Understanding replication of experiments in software engineering: A classification

Omar S. Gómez [a,*], Natalia Juristo [b,c], Sira Vegas [b]

[a] Facultad de Matemáticas, Universidad Autónoma de Yucatán, 97119 Mérida, Yucatán, Mexico
[b] Facultad de Informática, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain
[c] Department of Information Processing Science, University of Oulu, Oulu, Finland

A B S T R A C T

Context: Replication plays an important role in experimental disciplines. There are still many uncertain-ties about how to proceed with replications of SE experiments. Should replicators reuse the baseline experiment materials? How much liaison should there be among the original and replicating experiment-ers, if any? What elements of the experimental configuration can be changed for the experiment to be considered a replication rather than a new experiment?
Objective: To improve our understanding of SE experiment replication, in this work we propose a classi-fication which is intend to provide experimenters with guidance about what types of replication they can perform.
Method: The research approach followed is structured according to the following activities: (1) a litera-ture review of experiment replication in SE and in other disciplines, (2) identification of typical elements that compose an experimental configuration, (3) identification of different replications purposes and (4) development of a classification of experiment replications for SE.
Results: We propose a classification of replications which provides experimenters in SE with guidance about what changes can they make in a replication and, based on these, what verification purposes such a replication can serve. The proposed classification helped to accommodate opposing views within a broader framework, it is capable of accounting for less similar replications to more similar ones regarding the baseline experiment.
Conclusion: The aim of replication is to verify results, but different types of replication serve special ver-ification purposes and afford different degrees of change. Each replication type helps to discover partic-ular experimental conditions that might influence the results. The proposed classification can be used to identify changes in a replication and, based on these, understand the level of verification.

1. Introduction

Experimentation is an essential part of SE research. "[In SE] Experimentation can help build a reliable base of knowledge and thus reduce uncertainty about which theories, methods, and tools are adequate" [68]. Replication is at the heart of the experimental paradigm [61] and is considered to be the cornerstone of scientific knowledge [53].

* Corresponding author. Address: Anillo Periférico Norte, Tablaje Cat. 13615, (Room EA-7) Colonia Chuburna Inn, Mérida, Yucatán, Mexico. Tel.: +52 (999) 942 31 40x1117.
E-mail addresses: omar.gomez@uady.mx (O.S. Gómez), natalia@fi.upm.es (N. Juristo), svegas@fi.upm.es (S. Vegas).

To consolidate a body of knowledge built upon evidence, experimental results have to be extensively verified. Experiments need replication at other times and under other conditions before they can produce an established piece of knowledge [13]. Several replications need to be run to strengthen the evidence.

Most SE experiments have not been replicated. Sjøberg et al. [66] reviewed 5453 articles published in different SE-related journals and conference proceedings. They found a total of 113 controlled experiments, of which 20 (17.7%) are described as replications. Silva et al. [65] have conducted a systematic review of SE replications. They found 96 papers reporting 133 replications of 72 original studies run from 1994 to 2010.

If an experiment is not replicated, there is no way to distinguish whether results were produced by chance (the observed event occurred accidentally), results are artifactual (the event occurred

because of the experimental configuration but does not exist in reality) or results conform to a pattern existing in reality. Different replication types help to clarify which of these three types of results an experiment yields.

Most aspects are unknown when we start to study a phenomenon experimentally. Even the tiniest change in a replication can lead to inexplicable differences in the results. For immature experimental knowledge, the first step is replications closely following the baseline experiment to find out which experimental conditions should be controlled [10]. As Collins [16] explained for experiments in physics, "the less that is known about an area the more power a very similar positive experiment has to confirm the initial result. This is because, in the absence of a well worked out set of crucial variables, any change in the experiment configuration, however trivial in appearance, may well entail invisible but significant changes in conditions". For mature knowledge, the experimental conditions that influence results are better understood and artifactual results might be identified by running less similar replications. By using different experimental protocols, it is possible to check whether the results correspond to experiment-independent events. "As more becomes known about an area however, the confirmatory power of similar-looking experiments becomes less." [16]

The immaturity of experimental SE knowledge has been an obstacle to replication. Context differences usually oblige SE experimenters to adapt experiments for replication. As key experimental conditions are yet unknown, slight changes in replications have led to differences in the results which prevent verification. Attempts at combining replication results (Hayes [26], Miller [49–51], Hannay et al. [25], Jørgensen [35], Pickard et al. [55], Shull et al. [62], Juristo et al. [32]) have reported that it was not possible to verify results because of differences in experimental conditions.

There is no agreement in SE about what a replication is in terms of how many changes can be made to the baseline experiment and the purpose of such changes (as we will see in Section 2).

A classification of replications for SE may help form a better understanding of the particular verification purpose of each type of replication and what changes are valid for each type.

This paper is organized as follows. Section 2 discusses replication classifications proposed in SE. Section 3 describes different types of replication proposed in other disciplines. Section 4 outlines the research method that we have followed. The remainder of the paper reports each step of the research method. Section 5 describes the elements of an experimental configuration in SE. Section 6 introduces what specific verification purposes a replication can have. Section 7 describes a classification of replication types for SE experiments. Section 8 discusses the advantages of systematic changes in replications. Section 9 compares our proposal with other SE classifications proposed in the literature. Section 10 discusses researcher positions on SE replications. Finally, Section 11 presents the conclusions.

2. Related work

We have not found any research that specifically aims to classify replications in experimental SE. We have identified three works that have classified replications as part of the research.

Basili et al. [5] present a framework for organizing sets of related studies. They describe different aspects of the framework. One framework aspect defines a three-category classification of replications: (1) replications that do not vary any research hypothesis, (2) replications that vary the research hypotheses and (3) replications that extend the theory.

Basili et al. [5] identify two replication types that do not vary any research hypothesis:

- Strict replications, which duplicate as accurately as possible the original experiment.
- Replications that vary the manner in which the experiment is run. These studies seek to increase confidence in experimental results. To do this, they test the same hypotheses as previous experiments, but alter the details of the experiments in order to address certain internal threats to validity.

They identify three replication types that vary the research hypotheses:

- Replications that vary variables which are intrinsic to the object of study. These replications investigate what aspects of the process are important by systematically changing intrinsic properties of the process and examining the results.
- Replications that vary variables which are intrinsic to the focus of the evaluation. They may change the ways in which effectiveness is measured in order to understand the dimensions of a task for which a process results in most gain. For example, a replication might use a different effectiveness measure.
- Replications that vary context variables in the environment in which the solution is evaluated. They can identify potentially important environmental factors that affect the results of the process under investigation and thus help understand its external validity.

Replications that extend the theory are not further sub-divided. These replications help determine the limits to the effectiveness of a process by making big changes to the process, product, and context models to see if basic principles still hold.

In his master thesis, Almqvist [2] studies the use of controlled experiment replication in SE. He surveys 44 articles describing 51 controlled experiments and 31 replications. Categories are defined to organize the identified experiments. One of the categories develops a classification for pigeonholing the identified replications. As a reference, Almqvist [2] takes the concept of close and differentiated replication described in the accounting area by Lindsay and Ehrenberg [41] (depending on whether the replication attempts to keep almost all the known conditions of the study much the same or very similar at least, or have deliberate variations in fairly major aspects of the conditions of the study), to which he adds the internal and external replications used by Brooks et al. [11] (depending on whether the replication is run by the same experimenters or independently by other experimenters). Based on these classifications, Almqvist [2] defines the following four replication types:

1. Similar-Internal Replications.
2. Improved-Internal Replications.
3. Similar-External Replications.
4. Differentiated-External Replications.

Krein and Knutson [39] propose a unifying framework for organizing research methods in SE. They build a taxonomy of replications as part of such framework. The taxonomy defines four types of replication:

- Strict replication. An experiment that is meant to replicate a prior study as precisely as possible.
- Differentiated replication. An experiment that intentionally alters aspects of the prior study in order to test the limits of that study's conclusions.
- Dependent replication. A study that is specifically designed with reference to one or more previous studies, and is, therefore, intended to be a replication study.

- Independent replication. An experiment that addresses the same questions and/or hypotheses as a previous study, but is conducted without knowledge of, or deference to, that prior study either because the researchers are unaware of the prior work, or because they want to avoid bias.

There are other works in SE that mention replication types, albeit not for classification or inventory purposes. For example, Brooks et al. [11] and Mendonça et al. [48] refer to the differences between internal and external replication. Shull et al. [63] discuss some replication types (exact, independent, dependent and conceptual replication) to describe the role that replication plays in SE. Finally, Lung et al. [42] mention two types of replication (literal and theoretical replication) to explain the replication type that they conducted, and Mandić et al. [45] discuss two replication types: exact or partial replications and replications designed to improve the original experiment.

Other issues about replication have been discussed in SE literature. Some researchers like Miller [51] or Kitchenham [38] advise the use of different protocols and materials to preserve independence and prevent error propagation in replications by using the same configuration. This contrasts with recommendations by other researchers like Basili et al. [5] or Shull et al. [63] on the reuse of some experiment materials to assure that replications are similar enough for results to be comparable.

Today's SE knowledge on experiments replication has the following shortcomings:

(1) There is no agreement on what a replication is. Different authors consider different types of changes to the baseline experiment as admissible.
(2) None of the classifications are exhaustive in the sense that it is unclear whether the actual classification covers all possible changes that can be made to the experiment
(3) There is no agreement as to the terms used to refer to the different replication types.

In order to gain insight about replication, we have expanded the study of related work to other disciplines and examined replication classifications used in different branches of science.

3. Replication types in other disciplines

We used the Scopus® database to search for replication classifications in any discipline. We selected this database on Dieste et al.'s advice [19]: it covers a wide range of publications in different branches of science, and, the search engine is very robust.

We reviewed the titles, abstracts and keywords of 7343 documents returned by the search strings used and singled out promising documents that appeared to discuss how to run a replication. Appendix A describes the details of the search.

We read the promising documents and identified nine that describe replication types. Most of these nine articles reference other papers that describe different types of replications. We identified a total of 25 replication classifications, nine from Scopus® and 16 from references listed in the Scopus® documents. Table 1 shows the nine classifications retrieved from Scopus® (column 2), with their respective search strings (column 1), the 16 additional references found from the initial nine (column 3), and whether or not they were used in our research (column 4).

Of the 25 identified replication classifications, we selected 20 for analysis. We were unable to locate Reid et al.'s classification [58] either on the web or in the repository of historical archives referenced by the journal. We retrieved Sargent's classification [60] but decided to omit this document and not to retrieve the other three [37,7,57] belonging to the field of parapsychology, as this is a controversial field in terms of scientific research.

Table 1
References of identified classifications.

Search string using field code TITLE-ABS-KEY	Classifications from SCOPUS	Referenced classifications	Considered in our research
Replication AND classification	Beck 1994	Finifter [21]	✓
		Lykken [43]	✓
		Sidman [64]	✓
		La Sorte [40]	✓
		Blomquist [9]	✓
replication AND kind	Lüdtke [44]	✗	✗
Replication W/2 type	Schmidt [61]	Radder [56]	✓
		Lykken [43]	✓
		Hendrick [27]	✓
		Sargent [60]	✗
		Keppel [37]	✗
		Beloff [7]	✗
		Rao [57]	✗
	Adams et al. [1]	Kelly et al. [36]	✓
	La Sorte [40]	✗	✗
	Tsang and Kwan [69]	Brown and Coney [12]	✓
		Fuess [22]	✓
		Hendrick [27]	✓
		Lindsay and Ehrenberg [41]	✓
		Mittelstaedt and Zorn [52]	✓
		Reid et al. [58]	✗
Replication AND typology	Bahr et al. [3]	Lykken [43]	✓
		Finifter [20]	✓
		La Sorte [40]	✓
		Hyman and Wright [30]	✓
Replication AND taxonomy	Kelly et al. [36]	Sidman [64]	✓
		Lykken [43]	✓
		Barker and Gurman [4]	✓
		Finifter [20]	✓
	Finifter 1975	Finifter [20]	✓

The description of the 20 replication classifications studied is available at [18]. Most are also outlined in Gómez et al. [23].

We found that there are no standard intra- or interdisciplinary terms to refer to different replication types. However, as shown in Table 2, the replication types covered by the 20 classifications studied fall into three groups:

- Group I replications vary little or not at all with respect to the baseline experiment. Everything is (almost) exactly the same as in the baseline experiment. This type of replication aims to verify that the results are not a product of sampling error caused by type I error.[1]
- Group II replications do vary with respect to the baseline experiment. The experimental configuration is modified in the replication. There can be several degrees of similarity between the replication and the baseline experiment. Experimental elements that might be changed are: measurement instruments, metrics, protocol, populations, experimental design or researchers. This group shows that the experimental elements of a replication do not necessarily all have to be identical to the baseline experiment.

[1] Type I error occurs when the null hypothesis is rejected when it is true, i.e. when the analysis of the sample data gathered (experiment) shows a significant difference between the treatments that it compares, but no such difference exists in the population (reality).

Table 2
Identified ways of replicating an experiment.

Author(s)	Group I	Group II	Group III
Adams et al. [1]	Literal	Operational, Instrumental	Constructive
Bahr et al. [3]	Types A..D	Types E..H	Types I..P
Barker and Gurman [4]	Type I	Types II and III	Type IV
Beck [6]	Type I	Type II	Type III,V
Blomquist [9]	Literal	–	Construct
Brown and Coney [12]	–	Replication and replication with extension	–
Finifter [20]	–	Virtual	Systematic
Fuess [22]	Duplication	Replication with extension	–
Hendrick [27]	Strict	Partial	Conceptual
Hyman and Wright [30]	Type II	Type I	Type III
Kelly et al. [36]	Literal	Operational, Instrumental	Constructive
La Sorte [40]	–	Retest, Independent	Theoretical
Lindsay and Ehrenberg [41]	Close	Differentiated	Differentiated
Lüdtke [44]	Exact	Independent	Independent
Lykken [43]	Literal	Operational	Constructive
Radder [56]	Reproducibility of the material realization of an experiment	Reproducibility of an experiment under a fixed theoretical interpretation	Reproducibility of the result of an experiment
Schmidt [61]	Direct	Direct	Conceptual
Sidman [64]	Direct	Systematic	–
Tsang and Kwan [69]	Exact	Empirical generalization, generalization and extension	Conceptual

- The only thing that Group III replications have in common with the baseline experiment is that they are both based on the same theoretical structure, i.e. they share the same constructs and hypotheses. This type of replication aims to verify findings using a different experimental configuration.

Another two types of experiment replication described in the literature have not been included in Table 2. We decided not to consider these two types as true replications as the experiment is not run again. One of these types uses existing data for reanalysis using either the same or different procedures to the baseline experiment: Checking of Analysis [69], Reanalysis of Data [69], Internal Replication [40], Pseudoreplication [20], and Type I and II replications [52]. The other type uses different statistical models over the data generated in the baseline experiment. This type of replication is used to compare the statistical model used in the baseline study: type II and IV replications [52].

Although the overall objective of a replication is to verify results, the existence of different replication types suggests that each one has a specific aim. For example, according to Lykken [43], the goal of operational replication is to check that the experimental recipe outputs the same results with another experimenter. For this type of replication, therefore, the experimental protocol must be unaltered and the experimenter must change. Finifter's systematic replication [20] aims to output new findings using different methods to the baseline experiment. For this replication type, therefore, the experimental configuration must vary.

Summarizing our findings about replication classifications in other disciplines than SE:

(1) There is no generally accepted classification. All experimental disciplines (especially the least mature) have developed their own classifications.
(2) The bounds of replication are fuzzy, e.g. should a reanalysis be considered a replication?
(3) There is no agreement as to the terms used to refer to different replication types.
(4) Different replication types serve special verification purposes and afford different degrees of change.

Based on the results of this survey and the findings reported in Section 2, there appears to be a need to develop a specialized replication classification for SE that: (1) clearly defines the bounds of a replication and the admissible changes; (2) exhaustively states all possible changes to the original experiment; (3) uses self-explanatory terms; and (4) associates each replication type with its verification purpose(s).

4. Research method

The process we followed to achieve the research objectives was:

1. Identify the elements of an experimental configuration. Different authors report different elements of an experimental configuration that can be changed in a replication. We need to establish exactly which elements of the experimental configuration can be changed in SE. To do this, we start with the changes proposed by the authors surveyed in Section 3. This results in the identification of groups of elements, which we refer to as dimensions. Then, we examine each dimension and, based on our knowledge of SE experimentation, identify the elements of the experimental configuration that can be changed in a SE replication. This step is detailed in Section 5.
2. Identify the different replications purposes. In this step, we set down the possible purposes that replication can serve in SE which we associate with the changes to each dimension of the experimental configuration. We have compiled the purposes of replication from the proposals by authors surveyed in Section 3 and the dimensions of the experimental configuration identified in Step 1 of this process. We identify the purposes of the replications in SE based on our experience. This step is detailed in Section 6.
3. Establish a classification of experiment replications for SE. In this step, we propose the replication types for SE experiments. We establish the possible types based on the combinations of changes to the dimensions of the experimental configuration identified in Step 1 of this process. The changes are used to understand the purposes of the replication identified in Step 2 of this process. This step is detailed in Section 7.

5. Elements of an experimental configuration

The similarity between a replication and the baseline experiment varies depending on the changes that are made to the experiment setup. To understand which changes a replication can accommodate, we need to identify the elements of which an experimental configuration is composed and which of these elements can be changed and still be considered the same experiment. We have identified the elements of a SE experimental configuration based on the Group II replication types listed in Table 2. We have examined the elements of the experimental configuration identified by different researchers as being involved in each replication type, which we used to identify groups of elements. These groups are called dimensions. It has been necessary to adapt the dimensions to SE. To do this, we have explored each dimension from the viewpoint of their applicability to SE experiments. In this manner, we have established the elements of the SE experimental configuration. Finally, we have examined the possibility of there being

elements that are considered relevant for SE experiments that did not show up in the survey.

We consider that an experimental configuration has four dimensions: operationalization, population, protocol and experimenter. There follows a description and justification of the rationale for each dimension. For details of the dimensions that appear in each replication type of the survey see Table 3.

Operationalization. Operationalization describes the act of translating a construct into its manifestation. In a controlled experiment, we have cause and effect constructs.

A dimension that might be referred to as operationalization is suggested by Adams et al. [1] in their operational and instrumental replications, Barker and Gurman [4] in their Type III replication, Finifter [20] in his virtual replication, Kelly et al. [36] in their operational replication, or Schmidt [61] in his direct replication.

In an experiment cause constructs are operationalized into treatments. Due to the immaturity of SE, one and the same treatment can be applied differently, and such differences can influence the results of an experiment. By applying techniques differently, the same cause construct is being operationalized differently. For example, one might define a testing experiment as a white-box vs. black-box experiment or as a path-coverage vs. equivalence partitioning experiment. There are a many ways to operationalize treatments for path-coverage vs. equivalence partitioning, although there are more options for white-box vs. black-box testing treatments. Replications need to examine the limits of SE cause construct operationalization. We identify the following elements of cause operationalizations that we believe are worth studying in a replication because they gauge how similar the replication is to the baseline experiment. Based on the literature survey state, which we have adapted to the peculiarities of SE, this would mean:

- *Treatment according to literature.* Selected source or sources (e.g., literature) detailing the SE methods used as treatments in the experiment. Different sources may suggest slightly different applications of the same treatment.
- *Treatment transmission aspects.* How treatments are conveyed to the subjects who are to apply them. Different training in the same treatment may convey a different understanding of the treatment to subjects, who may then apply the treatment differently.
- *Treatment as instructions.* Instructions given to subjects on how to apply treatments during experiment sessions. Different instructions for the same treatment may lead subjects to apply the treatments differently.

Additionally, according to the specific characteristics of SE, it is necessary to take into consideration:

Table 3
Issues described in each Group II replication type.

Author	Type	Changed-issue	Our dimension
Adams et al. [1]	Operational	Operationalization of criterion variable	Operationalizations (effect)
	Instrumental	Treatments or independent variables	Operationalizations (both)
Bahr et al. [3]	Type E	Subjects	Populations
	Type F	Place, subjects	Populations
	Type G	Time, subjects	Populations
	Type H	Time, place, subjects	Populations
Barker and Gurman [4]	Type II	Methodology and procedures	Protocol
	Type III	Different dependent variables	Operationalization (cause)
Beck [6]	Type II	Original design	Protocol
Blomquist [9]	–	–	–
Brown and Coney [12]	Replication and Replication with extension	Not specified	–
Finifter [20]	Virtual	Initial methodological conditions (measuring devices, samples used, research personal)	Operationalizations (effect) Experimenters
Fuess [22]	Replication with extension	Not specified	–
Hendrick [27]	Partial	Procedural variables	Protocol
Hyman and Wright [30]	Type I	Not specified	–
Kelly et al. [36]	Operational	Dependent variables	Operationalizations (effect)
	Instrumental	Experimental manipulations in the procedure	Protocol
La Sorte [40]	Retest	Few changes in the research design	Protocol
	Independent	Significant modifications into the original research design, include independent samples drawn from related or different universes by different investigators. Independent replications differ in design and purpose	Experimenters Protocol Populations
Lindsay and Ehrenberg [41]	Differentiated	Not specified	–
Lüdtke [44]	Independent	Site and researchers	Experimenters
Lykken [43]	Operational	Variations in method	Protocol
Radder [56]	Reproducibility of an experiment under a fixed theoretical interpretation	Different members	Experimenters
Schmidt [61]	Direct	Contextual background, participants, dependent variable	Populations Operationalizations (effect)
Sidman [64]	Systematic	Context, subjects' characteristics, experimenter	Populations Experimenters
Tsang and Kwan [69]	Empirical generalization	Another populations	Populations
	Generalization and extension	Procedures and populations	Populations Protocol

- *Treatment application procedure.* How the treatment (i.e. technique) is applied during the experiment. Subjects are given instructions on how to apply the treatment, where the treatment according to literature is embedded. Different treatment application procedures for the same treatment may lead subjects to apply the treatments differently.
- *Treatment resources.* Any sort of software or hardware used to apply treatments. A (semi)-automated versus a manual application of treatment, or its application using different tools may result in the treatment being applied differently.

Different applications of the same treatment might cause differences between replication and baseline results. Since the treatment is still the same, the replication can be considered the same experiment and not a new one.

Effect constructs are operationalized into metrics and measurement procedures. Replications should study the limits within which the results are unchanged when different metrics and measurement procedures are used for the same response variable. We identify the following elements for effect operationalizations:

- *Metric.* Operationalization that the experimenter defines to measure a response variable. For example, time to detect the first fault can be used as a metric of testing technique efficiency. But also total time to detect all faults. Different metrics of the same response variable may result in differences in the data collected.
- *Measurement procedure.* The way in which response variables are calculated and observations are allocated. This can be a simple procedure, whereby the response variable values are directly observed, or a complex procedure, whereby a series of tasks are used to calculate response variable values from the observations. For example, different procedures can be used to measure the time to detect the first fault metric: a manual procedure using a chronometer or an automated procedure using a program that records times in a transparent manner (i.e. hidden to the subject). Different measurement procedures for the same raw data may result in differences in the data collected.

Differences in the data collected might cause differences between replication and baseline results. Since the response variable is still the same, the replication can be considered the same experiment and not a new one.

Replications should be run to study cause and effect operationalizations and find out the bounds within which results hold.

Population. There are two types of populations in controlled SE experiments for which results should be verified in replications: *subjects* and *objects*. If the experiment is run with subjects, replications should study how sensitive results are to the properties of subjects.

A dimension that might be referred to as population is suggested by Bahr et al. [3] in their E, F, G and H replications, La Sorte [40] in his independent replication, Schmidt [61] in his direct replication, Sidman [64] in his systematic replication, or Tsang and Kwan [69] in their empirical generalization and generalization and extension replications.

Replications should also study the properties of experimental objects. Specifications, design documents, source code, programs are all examples of experimental objects. Replications examine the limits of the properties of experimental objects for which results hold. The properties to be studied depend on the object used in the experiment. For example, possible properties of programs used as experimental objects are: programming language, complexity, type of functionality, etc. Replications are useful for understanding the object type (i.e. small C programs) for which the results hold.

Protocol. Apparatus, materials, experimental objects, forms and procedures used in the experiment. The protocol is the configuration of all these elements used to observe a particular effect.

A dimension that might be referred to as protocol is suggested by Barker and Gurman [4] in their Type II replication, Beck [6] in his Type II replication, Hendrick [27] in his partial replication, Kelly et al. [36] in their instrumental replication, La Sorte [40] in his retest and independent replications, Lykken [43] in his operational replication, or Tsang and Kwan [69] in their generalization and extension replication.

The elements of the experimental protocol that can vary in a replication are:

- *Experimental design.* The experimental design specifies how groups are allocated to treatments. Different sources of variation can be controlled depending on the design. Changing the design in a replication explores different ways of observing the same effect by controlling different irrelevant and confounding variables.
- *Experimental objects.* Replications should explore different instances of the same type of objects to guarantee that the results hold not only for one object but for several objects of the same type.
- *Guides.* Replications must explore different instructions provided to subjects to guarantee that results do not depend on the guides provided.
- *Measuring instruments.* Instruments for collecting response variables like questionnaires or any output generated by subjects when performing the experimental task.
- *Data analysis techniques.* How the set of observations are analyzed statistically. Different replications can use different data analysis techniques to verify that results are not affected by data analysis.

Protocol elements should be changed in replications in order to verify that the observed results are real and not artifactual (due only to one particular experimental configuration).

Experimenters. People involved in the experiment. A dimension that might be referred to as experimenters is suggested by Finifter [20] in his virtual replication, La Sorte [40] in his independent replication, Lüdtke [44] in his independent replication, Radder [56] in his reproducibility of an experiment under a fixed theoretical interpretation, or Sidman [64] in his systematic replication.

SE experiments may be run by more than one person. The same experimenters may perform more than one role during an experiment. Likewise, different people may perform different roles. For example, one person may participate in the design of the experiment, another during its execution and yet another in analysis. A replication should verify whether the observed results are independent of the experimenters by varying the people who perform each role. This is why we view roles rather than experimenters as an experimental element. Of course, only one person may perform all roles. We propose five roles that may influence the results of an experiment based on the tasks performed by the experimenters: designer, trainer, monitor, measurer and analyst. Not all roles will necessarily be present in an experiment. For instance, the trainer role will only exist if training is necessary, or the measurer might be unnecessary depending on the type of metric used.

- The *designer* devises strategies for increasing control to minimize the validity threats to the experiment. Different designers may conceive slightly different variations of the same experimental design which might affect the experimental results.
- The *trainer* is responsible for transmitting treatments to the subjects during training sessions. Depending on their knowledge and experience, different trainers could explain the same

treatment differently to the experimental subjects. This may influence how well subjects understand a given treatment. Therefore, subjects might apply the treatment differently, which could cause differences in the experimental results.

- The *monitor* looks after the subjects during the experiment session. The monitor gives instructions to subjects about how the session is to be run, hands out materials to subjects, collects in the materials and answers questions. Different monitors may give different instructions during the experiment operation, which could make it easier or harder for the experimental subjects to perform the experimental tasks. For example, one monitor might decide not to answer any questions. Another might decide to answer questions related to the experimental objects but not to the treatments. A third one might decide to answer all kinds of questions. Unresolved questions about the task, the experimental objects or the treatments could lead subjects to do things differently and cause differences in the results.
- The *measurer* enacts the measurement procedure on the data gathered during the experiment. Unless it is automated, different measurers may apply the same measurement procedure slightly differently, which could lead to slight variations in the raw data. For example, given an experiment on test case design techniques, where the test data do not serve the purpose for which the test case was designed, different measurers could interpret the test case differently. One measurer might decide to interpret the test case literally, whereas another one might decide that the experimental subject made a slight mistake, as, according to the goal of the test case, the treatment was properly applied.
- The *analyst* conducts the statistical analyses. Different analysts may apply differently the same data analysis techniques and get different results. Consequently, they will arrive at different conclusions.

Different experimenters should participate in a replication to verify whether results are independent of experimenters. Varying experimenters rules out experimenter influence on the results, increasing the objectivity of the results.

Note that we have not considered site in the experimental configuration. Bahr et al. [3] in their type F and type H replications and Lüdtke [44] in his independent replication suggest place and site respectively as a change to be made to a replication. By reproducing results at other sites the observed effect is verified as not being dependent on the physical conditions of the experiment. We do not expect that site influences the result in SE experiments. The fact that experimental subjects are at physically different locations does not necessarily influence the results of applying a SE treatment. In SE experiments, site might be interpreted as a mixture of researchers and physical conditions of the experiment, as suggested by Brooks et al. [11]. We consider more appropriate to interpret the SE site as a combination of two of the above dimensions of the experimental configuration: experimenters and a new sample of the same (or different) populations of experimental subjects.

Additionally, we have not considered time (suggested by Bahr et al. [3] in their type G and H replications) and samples (suggested by Lüdtke [44] in his independent replication) in the experimental configuration. In SE, a replication implies that it is run with different subjects in a different moment of time, and therefore, these issues cannot be kept the same.

Table 4 shows the elements of the experimental configuration that replications should examine.

Table 4
Elements of an experimental configuration in SE.

Dimension	Element
Operationalization	Cause (Treatments) Effect (response variable)
Population	Subject properties Object properties
Protocol	Design Experimental objects Guides Instruments Analysis
Experimenter	Designer Trainer Monitor Measurer Analyst

6. Replication functions

The general function of replication is to verify the results of an experiment. However, the fact that there are different types of replication implies that each one should have its own special purpose. Some authors (Schmidt [61] and Bahr et al. [3]) state that the changes to the elements of the experimental configuration are governed by what checks are to be run.

Schmidt [61] explicitly suggests that replications serve several different and more specific purposes depending on the changes that are introduced. The confirmatory power of replication increases with every difference, generalizing the phenomenon to a larger area of application. For example, replicating an experiment as closely as possible verifies that results are not accidental, whereas varying the population properties of the baseline experiment verifies the types of populations for which the results hold. For example, by altering subject experience we can study if experience is a condition that affects results. We understand that similarity between the baseline experiment and a replication in SE should vary depending on the verification purposes.

Based on the ideas of Schmidt [61] and Bahr et al. [3] and how the dimensions of the experimental configuration described in Section 5 are changed, we have identified six verification functions for SE experiment replications. Schmidt defines five functions, while Bahr et al. define only two. The equivalences are shown in Table 5. The functions we have identified are:

1. *Control sampling error.* The goal of such replications is to verify that the results of the baseline experiment are not a chance product of a type I error. For this verification purpose, all dimensions of the experimental configuration must resemble the baseline experiment as closely as possible. Additionally, they increase the sample size (number of observations) and provide an understanding of the natural (random) variation of the observed results. This is critical for being able to decide whether or not results hold in dissimilar replications.
2. *Control protocol independence.* The goal of such replications is to verify that the results of the baseline experiment are not artifactual. An artifactual result is due to the experimental configuration and cannot be guaranteed to exist in reality. Artifactual results could exist and under certain experimental conditions in the lab only. In other words, it is necessary to verify that reality and not an artifact is being observed. The experimental protocol needs to be changed for this purpose. If an experiment using particular materials is replicated several times using the same materials, the observed results may occur for those materials only but not for equivalent materials. Similarly, results observed in replications that are run using the same, defective measuring instrument could be influenced by that measuring instrument. The instrument anomaly can be detected if a different (but equivalent) instrument is used in some replications.

3. *Understand operationalization limits.* The goal of such replications is to learn how sensitive results are to different operationalizations. Variations in results in response to changes to treatment application procedures, treatment instructions, resources, and treatment transmission should be analyzed to verify cause operationalizations. On the other hand, changes to effect operationalizations verify if results hold using different metrics to measure the same construct or different (but equivalent) measurement procedures.
4. *Understand populations limits.* The goal of such replications is to learn the extent to which results hold for other subject types or other types of experimental objects. That is, learn to which specific population belongs the experimental sample and which are the characteristics of such population. Several replications that change different protocol elements are necessary to verify that the observed effects are not due to a particular protocol element or combination of elements.
5. *Control experimenters independence.* The goal is to verify that the results of the baseline experiment are independent of the experimenters. To do this, experimenters must be changed.
6. *Validate hypotheses.* Additionally, an experiment should also be replicated by retaining no more than the hypothesis. The aim here is to observe the same result using different experiments with identical goals.

The function of a replication varies depending on the changes made to the experimental configuration. Table 6 shows how the different replication types contribute to results verification.

We relate replication functions to the different types of experiment validity mentioned in [17]. If a replication is carried out to control sampling error, it increases conclusion validity (function 1). If a replication is run to control protocol independence, it increases internal validity (purpose 2). When a replication is carried out to understand operationalization limits, it increases construct validity (purpose 3). Finally, if a replication is run to understand population limits, it increases external validity (purpose 4).

Table 7 shows the different replication functions and dimensions of the experimental configuration that vary depending on the purpose of the replication.

Two types of changes go beyond the validity of the experiment itself. Changing the experimenters does not control any threat due to the experiment setup. This change controls a bias on results due to the experimenter who is obtaining the results. Changing all dimensions seeks a higher level of verification, a kind of double-checking finding some results with a different research method.

Replications provide different knowledge depending on the changes to the baseline experiment. When an experiment that is replicated identically at the same site by the same experimenters corroborates results, we learn that the observed result is not a chance outcome and gain knowledge of the random natural variation of the phenomenon under study; if the replication does not corroborate results, the observed result could be due to chance (more replications are needed). When replications with protocol changes corroborate results, we learn that results match reality, that is, the results are not artifactual; if the replications do not corroborate results, the results might be due to certain experimental conditions. When changes are made to the operationalizations of constructs, we learn the operationalization limits within which results hold or do not hold. When changes are made to the population properties, we learn the population properties that might or might not have a bearing on the results. When different replications with changes to the experimenters corroborate the results, we learn that they are not the product of the experimenters; if the replications do not corroborate results, the results might be biased by the experimenters.

7. SE experiment replication types

Now that we have identified which dimensions of an experimental configuration can be changed for verification purposes, as well as their replication functions, we can proceed with the generation of a replications classification for SE experiments.

First let us define the limits of a replication. The replication types range from not making any change to the baseline experi-

Table 5
Equivalences between our functions of replication and survey authors.

Author	Function of replication					
	Control sampling error	Control experimenters independence	Control protocol independence	Understand operationalizations limits	Understand populations limits	Validate hypotheses
Schmidt [61]	Control for sampling error	Control for fraud	Control for artifacts	Control for artifacts	Generalize results to a larger or to a different population	Verify the underlying hypothesis of the earlier experiment
Bahr et al. [3]	Check the findings of an earlier study	Assess whether they hold under altered conditions (the generality test)				

Table 6
Knowledge gained and validity threats addressed based on changes to baseline experimental configuration.

Dimension	Knowledge gained if changed	Knowledge gained if not changed	Validity threat addressed
None	Not applicable	Event not due to type I error Understanding of natural variation of results	Conclusion validity
Protocol	Real event	Artifactual event	Internal validity
Operationalization	Known operationalization limits	Event limited to this way of applying techniques and measuring results	Construct validity
Population	Known populations	Unknown population limits	External validity
Experimenters	Objective (inter-subjectively testable) event	Subjective event	Beyond the experiment threats
All of the above at the same time	Result observed using different experiments with identical goals	Not applicable	Beyond the experiment threats

Table 7
Functions of replication and changed dimensions.

Experimental configuration	Function of replication					
	Control sampling error	Control experimenters independence	Control protocol independence	Understand operationalizations limits	Understand populations limits	Validate hypotheses
Operationalization	=	=	=	≠	=	Unknown
Population	=	=	=	=	≠	Unknown
Protocol	=	=	≠	=	=	Unknown
Experimenter	=	≠	=	=	=	Unknown

LEGEND: = all the elements of the dimension are equal to, or as similar as possible to, the baseline experiment.
≠ some (or all) elements of the dimension vary with respect to the baseline experiment.
Unknown: the elements of the dimension are unknown.

ment to changing all the dimensions of the experimental configuration. However, we should establish a maximum level of change as of which a replication should be considered a new experiment. Our view is that a replication should:

- *Execute an experiment.* This omits activities that some authors define as replication types like reanalyzing existing data using the same procedures, different procedures, or different statistical models to the baseline experiment.
- *Retain at least some of the hypotheses of the baseline experiment.* Specifically, at least two treatments and one response variable need to be shared. Other treatments or response variables could be added. Note that if the exact hypotheses are not retained (treatments or response variables are added or removed, although keeping two treatments and one response variable in common), only a subset of the replication will be comparable with the baseline experiment (i.e. the part that corresponds to the treatments and response variables shared by the replication and baseline experiment).

Table 8 shows and describes the proposed types of replication. We propose to identify a replication by the dimensions of the experimental configuration that have been changed: protocol, operationalizations, populations, or experimenters. Based on changes to these four dimensions, we can establish the following three types of replications:

- *Literal.* The aim of a literal replication is to run as exact a replication of the baseline experiment as possible. The elements of all dimensions in the experimental configuration are kept unchanged: the replication is run by the same experimenters using the same protocol, the same operationalizations and different samples of the same populations. No deliberate changes are made to the experimental configuration; any change is made inadvertently.
- *Operational.* The aim of operational replications is to vary some (or all) of the dimensions of the experimental configuration. The replication could be run either by varying one or more experimenters, using a different protocol or operationalizations, or using a different population. In an operational replication, one or more of the dimensions of the experimental configuration may vary at the same time.
- *Conceptual.* In conceptual replications, experimenters have "nothing more than a clear statement of the empirical fact" [43], which the previous experimenter claims to have established. Consequently, new protocols and operationalizations are used by different experimenters to verify the results observed in the baseline experiment.

We propose to label a replication by the dimensions of the experimental configuration that have been changed (see last column of Table 9):

- *Literal and conceptual replications.* These two replication types go by the names of repetition and reproduction in the literature. We have tried to respect this idea. On the other hand, these two replication types have a totally different procedure to operational replications. Operational replications are based on the baseline experiment, which is modified to introduce the appropriate changes. This does not happen in either repetition or reproduction. No change is made in repetitions (they are new runs of the experiment on another sample of the same population), whereas reproductions retain no more than the hypotheses of the original experiment.
- *Operational replications.* A specific replication of this type can be defined by concatenating the different properties that have been changed in the replication. For example, a changed-population/-experimenter replication is run by different experimenters on a different population, with the same protocol and operationalizations.

8. Systematic replications

By making changes depending on the purpose of the verification defined by those very changes, we can increase confidence in a result not being artifactual, explore population bounds, discover relevant population, etc.

Some changes are made necessary by the new context. For example, the replication context in which a replication is run may oblige changes to the experimental design. For example, whereas the baseline experiment was run over a three-day period and all subjects applied three techniques, time is shorter in the new context, the experiment is run on one day and one third of subjects apply each technique. Far from being a threat to results verification, this major change to the experiment may turn out to be an opportunity for checking whether a particular experimental protocol element (design) affects the results.

If this is the first replication of the experiment and the results are different, the change cannot be reliably attributed to the design (unknown variables that have been accidently changed could be at work). In the long term, though, as more replications are run, this replication will play its role in verifying results.

However, if the baseline experiment had been repeated several times by the original experimenters (literal replications), there would already be a good estimation of the natural random variation of the results. This would be beneficial for the new operational replication to be able to help verify the results. If the baseline experiment has a larger sample of results thanks to the repetitions run by the same experimenters at their own laboratories, what appeared to be differences between the results of the new replication and baseline experiment might actually be within the bounds of the natural random variation of the results.

Systematically varying replications helps to increase the understanding of the conditions that may influence results. The replica-

Table 8
Experiment replication types proposed for SE experiments.

Replication type	Dimension	Description
Literal		SE's equivalent of what is defined as an exact replication in other disciplines. The aim is to run as exact a replication as possible of the baseline experiment. The replication is run by the same experimenters using the same protocol and the same operationalizations on different samples of the same population
Operational	Protocol	The experimental protocol elements are varied with the aim of verifying that the observed results are reproduced using equivalent experimental protocols
	Operationalization	The cause and/or effect operationalizations are varied in order to verify the bounds of the cause and/or effect construct operationalizations within which the results hold
	Population	The populations are varied to verify the limits of the populations used in the baseline experiment
	Experimenter	The experimenters are varied to verify their influence on the results
Conceptual		Different experimenters run the replication with new protocol and operationalizations

Table 9
Names proposed for SE experiments replication types.

Replication type	Protocol	Operationalizations	Populations	Experimenters	Replication name
Literal	=	=	=	=	Repetition
Operational	=	=	=	≠	Changed-experimenter
	=	=	≠	=	Changed-populations
	=	=	≠	≠	Changed-populations/-experimenters
	=	≠	=	=	Changed-operationalizations
	=	≠	=	≠	Changed-operationalizations/-experimenters
	=	≠	≠	=	Changed-operationalizations/-populations
	=	≠	≠	≠	Changed-operationalizations/-populations/-experimenters
	≠	=	=	=	Changed-protocol
	≠	=	=	≠	Changed- protocol /- experimenters
	≠	=	≠	=	Changed- protocol /-populations
	≠	=	≠	≠	Changed- protocol /-populations/-experimenters
	≠	≠	=	=	Changed- protocol /-operationalizations
	≠	≠	=	≠	Changed- protocol / -operationalizations/- experimenters
	≠	≠	≠	=	Changed- protocol /- operationalizations/- populations
	≠	≠	≠	≠	Changed- protocol /- operationalizations/- populations/-experimenters
Conceptual	Unknown	Unknown	Unknown	Unknown	Reproduction

tion types proposed for different verification purposes gain power if they are used through systematic replication.

The idea of systematically varying replications is not new. Hunt [28] suggested that a better procedure for running replications would be to systematically modify an element of the original experiment in each replication in order to study whether this change influences the results. As Hendrick [27] states, rigorously applied systematic replications can change negative attitudes toward replications. Similarly, Rosenthal [59] proposes running series of at least two replications, one more or less similar to the baseline experiment and the other moderately dissimilar.

Experimenters following a systematic approach can gradually verify results easier. The conditions that influence the experiments that we run in SE are not known. Using a systematic approach, where deliberate changes are made to replications in order to understand the dependence on the experimental protocol, the sensitivity to operationalizations and relevant population features would help to increase knowledge of experimental conditions (and, incidentally, of relevant software development variables).

After just one experiment, we do not know whether the observed results are a product of chance. The first step toward verifying the results is to repeat the experiment. In a repetition, the same experimenter verifies the observed result at the same site, using the same protocol, with the same operationalizations, on different samples of the same populations. The repetition of an experiment helps to determine the natural variation of the observed results, that is, the confidence interval within which the results are observed, reducing the likelihood of type I error.

After several repetitions, we do not know whether the results are artifactual or real. The observed events can be a product of the experimental setup. Once we have observed a pattern in literal replications, we can move on to verify whether results are or are not a product of the protocol. It is time to run operational replications that vary the experimental protocol.

If results observed in this series of replications are reproduced, they can be considered to be independent of the protocol, that is, the observed events are the result of a causal relationship in the real world.

After having identified real behavior patterns, it is possible to vary the populations and operationalizations to find out the bounds within which results are reproducible. By varying populations, new knowledge is obtained about critical population features. By varying operationalizations, new knowledge is obtained on the "active ingredients" of the factors (independent variables) under study.

Now we have several replications in which elements of the experimental protocol and properties of the populations and operationalizations have varied. We learn the regular natural variation of the results (confidence interval), as well as the conditions and bounds within which an experimental result is reproducible from these replications. This knowledge gives a better understanding of the results that external and foreign replications will produce.

When conditions influencing the experimental results are better known, it is possible to run a conceptual replication (reproduction) in order to verify results through different experiments. Reproduction is a more powerful way of confirming an experimen-

tal result, as the result is verified by different experimenters, at different sites, using a different protocol, with different operationalizations and on equivalent populations to the baseline experiment. However, this is the most risky type of replication, as, unless the results are similar, it is impossible to identify what causes the results of the replication and baseline experiment to differ. Hence, a less expensive and more reasonable approach is to start with minor changes and leave the major variations until the end of the process.

The possible threat of errors being propagated by the original and the replicating experimenters exchanging materials [38] is not such a serious thing, precisely because other replications that alter the design and other protocol details should be performed in order to assure that these elements are not influencing the results. Replications with identical materials and protocols (and possibly the same errors) are useful as a first and necessary step for verifying that an identical experiment reproduces results at another site by means of an identical experiment. Later replications will check whether the results are induced by the protocol. It is worthwhile replicating first with an identical protocol and materials (and exchanging experimental packages among experimenters) and then with different protocols in order to be able to identify sources of variation if the results do not hold (site in the first case and protocol in the second).

Obviously, a joint replication does not serve the purpose of verifying whether the results are independent of the researchers. However it is useful for finding out whether the replication is independent of the site. Again, if we do both things at the same time and the failure to interact and transmit tacit knowledge leads to unintentional changes, it will be impossible to decide what the source of the variation in the results is.

Therefore, the exchange of experimental packages or interaction among experimenters does not invalidate a replication. Quite the contrary, they produce two types of replication that are necessary and play a role in verifying results. However, they are not the only possible modus operandi; other replication types where materials are not exchanged and experimenters do not communicate are also necessary and play their own role in verifying results.

9. Comparing the proposed classification with other SE classifications

This section compares the proposed classification with other classifications existing in SE and reviewed in Section 2. Table 10 compares the replication types presented in the three works that propose SE replication classifications with the replication types that we propose.

We have classified Basili et al.'s [5] strict replications as our literal replications (in Basili et al.'s [5] words "duplicate as accurately as possible the original experiment"); their replications that vary the manner in which the experiment run are equated to our changed-protocol replications ("they test the same hypotheses but alter the details in order to address internal threats"); their replications that vary variables that are intrinsic to the object of study and focus of the evaluation are equivalent to our changed-operationalizations replications ("change intrinsic properties of the process and the ways in which the response variable is measured respectively"); their replications that vary context variables in the environment in which the solution is evaluated are catalogued as our changed-populations replications ("help to identify potentially important environmental factors and thus help understand its external validity"). We have not been able to classify the replications that extend theory, as they seem to refer more to the size of the change than to a definite change type. Finally, Basili et al.'s [5] replication types do not state whether there is any liaison between the replicating and original experimenters or whether replications with more than one type of change are possible.

Almqvist's differentiated-improved replications [2] can be equated to our operational replications with changed protocol and/or operationalizations and/or populations, and to our conceptual replications. We have classified his similar replications as our literal and changed-experimenters operational replications, since neither the protocol, nor the operationalizations or populations can be changed. Finally, Almqvist's [2] external/internal categories are comparable with our replications with or without changed experimenters. Therefore, we have classified Almqvist's [2] close-internal replications as our literal replications; his differentiated-internal replications as operational replications in which everything may vary except experimenters; his similar-external replications as changed-experimenters operational replications; and his differentiated-external as changed-experimenters operational replications in which the other dimensions may also vary.

Regarding Krein and Knutson's classification [39], we have classified their strict replication as our literal replication (in their words "replicate a prior study as precisely as possible"); we have equated their differentiated replication to our operational replication ("alters aspects of the prior study in order to test the limits of that study's conclusions"); their dependent replication is comparable to what we categorize as a literal, changed-protocol or changed-populations operational replication, whereas their independent replication is equivalent to our conceptual replication ("[it] is conducted without knowledge of, or, deference to, that prior study – either because the researchers are unaware of the prior work, or because they want to avoid bias").

Additionally, Table 11 compares the replication types mentioned in other SE works, albeit not for classification purposes, to the replication types that we propose.

Brooks et al.'s [11] and Mendonça et al.'s [48] internal replications can be equated to both our literal replication and any of our operational replications in which experimenters are unchanged, and their external replications are equivalent to our conceptual and changed-experimenter operational replications. Notice that the terms external and internal replications, originally used by Brooks et al. [11], have spread throughout experimental SE literature. It is surprising that we have found no reference to these two terms having been sourced from replication classifications in other disciplines[2].

We see Shull et al.'s [63] literal replications as our literal replications; their conceptual as our conceptual replications; their independent as our changed-experimenters operational replications; and their dependent replications as our operational replications that do not change experimenters.

Mandić et al.'s [45] exact/partial replications are comparable with our literal replications, and their replications that improve the original experiment are equivalent to our operational and conceptual replications.

10. Discussion about replication in SE: comparing researcher positions

This section examines, from the viewpoint of the proposed classification, some issues about experiment replication in SE that have been under debate for a long time.

There are currently two opposing views concerning SE replication. One of these standpoints is represented by the papers referenced in Section 9 and is totally compatible with the proposal

[2] Note that the publications that we have examined are listed in Section 2 (reporting the results of our literature survey) and refer to classifications of replications. This means that there could be articles that are not classifications of replications reporting internal or external replications.

Table 10
Comparison of SE replication classifications and the proposed classification.

Replication type	Prot.	Oper.	Popul.	Exp.	Replication name	Basili	Almqvist	Krein & Knutson
Literal	=	=	=	=	Repetition	Strict	Similar-internal	Strict
Operational	=	=	=	≠	Changed-experimenters	None	Similar-external	Differentiated-dependent
	=	=	≠	=	Changed-populations	Vary context variables	Improved-internal	Differentiated-dependent
	=	=	≠	≠	Changed-populations/-experimenters	None	Differentiated-external	Differentiated-dependent
	=	≠	=	=	Changed-operationalizations	Vary intrinsic object study Vary intrinsic focus evaluation	Improved-internal	Differentiated-dependent
	=	≠	=	≠	Changed-operationalizations/-experimenters	None	Differentiated-external	Differentiated-dependent
	=	≠	≠	=	Changed-operationalizations/-populations	None	Improved-internal	Differentiated-dependent
	=	≠	≠	≠	Changed-operationalizations/-populations/-experimenter	None	Differentiated-external	Differentiated-dependent
	≠	=	=	=	Changed-protocol	Vary the manner in which experiment is run Vary context variables	Improved-internal	Differentiated-dependent
	≠	=	=	≠	Changed-protocol/-experimenters	None	Differentiated-external	Differentiated-dependent
	≠	=	≠	=	Changed-protocol/-populations	Vary context variables	Improved-internal	Differentiated-dependent
	≠	=	≠	≠	Changed-protocol/ -populations/-experimenters	None	Differentiated-external	Differentiated-dependent
	≠	≠	=	=	Changed-protocol/-operationalizations	None	Improved-internal	Differentiated-dependent
	≠	≠	=	≠	Changed-protocol/-operationalizations/-experimenters	None	Differentiated-external	Differentiated-dependent
	≠	≠	≠	=	Changed-protocol/-operationalizations/-populations	None	Improved-internal	Differentiated-dependent
	≠	≠	≠	≠	Changed-protocol/-operationalizations/-populations/-experimenters	None	Differentiated-external	Differentiated-dependent
Conceptual	Unk.	Unk.	Unk.	Unk.	Reproduction	None	Differentiated-external	Independent
						Extend theory (size of change)	–	–

presented here. This current holds that there are different types of replications. On this ground, we will refer to as the *multiple-replication types approach*. As regards the other current, which is also quite widespread, there are not many publications stating its standpoint, even though it is often espoused in discussions at conferences, in reviewer comments, etc. The only publications that we have found representing this line of thought were published by Miller [51] and Kitchenham [38]. This current holds that SE replications should be confined to independent replications. These replications have in common with the original experiment the underlying hypothesis only; all the elements of the experimental configuration in the replication are different from the original experiment. These are what our classification denotes as conceptual replications or reproductions. As this current stands for only one type of replication, we will refer to as the *single-replication type approach*. The arguments upheld by the single-replication type approach to defend their view are:

- **The single-replication type approach avoids the possibility of error propagation.** When running the baseline experiment, researchers might make errors out of ignorance or by mistake. The single-replication type approach defends that the only way to stop this is by having other researchers do the replication without any interaction whatsoever (not even the exchange of materials) with the baseline experiment researchers. To be precise, they show the following situations:
  (1) The researcher is unaware of the validity threats of his/her design.
  (2) The preferred treatment is assigned to the best subjects.
  (3) The experimenter does not know how to analyze the data properly.
  (4) The experimenter unconsciously biases the results because he/she prefers one of the treatments

According to our proposal, however, researchers can check or learn about several issues if a replication shares materials with the baseline experiment or even if some of the researchers from the baseline experiment participate in the replication. To identify shortcomings in cases 1, 2 and 3, the researchers do not necessarily have to be different. A replication using a different design would identify the problems in cases 1 and 2. A replication changing the data analysis would show up the problem in case 3.

But there are circumstances in which it is strictly necessary to use different researchers, for example, in case 4, as the validity threats are intrinsic to the researchers. However, our approach proposes that researchers should be changed at some point during the replication process.

- **The single-replication type approach prevents invalid results from being repeated.** The single-replication type approach defends that invalid results due to possible errors made in the original experiment (for example, the choice of an unsuitable

Table 11
Comparison of other SE replication classifications and the proposed classification.

Replication type	Prot.	Oper.	Popul.	Exp.	Replication name	Brooks et al. & Mendonça	Shull et al.	Lung et al.	Mandić
Literal	=	=	=	=	Repetition	Internal	Exact	Literal	Exact/partial
Operational	=	=	=	≠	Changed-experimenters	External	Independent	None	Improve original exper.
	=	=	≠	=	Changed-populations	Internal	Dependent	None	Improve original exper.
	=	=	≠	≠	Changed-populations/-experimenters	External	Independent	None	Improve original exper.
	=	≠	=	=	Changed-operationalizations	Internal	Dependent	None	Improve original exper.
	=	≠	=	≠	Changed-operationalizations/-experimenters	External	Independent	None	Improve original exper.
	=	≠	≠	=	Changed-operationalizations/- populations	Internal	Dependent	None	Improve original exper.
	=	≠	≠	≠	Changed-operationalizations/-populations/-experimenter	External	Independent	None	Improve original exper.
	≠	=	=	=	Changed-protocol	Internal	Dependent	None	Improve original exper.
	≠	=	=	≠	Changed-protocol/-experimenters	External	Independent	None	Improve original exper.
	≠	=	≠	=	Changed-protocol/-populations	Internal	Dependent	None	Improve original exper.
	≠	=	≠	≠	Changed-protocol/ -populations/-experimenters	External	Independent	None	Improve original exper.
	≠	≠	=	=	Changed-protocol/- operationalizations	Internal	Dependent	None	Improve original exper.
	≠	≠	=	≠	Changed-protocol/-operationalizations/-experimenters	External	Independent	None	Improve original exper.
	≠	≠	≠	=	Changed-protocol/-operationalizations/-populations	Internal	Dependent	None	Improve original exper.
	≠	≠	≠	≠	Changed-protocol/-operationalizations/-populations/-experimenters	External	Independent	None	Improve original exper.
Conceptual	≠	≠	Unknown	≠	Reproduction	External	Conceptual	Theoretical	Improve original exper.

experimental design) will not be propagated thanks to the fact that the experimental configuration of a conceptual replication is completely different. However, this not absolutely true. Any replication could very well repeat invalid results. The researchers running the replication could, by chance, make the same error made by the baseline experiment researchers (or even different ones that could repeat the same invalid results). This is an unavoidable possibility in any type of replication. The only way of making out the correct results is running a large number of replications of different types.

Our approach suggests that replications should make small changes to the original experiment. Thanks to this iterative approach, it will be possible to **empirically** identify what elements caused the incorrect results. For example, an inappropriate experimental design[3] will eventually be detected when the results of the replications using other designs match up and yield different results than the baseline experiment. Using our approach, there is a possibility of spurious results being propagated across replications, but we do not regard this as a problem because the replication series (which we propose) will detect this circumstance (and researchers will learn things about the phenomenon under study).

In case a researcher suspects that a design is inappropriate, our approach suggests to use another design to replicate the experiment in order to **empirically** demonstrate which variables the original design failed to control. The design appraisals output by this procedure will be more useful than theoretical criticisms as they will uncover other relevant variables (whose existence the discerning researcher **suspected**).

[3] Additionally, we believe that there is no such thing as a perfect experimental design. Different designs have different validity threats. According to Christensen [15], "...we can never be certain that complete control has been effected in the experiment. All we can do is increase the probability that we have attained the desired control of the confounding extraneous variables that would be sources of rival hypotheses". Therefore, we do not think that a design should be rejected outright until the experiment has been replicated using other designs.

- **The single-replication type approach is useful for planning experiments to test reasoned hypotheses about what could be happening if the results are not confirmed.** This conception of experimentation is not exclusive to the single-replication type approach. Several authors ([14,31,33,34], and us in this research) claim that the results of a replication do not put an end to the replication process; rather they are the start of a learning process about variables that are possibly influencing the phenomenon under study. Replication is really an exercise in questioning and hypothesizing about why things happen, conjectures which are then tested by running experiments.
- **The single-replication type approach avoids confirmation bias.** The single-replication type approach holds that a failure to falsify is more convincing than verifying results to avoid confirmation bias. Looking at other branches of science, we find that the falsification approach is naïve. This view is fine for mature experimental topics. But, as highlighted by many authors (for example, Brinberg and McGrath [10] or Collins [16]), the aim at the start of the research (early replications) is not to falsify but to learn which variables are influencing the results: "In the early stages, failure to get the expected results is not falsification, but a step in the discovery of some interfering factor. For immature experimental knowledge, the first step is [...] to find out which experimental conditions should be controlled".

Thompson and McConnell's experiments clearly illustrate this view. In their experiments with flatworms, Thompson and McConnell found out that by cutting a planarian across the middle into head and tail sections, each part would not only regenerate its missing half, but also retain what it had previously learned. The regenerated tails showed as much retention—and in some cases more—than the regenerated heads [67]. These results led them to think more seriously about the chemical nature of memory. To test this notion, they transferred the putative molecules from a trained to an untrained animal, by using cannibalistic worms. They fed pieces of trained worm to hungry untrained worms, obtaining promising results [47]. However, Halas et al. [24] ran several repli-

cations of this experiment and were unable to confirm Thompson and McConnell's results. As they ran more replications of the original experiment Thompson and McConnell became aware of a range of conditions that influenced the result of their experiments. At some point during this research (after several hundreds of experiments over more than ten years), McConnell managed to detect around 70 conditions that influence the behavior of flatworms [46].

- **The single-replication type approach is viable and efficient.** In the literature review that we have conducted, we have not found anything to support the claim that it is not viable or efficient to change one thing at a time. In fact, Thomson and McConnell's experiments with flatworms ([46]) suggest that the right thing to do is to make small changes. Additionally, Hendrick [27], Hunt [28] and Rosenthal [59] support this idea.

Besides, it is not at all advisable to completely change the original experiment during replication (as suggested by the single-replication type viewpoint), because, if so, it would be impossible to find out why inconsistent results are not consistent [8,41]. Only by means of a series of controlled changes would it be possible to identify the variables interfering with the phenomenon under study. According to Collins [16], "The less that is known about an area, the more power a very similar experiment has ... This is because, in the absence of a well worked out set of crucial variables, any change in the experiment configuration, however trivial in appearance, may well entail invisible but significant changes in conditions". Or, according to Brinberg and McGrath [10], "Most aspects are unknown when we start to study a phenomenon experimentally. Even the tiniest change in a replication can lead to inexplicable differences in the results".

- **The single-replication type approach obtains conclusive results quickly.** It does not take as long to get conclusive results if the replications are completely different from the original experiment.

It may appear that it takes longer to get conclusive results using our proposal, because we suggest that small changes should be made to the original experiment, and therefore it takes a lot of replications to arrive at conclusive results. However, we believe that SE has not yet grasped how difficult it is (and therefore how long it takes) to establish a scientific fact. It takes years if not decades to achieve conclusive results in other disciplines. Unfortunately, scientific progress is slow. Other disciplines illustrate this point:

- The builders of the Transversely Excited Atmospheric (TEA) laser [16] could not explain, based on their understanding of why the laser was supposed to work, why their laser worked but a replication of it did not work. Around 10 years (and many replications) later, it was revealed that their understanding of why their laser worked was incomplete.
- Bisell [8] claims that it takes her from four to six years, and at times much longer, to get results conclusive enough to be able to write a paper about her experiments on the roles of the microenvironment and extracellular matrix in cancer.
- Moyer [54] tells how recent studies have found that the benefits attributed to fish oil supplement at the end of the 20th century are not true. It has taken over 10 years to arrive at this result.

11. Conclusions

Replication plays an important role in scientific progress. In science, facts are at least as important as ideas [29]. Experiment replication is necessary to identify facts. To build an empirical body of knowledge in SE, it is necessary to run several types of replication. It is necessary to understand that different types of replication are able to verify different aspects of the results and what these aspects are.

Replications can either use the same or vary the elements of the experimental configuration of the baseline experiment. It is just as valid to reuse as it is to vary the baseline experimental configuration or for original and replicating experimenters to run the replication independently or jointly. Each of these approaches to replication serves to verify a particular aspect.

The proposed replication classification should give experimenters guidance about what different types of replication they can run. A classification of replication types helps experimenters to plan and understand their replications. Depending on the changes, experimenters can opt for different replication types.

By systematically performing different types of replication, experimental results can be gradually verified. This furthers and speeds up the generation of pieces of knowledge. Considering the state of the practice in experimental SE, a systematic approach is the best replication process.

The proposed classification helps to accommodate opposing views within a broader framework. Thus, our classification is capable of accounting for replications as far apart as Basili et al.'s strict replications [5] or replications that retain no more than the hypothesis, which Kitchenham [38] or Miller [51] suggest are the only replications that are of any use. According to the findings reported in this paper, such contrary stances are really tantamount to different types of replication conducted for different purposes. The proposed classification embraces different ways of running replications that are useful for gradually advancing toward verified experimental results.

Acknowledgment

This work was supported by research grant TIN2011-23216 of the Spanish Ministry of Economy and Competitiveness, and research grant 206747 of the Mexico's National Council of Science and Technology (CONACyT).

Appendix A. Literature search details

We selected the terms replication, classification, kind, type, typology and taxonomy for use as keywords in the search string.

The string was compiled so that Scopus® searched the term 'replication' plus one of the other terms in article titles, abstracts and keywords, i.e. TITLE-ABS-KEY(replication AND (classification OR kind OR type OR typology OR taxonomy)). The terms were specified in the singular as Scopus® automatically searches singular and plural terms. This search string returned 46,783 documents.

As the search returned such a huge number of results, it had to be refined. Our experience has shown that shorter strings (in this case, two-term strings) help to more efficiently identify relevant documents, because they require less effort (as fewer documents are returned) to assess the unpromising terms that are then not used (as opposed to mixing promising and unpromising terms in a long query and obtaining an enormous number of documents as a result). We divided the search string into five two-term strings: the term replication plus one of the other five terms (classification, kind, type, typology and taxonomy). Table 12 shows the number of documents returned by these new search strings. Search number 3 returns more results because the term "type" is more common (and, as such, not such a good keyword) than the other terms.

According to the search string configuration, Scopus® locates the pairs of terms in the titles, abstracts and keywords of the documents irrespective of how far apart they are. This means that

Table 12
Search strings using the term replication.

Search	Search String	Documents
1	TITLE-ABS-KEY(replication AND classification)	2.541
2	TITLE-ABS-KEY(replication AND kind)	1.290
3	TITLE-ABS-KEY(replication AND type)	43.592
4	TITLE-ABS-KEY(replication AND typology)	80
5	TITLE-ABS-KEY(replication AND taxonomy)	289

Table 13
Search strings using the term reproduction.

Search	Search string	Docs.
1	TITLE-ABS-KEY(reproduction AND classification)	2.381
2	TITLE-ABS-KEY(reproduction AND kind)	1.888
3.1	TITLE-ABS-KEY(reproduction AND type)	14.759
3.2	TITLE-ABS-KEY(reproduction W/2 type)	466
4	TITLE-ABS-KEY(reproduction AND typology)	141
5	TITLE-ABS-KEY(reproduction AND taxonomy)	819

many of the resulting documents do not mention or discuss how to run a replication. Even so, we considered that it was worthwhile examining the titles, abstracts and keywords of searches 1, 2, 4 and 5, as their size is reasonable; not so, the results of search 3. We refined string 3 by limiting the distance between the terms "replication" and "type". We used the proximity operator 'within'. 'Within' (represented as W/d) searches a pair of terms within a particular distance "d" irrespective of which term comes first. In this new string, we specified a maximum distance of two terms between the words "replication" and "type" (TITLE-ABS-KEY(replication W/2 type)). This new search returns "replication type", but also "type of replication" (one-word separation among terms) or "replication of some type" (two-word separation among terms). The search using this new string returned 3143 documents. This is a more manageable size for inspecting titles, abstracts and keywords to identify relevant articles.

The five search strings used (shown in Table 12, except that search 3 is replaced by the string containing the proximity operator 'within') returned 7343 documents.

Replication is the keyword used in our searches. However, the terms reproduction and repetition are sometimes used to denote a replication. To locate more replication classifications, we reran the searches using the same search strings with the terms reproduction and repetition in place of replication. Tables 13 and 14 show the number of documents returned for these searches. We refined the 'reproduction AND type' and 'repetition AND type' strings as we did for the 'replication AND type' string.

We examined the title, abstract and keywords of the resulting documents for the reproduction and repetition searches, but we did not find any papers mentioning different ways of running a replication. For example, we found that the term reproduction is often used in combination with the other terms (classification, kind, type, typology and taxonomy) to refer to the process whereby living beings engender other living beings, and the term repetition is used, albeit less so, in areas akin to linguistics. Both terms are used to a greater or lesser extent to denote concepts other than experiment replication in different branches of science. Therefore, we suggest that, in SE, we do not use repetition or reproduction as synonyms of replication.

Our search of the Scopus® database using these three terms confirmed empirically that "replication" is the most widespread term used in the sciences to refer to the repetition of an experiment.

Table 14
Search strings using the term repetition.

Search	Search String	Docs.
1	TITLE-ABS-KEY(repetition AND classification)	694
2	TITLE-ABS-KEY(repetition AND kind)	1.034
3.1	TITLE-ABS-KEY(repetition AND type)	5.902
3.2	TITLE-ABS-KEY(repetition W/2 type)	266
4	TITLE-ABS-KEY(repetition AND typology)	88
5	TITLE-ABS-KEY(repetition AND taxonomy)	70

References

[1] A.T. Adams, H. Ajrouch, J. Kristine, Henderson, H. Irene, Service-learning outcomes research: the role and scarcity of replication studies, J Appl Sociol 22 (2) (2005) 55–74.
[2] J.P.F. Almqvist, Replication of controlled experiments in empirical software engineering – a survey. Master's thesis, Department of Computer Science, Faculty of Science, Lund University, 2006. Supervisors: Amela Karahasanovic, Simula RL and Goran Fries, Lund University.
[3] H.M. Bahr, T. Caplow, B.A. Chadwick, Middletown iii: problems of replication, longitudinal measurement, and triangulation, Annu. Rev. Sociol. 9 (1) (1983) 243–264.
[4] H.R. Barker, E.B. Gurman, Replication versus tests of equivalence, Percept. Motor Skills 35 (1972) 807–814.
[5] V. Basili, F. Shull, F. Lanubile, Building knowledge through families of experiments, IEEE Trans. Softw. Eng. 25 (4) (1999) 456–473.
[6] C.T. Beck, Replication strategies for nursing research, J. Nurs. Scholarship 26 (3) (1994) 191–194.
[7] J. Beloff, Research strategies for dealing with unstable phenomena, in: B. Shapin, L. Coly (Eds.), The Repeatability Problem in Parapsychology, The Parapsychology Foundation, 1985.
[8] M. Bisell, The risks of the replication drive, Nature 503 (November) (2013) 333–334.
[9] K.B. Blomquist, Replication of research, Res. Nurs. Health 9 (3) (1986) 193–194.
[10] D. Brinberg, J.E. McGrath, Validity and the Research Process, Sage Publications, Inc., Newbury Park, Calif., 1985. June.
[11] A. Brooks, J. Daly, J. Miller, M. Roper, M. Wood, Replication of experimental results in software engineering, Technical Report ISERN-96-10, Univ. of Strathclyde, Scotland, 1996.
[12] S.W. Brown, K.A. Coney, Building a replication tradition in marketing, in: K. Bernhardt (Ed.), Marketing 1776–1976 and Beyond, American Marketing Association, Chicago, 1976.
[13] D.T. Campbell, J.C. Stanley, Experimental and Quasi-Experimental Designs for Research, Houghton Mifflin Company, 1963. June.
[14] J. Carver, N. Juristo, T. Baldassarre, S. Vegas, Replications of software engineering experiments. Guest's editor introduction to special issue on replication, Empirical Softw. Eng. 19 (2) (2014) 267–276.
[15] L.B. Christensen, Experimental Methodology, eighth ed., Allyn and Bacon, 2001.
[16] H.M. Collins, Changing Order: Replication and Induction in Scientific Practice, Sage Publications, 1985.
[17] T. Cook, D. Campbell, The Design and Conduct of Quasi-experiments and True Experiments in Field Settings, Rand McNally, Chicago, 1976.
[18] G. de Investigación en Ingeniería del Software Empírica (GrISE). Descriptions of the replications typologies, 2013.
[19] O. Dieste, A. Grimán, N. Juristo, Developing search strategies for detecting relevant experiments, Empirical Softw. Eng. 14 (October 2009) 513–539.
[20] B. Finifter, The generation of confidence: evaluating research findings by random subsample replication, Sociol. Methodol. 4 (1972) 112–175.
[21] B. Finifter, Replication and Extension of Social Research through Secondary Analysis, Soc. Sci. Inform. 14 (1975) 119–153.
[22] S.M. Fuess, On replications in business and economics research: the qjbe case, Quart. J. Business Econ. (1996). March.
[23] O.S. Gómez, N. Juristo, S. Vegas, Replication types in experimental disciplines, in: 4th International Symposium on Empirical Software Engineering and Measurement (ESEM'2010), Bolzan, Italy, 2010, pp. 71–75.
[24] E.S. Halas, R.L. James, L.A. Stone, Types of responses elicited in planaria by light, J. Comp. Physiol. Psychol. 54 (3) (1961) 302–305.
[25] J. Hannay, T. Dybå, E. Arisholm, D. Sjøberg, The effectiveness of pair programming: a meta-analysis, Inf. Softw. Technol. 51 (7) (2009) 1110–1122.
[26] W. Hayes, Research synthesis in software engineering: a case for meta-analysis, in: METRICS '99: Proceedings of the 6th International Symposium on Software Metrics, IEEE Computer Society, Washington, DC, USA, 1999, p. 143.
[27] C. Hendrick, Replications, Strict Replications, and Conceptual Replications: Are They Important?, Sage, Newbury Park, California, 1990
[28] K. Hunt, Do we really need more replications?, Psychol Rep. 36 (2) (1975) 587–593.
[29] J. Hunter, The desperate need for replications, J. Consumer Res. 28 (1) (2001) 149–158.
[30] H.H. Hyman, C.R. Wright, Evaluating social action programs, in: H.L. Lazarsfeld, H.L. Wilensky (Eds.), The Uses of Sociology, NY Basic, 1967, pp. 769–777.

[31] N. Juristo, Towards Understanding Replication of Software Engineering Experiments, Keynote, in: International Symposium on Empirical Software Engineering and Measurement (ESEM'13), Baltimore, USA, September 2013.
[32] N. Juristo, A. Moreno, S. Vegas, Reviewing 25 years of testing technique experiments, Empirical Softw. Eng. 9 (1–2) (2004) 7–44.
[33] N. Juristo, S. Vegas, The role of non-exact replications in software engineering experiments, Empirical Softw. Eng. 16 (3) (2011) 295–324.
[34] N. Juristo, S. Vegas, M. Solari, S. Abrahao, I. Ramos, A process for managing interaction between experimenters to get useful similar replications, Inf. Softw. Technol. 55 (2) (2013) 215–225.
[35] M. Jørgensen, A review of studies on expert estimation of software development effort, J. Syst. Softw. 70 (1–2) (2004) 37–60.
[36] C. Kelly, L. Chase, R. Tucker, Replication in experimental communication research: an analysis, Human Commun. Res. 5 (4) (1979) 338–342.
[37] G. Keppel, Design and Analysis. A Researcher's Handbook, Prentice Hall, Englewood Cliffs, NJ, 1982.
[38] B. Kitchenham, The role of replications in empirical software engineering – a word of warning, Empirical Softw. Eng. 13 (2) (2008) 219–221.
[39] J.L. Krein, C.D. Knutson, A case for replication: Synthesizing research methodologies in software engineering, in: 1st International Workshop on Replication in Empirical Software Engineering Research (RESER'2010), Cape Town, South, Africa, 2010.
[40] M.A. La Sorte, Replication as a verification technique in survey research: a paradigm, Sociol. Quart. 13 (2) (1972) 218–227.
[41] R.M. Lindsay, A.S.C. Ehrenberg, The design of replicated studies, Am. Stat. 47 (3) (1993) 217–228.
[42] J. Lung, J. Aranda, S. Easterbrook, G. Wilson, On the difficulty of replicating human subjects studies in software engineering, in: ICSE '08: Proceedings of the 30th International Conference on Software Engineering, ACM, New York, NY, USA, 2008, pp. 191–200.
[43] D.T. Lykken, Statistical significance in psychological research, Psychol. Bull. 70 (3) (1968) 151–159.
[44] R. Lüdtke, Do that to me one more time! – what kind of trial replications do we need?, Complementary Therapies Med 16 (4) (Aug. 2008) 181–182.
[45] V. Mandić, J. Markkula, M. Oivo. Towards multi-method research approach in empirical software engineering, vol. 32, 2009, pp. 96–110.
[46] J.V. McConnell, Specific factors influencing planarian behavior, in: W.C. Corning, S.C. Ratner (Eds.), Chemistry of Learning, Plenum, 1967, pp. 217–233.
[47] J.V. McConnell, R. Jacobson, B. Humphries, The effects of ingestion of conditioned planaria on the response level of native planaria: a pilot study, Worm Runner's Digest. 3 (1) (1961) 41–47.
[48] M. Mendonça, J. Maldonado, M. de Oliveira, J. Carver, S. Fabbri, F. Shull, G. Travassos, E. Höhn, V. Basili. A framework for software engineering experimental replications, in: ICECCS '08: Proceedings of the 13th IEEE International Conference on Engineering of Complex Computer Systems (iceccs 2008), IEEE Computer Society, Washington, DC, USA, 2008, pp. 203–212.
[49] J. Miller,. Can results from software engineering experiments be safely combined? in: METRICS '99: Proceedings of the 6th International Symposium on Software Metrics, IEEE Computer Society, Washington, DC, USA, 1999, p. 152.
[50] J. Miller, Applying meta-analytical procedures to software engineering experiments, J. Syst. Softw. 54 (1) (2000) 29–39.
[51] J. Miller, Replicating software engineering experiments: a poisoned chalice or the holy grail, Inf. Softw. Technol. 47 (4) (Mar. 2005) 233–244.
[52] R. Mittelstaedt, T. Zorn, Econometric replication: lessons from the experimental sciences, Quart. J. Business Econ. 23 (1) (1984).
[53] R. Moonesinghe, M.J. Khoury, A.C. Janssens, Most published research findings are false – but a little replication goes a long way, PLoS Med. 4 (2) (2007) 0218–0221.
[54] M.W. Moyer, Fish oil supplement research remains murky, Sci. Am. 24 (September) (2012).
[55] L. Pickard, B. Kitchenham, P. Jones, Combining empirical results in software engineering, Inf. Softw. Technol. 40 (14) (1998) 811–821.
[56] H. Radder. Experimental reproducibility and the experimenters' regress, in: PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol. 1, 1992, pp. 63–73.
[57] K. Rao, On the question of replication, J. Parapsychol. 45 (1981) 311–320.
[58] L.N. Reid, L.C. Soley, R.D. Wimmer, Replication in advertising research: 1977, 1978, 1979, J. Advert. 10 (1) (1981) 3–13.
[59] R. Rosenthal, Replication in behavioral research, J. Social Behav. Personality 4 (4) (1990) 1–30.
[60] C.L. Sargent, The repeatability of significance and the significance of repeatability, Eur. J. Parapsychol. 3 (1981) 423–433.
[61] S. Schmidt, Shall we really do it again? The powerful concept of replication is neglected in the social sciences, Rev. Gener. Psychol. 13 (2) (2009) 90–100.
[62] F. Shull, V. Basili, J. Carver, J. Maldonado, G. Travassos, M. Mendonça, S. Fabbri, Replicating software engineering experiments: Addressing the tacit knowledge problem, in: ISESE '02: Proceedings of the 2002 International Symposium on Empirical Software Engineering, IEEE Computer Society, Washington, DC, USA, 2002, p. 7.
[63] F. Shull, J. Carver, S. Vegas, N. Juristo, The role of replications in empirical software engineering, Empirical Softw. Eng. 13 (2) (2008) 211–218.
[64] M. Sidman, Tacticts of Scientific Research, NY Basic, 1960.
[65] F. Silva, M. Suassuna, A. França, A. Grubb, T. Gouveia, C. Monteiro, I. Santos, Replication of empirical studies in software engineering research: a systematic mapping study, Empirical Softw. Eng. (2012) 1–57.
[66] D. Sjøberg, J. Hannay, O. Hansen, V. Kampenes, A. Karahasanovic, N.-K. Liborg, A. Rekdal, A survey of controlled experiments in software engineering, IEEE Trans. Softw. Eng. 31 (9) (Sept. 2005) 733–753.
[67] R. Thompson, J. McConnell, Classical conditioning in the planarian, *Dugesia dorotocephala*, J. Comp. Physiol. Psychol. 48 (1) (1955) 65–68.
[68] W. Tichy, Should computer scientists experiment more?, Computer 31 (5) (1998) 32–40
[69] E. Tsang, K.-M. Kwan, Replication and theory development in organizational science: a critical realist perspective, Acad. Manage. Rev. 24 (4) (1999) 759–780.