.795
Standardized Test Statistic	–0.343
Asymptotic Sig. (2-sided test)	0.731

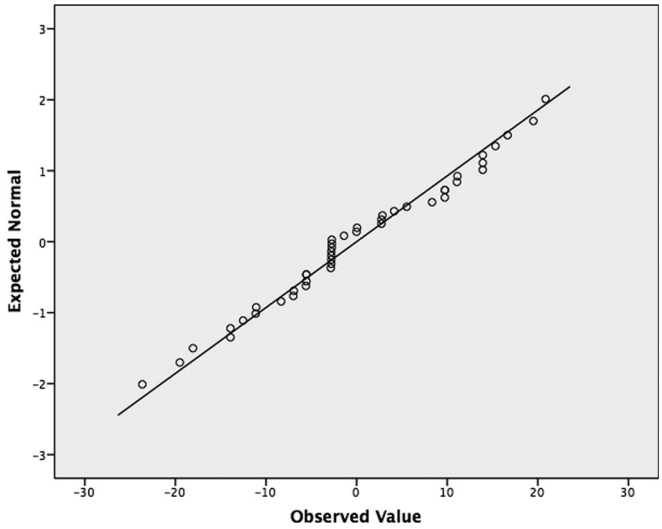

Fig. 6. Normal probability plot of residuals.

treatment for the response variable for normality. The Shapiro-Wilk test shows a significance value of 0.021 (smaller than 0.05), which indicates that the differences do not conform to a normal distribution.

As the paired-samples t-test cannot be used, we apply the **non-parametric Wilcoxon test**. We find that the null hypothesis cannot be rejected, and therefore **both periods are equally effective**, as shown in Table 15 (sig. 0.086). This means that none of the threats to the validity of the experiment associated with the session (learning by practice, history, tiredness/boredom, are actually occurring). Since in this experiment period and program are confounded, **program is not significant either**.

In order to apply the **independent-samples t-test** to the between-subjects factor sequence,[17] we need to check each treatment sample for normality. The Shapiro-Wilk test shows a significance value of 0.050 (equal to 0.05) for the sequence BT-EP, and 0.444 (greater than 0.05) for the sequence EP-BT. Therefore the sequence BT-EP does not conform to a normal distribution, while the sequence EP-BT does.

As the independent-samples t-test cannot be used, we apply the **non-parametric Mann-Whitney test**. We find that the null hypothesis cannot be rejected, and therefore **both sequences are equally effective**, as shown in Table 16 (sig. 0.086). Since sequence was confounded with carryover and the period*technique interaction, we can infer that none of these three variables (sequence, carryover and period*technique interaction) are significant. **There is no carryover effect between treatments**. Additionally, there is no optimal sequence validity threat.

7.4.4 *Correct Data Analysis: Analysing Period, Sequence and Carryover Jointly*

As already mentioned, the best method for analysing models with random coefficients and data dependency due to repeated measures is the **linear mixed model**. The model includes the following terms: technique (treatment), period (confounded with program in this experiment) and sequence (confounded with period*technique and carryover in this experiment) as fixed factors, and subject as random factor nested within sequence.

17. Note that this is not a repeated measures factor. Each subject is assigned to one and only one sequence.

TABLE 17
Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	8.369	13.158	0.006
Sequence	1	20	0.085	0.774
Period/Program	1	20	3.406	0.080
Technique	1	20	4.756	0.041

To apply the linear mixed model, the residuals must meet the condition of normality described in Section 5.3. Fig. 6 and the Shapiro-Wilk test (sig. 0.540) show that the model used is valid, as it meets the condition of normality. Note that each analysis technique is based on a different mathematical hypothesis testing approach. Consequently, not all techniques have the same requirements, and the same data may meet the requirements of some techniques but not others. This applies to these experiment data, where the sample formed by the difference in the scores of each subject per treatment for the response variable is not normal in the case of the paired-samples t-test, but the model residuals are normal for the linear mixed model.

According to the tests of fixed effects shown in Table 17, the effectiveness of the equivalence partitioning technique is significantly different from branch testing (sig. 0.041). **Equivalence partitioning is less effective than branch testing** (effectiveness of 46.53 and 55.56 percent, respectively).

Regarding the other two fixed factors checked (period and sequence), the tests reported in Table 17 show:

- **Period is not significant** (sig. 0.080) (meaning it is not affecting the response variable), that is, none of the threats to the validity of the experiment associated with session (learning by practice, history, tiredness/boredom) are actually occurring. **Program is**

TABLE 18
Experiment Results Using Different Analysis Techniques

Analysis technique	Should be applied	Effect of the variable on effectiveness		
		Treatment	Sequence/Carryover	Period/ Program
Mann-Whitney (treatment)	No	No	Not considered	Not considered
Wilcoxon (treatment)	No	No	Not considered	Not considered
Wilcoxon (treatment) + Wilcoxon (session) + Mann-Whitney (sequence)	No	No	No	No
Linear mixed model	Yes	Yes	No	No

not significant either (period and program are confounded).

- **Sequence is not significant** (sig. 0.774). Since sequence was confounded with carryover and the period*technique interaction, we can infer that none of these three variables (sequence, carryover and period*technique interaction) are significant. **There is no carryover effect between treatments**. Additionally, there is no optimal sequence validity threat.

As carryover, sequence or period turned out not to be significant, we now can safely attribute the observed effectiveness difference to the testing technique rather than to any other variable.

7.5 Discussion

We have used four different methods to analyse the data of the experiment from four different viewpoints: wrongly assuming data independency; wrongly omitting critical crossover design variables (sequence, period and carryover); wrongly analysing sequence, period and carryover separately; and rightly modelling sequence, period and carryover. The analysis techniques have revealed different results. This illustrates how using incorrect data analysis techniques[18] may lead to invalid results. Table 18 summarizes the results, showing which factors (technique, carryover, sequence, and period) do or do not have an effect on testing technique effectiveness.

Notice that the result of analyses that wrongly assume either data independency (Mann-Whitney test) or omit the effects of sequence, period and carryover (Wilcoxon test on treatments only) is that both testing techniques are equally effective, which is incorrect. The results of the analyses that wrongly study the effects of sequence, period and carryover with separate analyses (Mann-Whitney test for sequence and Wilcoxon test for period) are that both techniques are equally effective (which is again incorrect), and sequence (carryover) and period (program) do not influence the results. Finally, the result of the correct analyses that account for the effect of sequence, period and carryover in a single model (linear mixed model) is that equivalence partitioning is less effective than branch testing. These analyses find that sequence, period and carryover do not influence testing technique effectiveness.

8 Conclusions

In this paper, we have studied a common design in SE experiments: the crossover design. Crossover is such a complex design [38] that its use is sometimes discouraged based on the risk of incorrect analysis. The use of this type of design has been criticized in SE and other disciplines as being susceptible to the carryover threat and usually poorly analysed. But crossover designs also have big advantages, such as requiring fewer subjects or reducing variability due to differences among subjects. These strengths can benefit SE experiments, so it is inconvenient for researchers to have to steer clear of this design just because of the risk of incorrect analysis. We take the view that proper analysis can be ensured if researchers are acquainted with the risks and threats.

We have examined how another discipline (medicine) deals with crossover experiments and have adapted good practices to SE. We have surveyed the SE literature to check the state of design and analysis practices with regard to crossover experiments. We describe the perils of bad design and analysis practices for crossover designs and provide good practices for SE researchers using this type of experiments.

To show the hazards in practice we present a real case: an experiment that we ran to compare the effectiveness of two testing techniques. We analyse the data collected from this experiment using the wrong (all too often applied in SE literature) and right (following the good practices presented in this paper) analysis techniques. The different analysis techniques yield different results. This real example should raise the awareness of SE researchers regarding the importance of improving crossover experiment design and analysis.

18. Violating the required test conditions.

Acknowledgments

Research funded by the Spanish Ministry of Economy and Competitiveness research grant TIN2011-23216.

References

[1] S. Abrahao, C. Gravino, E. Insfran, and G. Scanniello, "Assessing the effectiveness of sequence diagrams in the comprehension of functional requirements: Results from a family of five experiments," *IEEE Trans. Softw. Eng.*, vol. 39, no. 3, pp. 327–342, Mar. 2013.

[2] V. R. Basili and R. W. Selby, "Comparing the effectiveness of software testing strategies," *IEEE Trans. Softw. Eng.*, vol. SE-13, no. 12, pp. 1278–1296, Dec. 1987.

[3] B. Beizer, *Software Testing Techniques*, 2nd ed. Boston, MA, USA: Int. Thomson Comput. Press, 1990.

[4] D. Binkley, M. Davis, D. Lawrie, J. I. Maletic, C. Morrell, and B. Sharif, "The impact of identifier style on effort and comprehension," *Empirical Softw. Eng.*, vol. 18, pp. 219–276, 2014.

[5] M. Ceccato, M. Di Penta, P. Falcarin, F. Ricca, M. Torchiano, and P. Tonella, "A family of experiments to assess the effectiveness and efficiency of source code obfuscation techniques," *Empirical Softw. Eng.*, vol. 19, pp. 1040–1074, 2014.

[6] T. J. Cleophas. *Human Experimentation. Methodologic Issues Fundamental to Clinical Trials*. Norwell, MA, USA: Kluwer, 1999.

[7] J. Cornfield and R. T. O'Neill, "Minutes of the food and drug administration," *Biostatics and Epidemiology Advisory Committee Meeting*, Jun. 23, 1976.
[8] G. D'Angelo, D. Potvin, and J. Turgeon, "Carryover effects in bioequivalence studies," *J. Biopharmaceutical Statist.*, vol. 11, pp. 27–36. 2001.
[9] A. C. Dias-Neto and G. H. Travassos, "Supporting the combined selection of model-based testing techniques," *IEEE Trans. Softw. Eng.*, vol. 40, no. 10, pp. 1025–1041, Oct. 2014.
[10] P. D. Ellis, *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[11] J. L. Fleiss, "A critique of recent research on the two treatment crossover design," *Controlled Clinical Trials*; vol. 10, pp. 237–244, 1989.
[12] P. Freeman, "The performance of the two-stage analysis of two-treatment, two-period crossover trials," *Statist. Med.*, vol. 8, pp. 1421–1432, 1989.
[13] O. S. Gómez, N. Juristo, and S. Vegas, "Understanding replication of experiments in software engineering: A Classification," *Inf. Softw. Technol.*, vol. 56, no. 8, pp. 1033–1048. 2014.
[14] A. P. Grieve, "A Bayesian analysis of the two-period crossover design for clinical trials," *Biometrics*, vol. 41, pp. 979–990, 1985.
[15] J. E. Grizzle, "The two-period change-over design and its use in clinical trials," *Biometrics*, vol. 21, no. 2, pp. 467–480, Jun. 1965.
[16] "Guidance for industry E9 statistical principles for clinical trials," U. S. Dept. Health and Human Services, Food and Drug Administration, Sep. 1998.
[17] M. Hills and P. Armitage, "Two-period crossover clinical trial," *Brit. J. Clinical Pharmacol.*, vol. 8, pp. 7–20, 1979.
[18] P. K. Ito, "Robustness of ANOVA and MANOVA test procedures," in *Handbook of Statistics*, vol. 1, P. R. Krishnaiah, Ed. Amsterdam, The Netherlands: Elsevier, 1980, pp 199–236.
[19] A. Jedlitschka, M. Ciolkowski, and D. Pfahl, "Reporting experiments in software engineering," in *Guide to Advanced Empirical Software Engineering*, F. Shull, J. Singer, and D. I. K. Sjøberg, Eds. London, U.K.: Springer, 2008, pp. 201–228.
[20] B. Jones and M. G. Kennard, *Design and Analysis of Crossover Trials*. London, U.K.: Chapman & Hall, 1989.
[21] N. Juristo and A. M. Moreno, *Basics of Software Engineering Experimentation*. New York, NY, USA: Springer, 2001.
[22] E. Kamsties and C. M. Lott, "An empirical evaluation of three defect-detection techniques," in *Proc. 5th Eur. Softw. Eng. Conf.*, 1995, pp. 362–383.
[23] B. Kitchenham, J. Fry, and S. G. Linkman, "The case against cross-over designs in software engineering," in *Proc. 11th Int. Workshop Softw. Technol. Eng. Practice*, 2003, pp. 65–67.
[24] T. Kosar, M. Mernik, and J. C. Carver, "Program comprehension of domain-specific and general-purpose languages: Comparison using a family of experiments," *Empirical Softw. Eng.*, vol. 17, pp. 276–304, 2012.
[25] R. O. Kuehl. *Design of Experiments: Statistical Principles of Research Design and Analysis*, 2nd ed. Pacific Grove, CA, USA: Duxbury Thomson Learning, 2000.
[26] R. Latorre, "Effects of developer experience on learning and applying unit test-driven development," *IEEE Trans. Softw. Eng.*, vol. 40, no. 4, pp. 381–395, Apr. 2014.
[27] O. A. L. Lemos, F. C. Ferrari, F. F. Silveira, and A. Garcia, "Development of auxiliary functions: Should you be agile? An Empirical assessment of pair programming and test-first programming," in *Proc. 34th Int. Conf. Softw. Eng.*, Zurich, Switzerland, 2012, pp. 529–539.
[28] G. Levine and S. Parkinson, *Experimental Methods in Psychology*. Hove, U.K.: Psychol. Press, 1993.
[29] C. E. McCulloch, S. R. Searle, and J. M. Neuhaus, *Generalized, Linear, and Mixed Models*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2008.
[30] G. J. Myers, T. Badgett, and C. Sandler, *The Art of Software Testing*, 2nd ed. New York, NY, USA: Wiley-Interscience, 2004.
[31] V. Pankratius, F. Schmidt, and F. Garretón, "Combining functional and imperative programming for multicore software: An empirical study evaluating scala and java," in *Proc. Int. Conf. Softw. Eng.*, Zurich, Switzerland, 2012, pp. 123–133.
[32] M. Roper, M. Wood, and J. Miller, "An empirical evaluation of defect detection techniques," *Inf. Softw. Technol.*, vol. 39, pp. 763–775, 1997.
[33] N. J. Salkind, *Encyclopedia of Research Design*, vol. 1. Newbury Park, CA, USA: SAGE, 2010.
[34] S. Senn, *Crossover Trials in Clinical Research*. New York, NY, USA: Wiley, 2002.
[35] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed. Boston, MA, USA: Cengage Learning, 2002.
[36] M. Shepperd, D. Bowes, and T. Hall, "Researcher bias: The use of machine learning in software defect prediction," *IEEE Trans. Softw. Eng.*, vol. 40, no. 6, pp. 603–616, Jun. 1, 2014.
[37] W. Trochim, J. P. Donnelly, and K. Arora, *Research Methods: The Essential Knowledge Base*. Boston, MA, USA: Cengage Learning, 2015.
[38] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering*. New York, NY, USA: Springer, 2010.

Sira Vegas received the PhD degree from the Universidad Politécnica de Madrid in 2002. She is currently associate professor of software engineering at the Universidad Politécnica de Madrid. Her main research interests include experimental software engineering and software testing. She is a reviewer of highly ranked journals such as the *IEEE Transactions on Software Engineering*, *Empirical Software Engineering Journal*, *ACM Transactions on Software Engineering and Methodology* and *Information and Software Technology*. She was a program chair for the International Symposium on Empirical Software Engineering and Measurement (ESEM) in 2007. She began her career as a summer student at the European Centre for Nuclear Research (CERN, Geneva) in 1995. She was a regular visiting scholar in the Experimental Software Engineering Group, University of Maryland from 1998 to 2000, and a visiting scientist at the Fraunhofer Institute for Experimental Software Engineering in Germany in 2002.

Cecilia Apa received the MsC degree from the Universidad de la República, Montevideo, Uruguay, in 2014. She is currently professor of software engineering at the Universidad de la República and consultant at Scanntech. Her main research interests include software verification and validation and software process improvement. She was a regular visiting scholar at the Software Engineering Research Group, Universidad Politécnica de Madrid.

Natalia Juristo received the PhD degree from the Universidad Politécnica de Madrid in 1991. She is currently a full professor of software engineering at the Universidad Politecnica de Madrid. She received a Finland Distinguished Professor Program (FiDiPro) professorship, starting in January 2013. She was the director in the UPM MSc in Software Engineering from 1992 to 2002 and a coordinator in the Erasmus Mundus European Master on SE (with the participation of the University of Bolzano, the University of Kaiserslautern and the University of Blekinge) from 2006 to 2012. Her main research interests include experimental software engineering, requirements, and testing. She coauthored the book *Basics of Software Engineering Experimentation* (Kluwer, 2001). She is a member of the editorial boards of the *IEEE Transactions on SE* and *Empirical SE Journal*. She began her career as a developer at the European Space Agency, Rome, and the European Center for Nuclear Research, Geneva. She was a resident affiliate at the CMU Software Engineering Institute in Pittsburgh in 1992.