Information and Software Technology 64 (2015) 76–101

Contents lists available at ScienceDirect

Information and Software Technology

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

Investigations about replication of empirical studies in software engineering: A systematic mapping study ☆

Cleyton V.C. de Magalhães *, Fabio Q.B. da Silva, Ronnie E.S. Santos, Marcos Suassuna

Centre for Informatics, Federal University of Pernambuco, Recife 50.740-560, Brazil

ARTICLE INFO

Article history:
Received 11 September 2014
Received in revised form 28 January 2015
Accepted 2 February 2015
Available online 14 February 2015

Keywords:
Replications
Experiments
Empirical studies
Mapping study
Systematic literature review
Software engineering

ABSTRACT

Context: Two recent mapping studies which were intended to verify the current state of replication of empirical studies in Software Engineering (SE) identified two sets of studies: empirical studies actually reporting replications (published between 1994 and 2012) and a second group of studies that are concerned with definitions, classifications, processes, guidelines, and other research topics or themes about replication work in empirical software engineering research (published between 1996 and 2012).
Objective: In this current article, our goal is to analyze and discuss the contents of the second set of studies about replications to increase our understanding of the current state of the work on replication in empirical software engineering research.
Method: We applied the systematic literature review method to build a systematic mapping study, in which the primary studies were collected by two previous mapping studies covering the period 1996–2012 complemented by manual and automatic search procedures that collected articles published in 2013.
Results: We analyzed 37 papers reporting studies about replication published in the last 17 years. These papers explore different topics related to concepts and classifications, presented guidelines, and discuss theoretical issues that are relevant for our understanding of replication in our field. We also investigated how these 37 papers have been cited in the 135 replication papers published between 1994 and 2012.
Conclusions: Replication in SE still lacks a set of standardized concepts and terminology, which has a negative impact on the replication work in our field. To improve this situation, it is important that the SE research community engage on an effort to create and evaluate taxonomy, frameworks, guidelines, and methodologies to fully support the development of replications.

1. Introduction

Replications of empirical studies play important roles in the construction of knowledge. According to Schmidt, a replication that demonstrates the same findings obtained by other experiment "... is the proof that the experiment reflects knowledge that can be separated from the specific circumstances (such as time, place, or persons) under which it was gained" [2]. Replications are also important to identify the range of conditions under which findings from one experiment hold and the possible exceptions [3].

Considering the importance of replications in the advance of science in general, Schmidt [2] expected that one would find a body of knowledge that provide clear and unambiguous definitions for central questions like 'what exactly is a replication experiment?', 'what exactly is a successful replication?', and 'what are all types of replication and their corresponding roles?'. Furthermore, one would expect to find empirically evaluated guidelines on how to perform and report replications complementing existing guidelines to perform experiments and other empirical studies.

However, Schmidt argues that this is not true for most of scientific disciplines [2]. The published replications and the theoretical works about replication research have not used clear-cut definitions of terms and concepts, and there is no generally accepted taxonomy to distinguish between types of replications and their roles in generating scientific knowledge. According to Schmidt, "the word replication is used as a collective term to describe various meanings in different contexts" [2]. Carver et al. [4] report that a similar situation is also found in empirical software engineering

☆ *Article Notes:* Preliminary and partial results of this study have been presented at the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE'2014) and published in the Conference Proceedings [1].

* Corresponding author.

http://dx.doi.org/10.1016/j.infsof.2015.02.001
0950-5849/

research. Our findings reinforce the need to address these issues in software engineering.

The goal of this article is to contribute to the advance of the replication work in empirical software engineering. We expect that the results presented in our study will stimulate and provide support for a debate in the scientific community to central questions related to replications. Although we do not expect to fully answer these questions in this article, we believe our work will contribute to some of the answers:

What should be considered a replication?
What should be considered a successful replication?
What are the types of replications and their functions?
How should replications be performed?
How should replications be reported?

In a recent mapping study, da Silva et al. [5] studied the current state of published replications of empirical studies in software engineering research. The mapping study selected and analyzed papers reporting replications of empirical studies published until 2010 and also found a second set of studies addressing several topics about replication work. The papers about replication were not further analyzed by da Silva et al. [5]. More recently, the same research group performed an update of the mapping study previously published, covering material published in 2011 and 2012 [6]. Also in this update, the same type of papers about replication were collected and saved for future analysis.

In this current article, we analyze and discuss the content of the papers about replications (hereafter referred to as ABO papers) published in the Software Engineering literature to increase our understanding about the current state of the work on replication in empirical software engineering research. We expect that this analysis will shed some light in the issues related to the five questions raised above.

Our goal is twofold. First, to classify the set of ABO studies in Software Engineering into categories related to the topics in which the articles focused on (recommendations, frameworks, guidelines, among others). Second, to analyze how the replications performed between 1994 and 2012 have cited and used the ABO studies, in order to verify the impact of these studies in recent replication work.

The set of papers analyzed in this article is composed of those selected by da Silva et al. [5], those found in the update of the mapping study [6], and papers found through a search process performed to cover work published in 2013. We systematically structured and analyzed data extracted from these articles to answer the following six research questions:

- *RQ1: What was the evolution in the number of ABO studies over the years?*
- *RQ2: Which individuals and organizations are most active in publishing ABO studies?*
- *RQ3: How the ABO studies define replication?*
- *RQ4: What topics or themes have been addressed by the ABO studies?*
- *RQ5: Which ABO studies are cited by the papers that reported replications?*
- *RQ6: How the results or propositions presented in the cited ABO studies have been used in papers that report replications?*

This article is organized as follows. In Section 2, we present a background with discussion on concepts and related works. In Section 3, we present the method used in this study. In Section 4, we present a comprehensive set of results of our review and in Section 5 we discuss these results. Finally, in Section 6, we present some conclusions and proposals for future works.

2. Background and related work

As briefly discussed in the Introduction, there is little agreement about nomenclature and definition of concepts about replication in many empirical sciences and also in empirical software engineering. In this article, we expect to shed some light on the debate about some theoretical and practical issues related to performing, classifying, and reporting replications in SE research. In this section, we start by providing some preliminary definitions, we then briefly describe the two mapping studies on replication that originated this current study and clarify some terminology issues. Finally, we show how this article improves the preliminary results published by Magalhães et al. [1].

2.1. Definition of replication

According to La Sorte, "replication refers to a conscious and systematic repeat of an original study" [7]. This definition implies that a replication must be explicitly related (conscious repetition) to a previous study. Similarly, A Dictionary of Social Sciences [8] defines replication as "a repetition of a research procedure to check the accuracy or truth of the findings reported". In fact, most definitions found in the scientific literature consider a replication to be a repetition of a research procedure already performed in another study, usually called the *original or the baseline study*.

This definition is a starting point in precisely characterizing what should be considered a replication. According to this characterization, empirical studies that address similar questions or hypothesis, but without explicit reference to a previous study that can be considered the original study, should not be considered replications. For this reason, da Silva et al. [5] do not consider as replications the studies that Krein and Knutson (2010) [ABO022] classify as independent replications. Similarly, we also do not consider replications the type of study that Baldassarre et al. [20] call conceptual replications. The reason in both cases is that the (very similar) definitions of independent and conceptual replication admit studies to be called replications without a reference (direct or indirect) to an original study.

However, because of the variations that may be intended or unintended introduced in the replication design, the definitions presented above are not precise enough to characterize unambiguously what should be considered a replication and what should be seen as an entirely different study. We expect that this article motivates the research community to engage on an effort in building standardized and consistent set of definitions and corresponding terminology related to replication work in SE research.

2.2. A brief summary of the mapping studies

The first article that explicitly reported a replication of an empirical software engineering study was published in 1994 [9]. The mapping study presented by da Silva et al. [5] analyzed 96 articles reporting 133 unique replications of 72 original studies published between 1994 and 2010. Bezerra and da Silva [6] updated da Silva's work and found 39 new articles, reporting 51 replications of 35 original studies, published in 2011 and 2012.

Using the definition of internal and external replication proposed by Brooks et al. [ABO036] to classify the replications, Fig. 1 shows the evolution of the number of replications found in the two mapping studies (da Silva et al. [5] is presented in blue and Bezerra and da Silva [6] in red).

da Silva et al. [5] and Bezerra and da Silva [6] raise several questions about the replication work in SE research. According to both studies, no clear cut definition of replication has been used in the studies, there is little standardization on how to report the

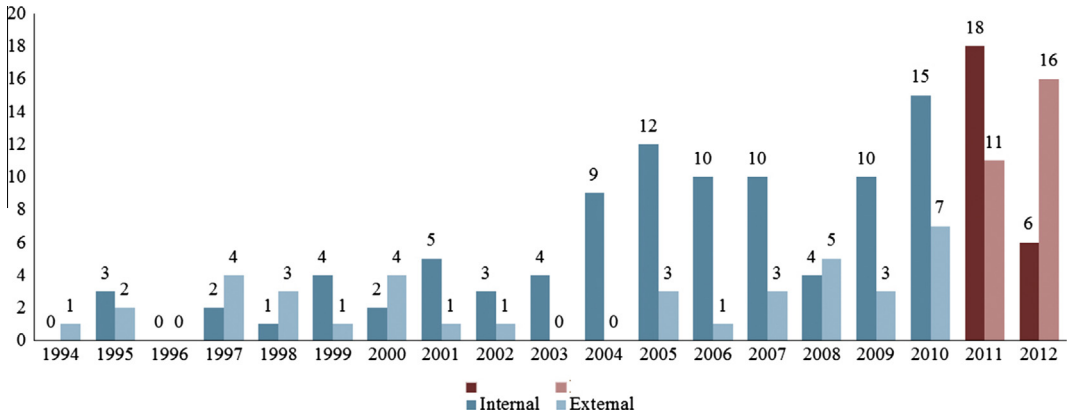

Fig. 1. Evolution of replications over the years.

replications, and there is no common characterization of replication types. Further, the vast majority of the papers reporting replications do not explicitly mention the specific motivation for performing a replication.

The two mapping studies also identified and selected several articles that report empirical and theoretical studies about issues related to replication. These articles do not present any replication, but describe conceptual frameworks, guidelines, processes, and lessons learned or recommendations about how to perform and report replications. Some of these articles also discuss different types of replications and their roles.

The mapping studies argued that there was a growing interest of the research community in performing and studying replications in the past decade. This interest resulted in an increase in the number of ABO studies published in the last 3 years. Another sign of this interest in the topic is the organization of three editions of the International Workshop on Replication in Empirical Software Engineering Research (RESER), in 2010 [10], 2011 [11], and 2013 [12]. This workshop is responsible for nearly a third (10/37) of the papers analyzed in this article and we provide a brief summary of its results in Section 5.3.

2.3. Naming conventions

In the rest of this article, we use the term *paper* to refer to the published work (article or other form of publication) analyzed in this review. We use *ABO study* to refer to the study that is reported in the paper. The references to the papers reporting ABO studies are numbered [ABOnnn], where *n* is a number between 0 and 9.

We use *replication* to refer to an experiment or other type of empirical study reported in the papers analyzed in the two mapping studies and use *replication paper* to refer to the article in which the replication is published, if the distinction is necessary. Replication papers are referenced consistently with the numbering systems used by da Silva et al. [5] and Bezerra and da Silva [6]. [REPnnn] is for replications from the former and [REPnnnFE] is used for the latter.

2.4. Improvements from the preliminary results of this mapping study

We presented some preliminary results of this mapping study at the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE2014) and the article was published in the conference proceedings [1]. In the conference article, we presented the results from the review of 36 papers and analyzed their use in a set of 46 replications papers published between 2010 and 2012. For this current article, we improve and extend those results in seven ways:

- We analyzed the use of the ABO studies over the entire set of 135 replication papers replications found in the two mapping studies between 1994 and 2012.
- We added three new research questions to guide important discussions about the ABO studies and their use in the replication papers.
- We performed the thematic analysis on the research topics addressed by the ABO studies to build a more refined categorization of the papers. As a result, we added three new categories to the five categories found by Magalhães et al. [1] and organized the set of papers accordingly.
- We largely extended and improved the presentation of the results and the discussion about the answers of each research question.
- We constructed new mappings to connect the topics addressed by the ABO studies and their use in the replication papers.
- We improved the discussions of our results by adding a new section on implications for research in which five relevant questions related to research about replication in empirical software engineering are discussed.
- We provided the complete list of references to the papers reporting the ABO studies (Appendix A) and the replication papers that cited at least one ABO study (Appendix B), and numbered the references to replication papers consistently with the two mapping studies.

Although the results presented in this article are based on essentially the same data set as Magalhães et al. [1], the new added information and the derived discussions represent significant improvements to the preliminary results.

3. Method

The scientific literature differentiates at least two types of systematic reviews: conventional systematic reviews and mapping studies [13]. The former aims to aggregate results about the effectiveness of a treatment, intervention, or technology, and therefore seeks answers to causal or relational research questions (e.g., *Is intervention I on population P more effective for obtaining outcome O in context C than comparison treatment C?*). The latter, aims to identify all research related to a specific topic and to answer broader and exploratory questions related to trends in research (e.g., *What do we know about topic T?*).

In this work, we studied the research work about replication research in software engineering. Our study analyzed and synthesized results from other published scientific articles. It is, therefore, classified as secondary study. The papers used in our study were selected from the two mapping studies discussed in Section 2.2,

covering the period of 1996–2012 and from a manual search procedure performed on journal and conference proceedings from 2013.

In this section, we describe the review method used by da Silva et al. [5] and Bezerra and da Silva [6], from which we selected the articles analyzed in our study. The work on conventional systematic literature reviews (SLR) [13] and the guidelines for performing SLR in software engineering presented by Kitchenham and Charters [14] were followed to plan and execute the mapping studies performed by da Silva et al. [5] and Bezerra and da Silva [6].

3.1. Inclusion and exclusion criteria

The protocol used to conduct the two mapping studies selected papers that met at least one of following two inclusion criteria:

(1) Papers reporting replications of empirical studies in Software Engineering.
(2) Papers reporting studies that were concerned with concepts, classifications, guidelines, and other themes about replication.

The protocol excluded papers that met at least one of these seven exclusion criteria:

(1) Written in any language but English.
(2) Not accessible on the Web.
(3) Invited papers, keynote speeches, workshop reports, theses, and dissertations.
(4) Incomplete documents, drafts, slides of presentations, and extended abstracts.
(5) Secondary and tertiary studies, and meta-analyses.
(6) Addressing areas of computer science that were clearly not Software Engineering (e.g., database systems, human–computer interaction, computer networks, etc.).
(7) Addressing replication only as part of future work.

We allowed one exception to the exclusion criteria for the work of Almqvist [ABO037]. This study is a Master Thesis and, therefore, should not be included according to exclusion criteria 3. However, this is the first work to review replications of experiments in SE, and has been the reference for various other studies selected in this review as well as in several replications analyzed in the two mapping studies. Further, Almqvist also proposes one of the few classification schemes for replications in SE research. Therefore, we decided to include this study in our review.

3.2. Data sources and search strategy

The search process combined automatic and manual search. Similar search processes were performed by the two mapping studies, covering papers published until 2012. The manual search was performed on the following relevant journals, conference proceedings, and on the list of primary studies analyzed in the three reports of related reviews:

- ACM Transactions on Software Engineering Methodologies.
- IEEE Transactions on Software Engineering.
- Empirical Software Engineering Journal.
- Information and Software Technology Journal.
- Int. Conference on Software Engineering.
- Int. Conference on Evaluation and Assessment of Software Engineering.
- Int. Symposium on Empirical Software Engineering and Measurement.
- Int. Ws. on Replication in Empirical Software Engineering Research.
- Related reviews (Almqvist, 2006; Carver, 2010; Sjøberg, 2005).

The researchers looked in the titles and abstracts of all papers in each source used in the manual search, using the same procedure applied to the list of papers returned in the automatic search. Therefore, both searches were compatible.

The automatic searches of the two mapping study were performed using the following five search engines and indexing systems:

- ACM Digital Library – http://portal.acm.org.
- IEEEXplore Digital Library – http://www.ieeexplore.ieee.org/Xplore.
- ScienceDirect – http://www.sciencedirect.com.
- Scopus – http://www.scopus.com.
- JSTOR – http://www.jstor.org.

Automatic searches were performed on the entire paper on all engines but Scopus, which did not perform full-text search at the time the search was conducted. For this engine, the search was performed on Title and Abstract.

da Silva et al. [5] performed the automatic search using a search string constructed based on three search terms and their synonyms: replication, empirical study, and software engineering (Fig. 2).

Bezerra and da Silva [6] searched for papers in almost the same engines, but using Springer instead of JSTOR, while performing the automatic search. The most significant change in this process between the two mapping studies was in the search string. da Silva et al. [5] had realized that the first string was retrieving a great number of undesirable papers and Bezerra and da Silva [6] (together with the researchers of the first mapping study) simplified the string. Fig. 3 presents the string used by Bezerra and da Silva [6].

The adequacy of the new string was tested by performing the automatic search on publications prior to 2011 to check if the papers found with the first string were also retrieved with the second string. The new string retrieved 100% of the articles selected by da Silva et al. [5].

In addition to the searches performed in the two mapping studies, for our review we conducted an automatic search in the same digital libraries mentioned above and manual search in the proceedings of the International Workshop on Replication in Empirical Software Engineering Research (RESER 2013) seeking for ABO studies published in 2013. We also searched the references of all replication papers selected in the two mapping studies [5,6] seeking for ABO papers that could have been missed in the previous searches.

3.3. Study selection

The study selection process executed by the two mapping studies identified 89 potentially relevant articles reporting ABO studies: 43 papers were identified by da Silva et al. [5] and 46 papers were identified by Bezerra and da Silva [6]. In both mapping studies, the researchers were searching for papers that have performed a complete replication. Articles that did not present a replication, but addressed replication issues somehow, were classified as potential ABO studies and no further analysis was performed on those papers. In particular, the inclusion/exclusion criteria presented in Section 3.1 were not applied on these potentially relevant articles. In our study, when we applied the inclusion and exclusion criteria (Section 3.1), we realized that several of these articles were not ABO studies because they did not addressed issues about

("replication" OR "replications" OR "reproduction" OR "reproductions" OR "reanalysis" OR "re-analysis" OR "empirical generalization" OR "generalization and extension" OR ("reproducibility" AND "experiment") OR "conceptual extension" OR "corroboration" OR "checking of analysis" OR "complete secondary analysis" OR "restricted secondary analysis" OR "pseudoreplication" OR "duplication" OR "conceptual extension" OR "empirical generalization" OR "model comparisons")
AND
("empirical" OR "experimental" OR "experiment" OR "experiments" OR "case study" OR "case studies" OR "ethnography" OR "ethnographic study" OR "ethnographic studies" OR "survey" OR "surveys" OR "action research" OR "quasi-experiment" OR "field research")
AND
("software engineering")

Fig. 2. Search string in first mapping study.

replication. Most of them just mention replications without actually presenting studies or findings about the issues addressed in this current study. Therefore, 62 articles from the set of 89 were excluded.

Therefore, the final set of selected ABO studies includes 37 articles: 28 from da Silva et al. [5], 3 from Bezerra and da Silva [6], and 6 from the automatic and manual search performed on papers published in 2013. Fig. 4 illustrates the process.

3.4. Data extraction

Data extraction was carried out guided by an extraction form implemented in MS Excel™. In this step, two researchers, working independently, analyzed each paper in order to answer the research questions previously defined. Conflicts of extraction of information were discussed and solved in consensus meetings, which involved at least three researchers. The results from data extraction were analyzed with support of MS Excel™, which was also used to generate graphics and tables. Table 1 shows the information extracted from the articles reporting ABO studies.

After extracting and partially analyzing the information about the ABO studies, we identified the articles reporting replication that cited ABO studies. The goal was to analyze the use of the ABO studies in the papers reporting replications. Table 2 shows the data extracted from articles reporting replications that cited the ABO studies.

3.5. Synthesis of results

The results from data extraction were integrated in spreadsheets, which were also used to generate graphs and tables. All descriptive information was calculated and organized using MS Excel™. The answers to RQ4 were constructed by analyzing the data extracted from each paper and synthesizing them using thematic analysis and qualitative coding techniques. Each article was read by two researchers and coded with respect to the topics or themes addressed. These processes looked at the research problem explicitly stated and the results reported by the articles. Categories that emerged from each paper were combined using constant comparison techniques.

("software engineering")
AND ("replication"
OR "replications"
OR "replicate"
OR "replicated"
OR "replicating")

Fig. 3. Search string in updated study.

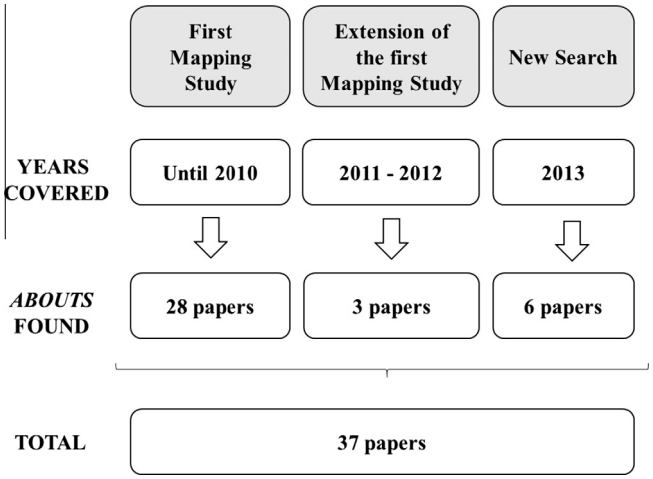

Fig. 4. Selection process.

4. Results

Our results naturally fall into two groups. The first group of research questions (RQ1–RQ4) deals with the descriptive nature of ABO studies and the second group of research questions (RQ5 and RQ6) describe how the papers reporting replications use of the results or propositions presented in the ABO studies.

4.1. Descriptive information about replications

In this section, we provide the answers to research questions RQ1–RQ4, summarizing the descriptive information about the ABO studies.

RQ1: What was the evolution in the number of ABO studies over the years?

We analyzed the evolution of publications of ABO studies over the years (Fig. 5). The first ABO study was published by Brooks et al. [ABO036] in 1996, the same group of researchers that performed the first published replication, in 1994, found by da Silva et al. [5]. This first ABO study proposes a framework for performing external replications and uses the first replication as an example of its application.

In our analysis, we identified 37 articles reporting studies about replication published in 17 years, between 1996 and 2013. In

Table 1
Data extracted from each ABO paper.

Data	Description
Title	Title of the paper
Year	Year of publication of the paper
Publisher	Journal or conference where the paper was published
Topic	Main topic or theme addressed by or discussed in the ABO paper
Replication definition	Definition of replication used and the reference from which it was extracted (when applicable)
Research problem	Research problem addressed by the paper
Proposal	Proposed solution to the problem addressed
Contribution	Main contribution of the paper

Table 2
Data extracted from replications that cite the ABO papers.

Data	Description
Title Rep	Title of the replication (paper)
Year Rep	Year of publication of the replication
Venue Rep	Journal or conference where the replication was published
ABO referenced	Papers in the ABO set cited in the replication paper
Level of use	Three levels of use (Level 0 (cited), Level 1 (used), and Level 2 (fully-used)) as describe in the answer to RQ6 in Section 4.2

Appendix A, we present the complete list of references of these articles.

We consider the total number of 37 articles on such an important topic to be very small. The average number of publications is only just above two per year. The growth in the publications after 2010 strongly influenced this average and coincided with the Workshop on Replication in Empirical Software Engineering Research (RESER), which started in 2010 and had three editions until 2013 [10–12].

Regarding the types of publication venue, over 67% (25/37) of the papers were published in conferences and workshops (19 full and 6 short papers). Just under 25% (9/37) were published in journal papers. RESER is the venue in which almost 27% (10/37) of the ABO studies were published (8 in RESER 2010 and 2 in RESER 2013), followed closely by International Symposium on Empirical Software Engineering and Metrics (ESEM) with 16% (6/37) of the papers. Considering journal papers, 16% (6/37) of the papers were published at the Empirical Software Engineering Journal. Together, these three venues are responsible for nearly 60% of the published ABO studies. The remaining papers were published in 13 distinct venues: 9 conferences, 2 journals (IST Journal and IEEE Transactions on Software Engineering), 2 books, and one Master Thesis.

RQ2: Which individuals and organizations are most active in publishing ABO studies?

We analyzed the ABO studies looking for the researchers most active in publishing studies about replication research in SE and found 76 distinct authors. In Table 3, we rank the most active researchers, their organizations, and the papers they co-authored.

In Table 3, we can identify the two most active groups working on research about replications: the group of the researchers collaborating with Natalia Juristo, and the researchers collaborating with Forrest Shull and Victor Basili. The former group produced almost 25% (9/37) of the ABO studies and the latter has been responsible for 17% (6/37) of the papers. Altogether, these two groups published nearly half of the ABO studies. Among the 16 most active researchers presented in Table 3, only five do not participate in the publications of one of the two groups and one researcher participate in one publication in both groups: Jeffrey Carver in [ABO003].

Table 4 presents the most active authors who have produced an ABO study and also published a replication paper analyzed in one of the two mapping studies. Only five of the most active authors of ABO studies did not participate in the publication of at least one replication paper.

Overall, just under 39% (30/76) of the authors published both ABO studies and replications papers. In the two mapping studies, 274 distinct authors were found and only 11% (30/274) of them also published an ABO study.

We can make three important observations from these data. First, few researchers and research groups are responsible for most of the effort related to studies about replication. Second, the majority of the researchers working on topics related to replications never participated in the publication of a replication paper. This may imply that these researchers have little practical experience in actually performing replications. Third, just above 10% of the researchers performing replications also published a paper about replication. Thus, there is only a small intersection between the researchers performing ABO studies and replications.

RQ3: How the ABO studies define replication?

We investigated the ABO studies looking for explicit definitions of what constitutes a replication and distinguished four categories of studies related to this issue. First, there are studies that use a definition from another published article, in SE or in another discipline. We found four ABO studies in this category and Table 5 shows the ABO study, the definition of replication used, and the reference from which it was extracted.

The second category includes 11 papers that developed their own definition and in Table 6 we present the ABO studies and the definition provided by each one. It is possible to notice that these definitions are not necessarily in contradiction with those presented in Table 5.

The third category includes papers that define or use classifications or taxonomies of replications, provide definitions for the individual replication types in the classification or taxonomy, but do not provide an explicit definition of replication in general. We found seven papers providing classifications and taxonomies (which are discussed further in Section 4.1, when we answer RQ4) and four papers that use a classification or taxonomy defined by another paper [ABO010] [ABO013] [ABO023] [ABO024]. In both cases, the studies may have implicitly used definitions of replication from other studies.

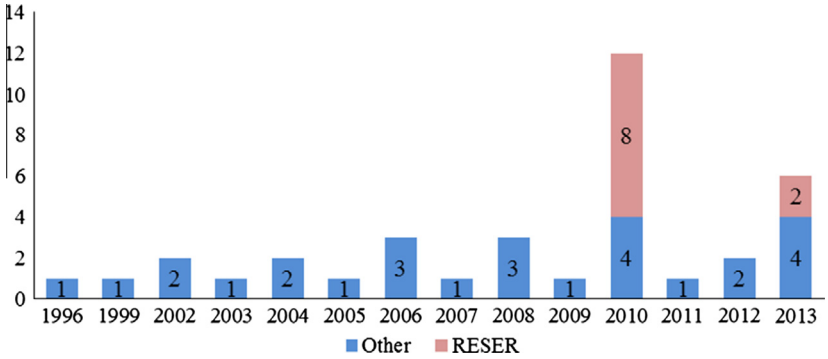

Fig. 5. Number of ABO published in each year.

Table 3
Researchers involved in ABO studies.

Author/organization	ABO published by the author	Total
Natalia Juristo *Universidad Politécnica de Madrid*	[ABO003], [ABO010], [ABO011], [ABO012], [ABO014], [ABO019], [ABO021], [ABO027], [ABO034].	9
Sira Vegas *Universidad Politécnica de Madrid*	[ABO003], [ABO010], [ABO011], [ABO014], [ABO021], [ABO034].	6
Jeffrey C. Carver *University of Alabama*	[ABO001], [ABO007], [ABO012], [ABO015], [ABO016], [ABO018].	6
Forrest Shull *Fraunhofer Centre for Experimental Software Engineering*	[ABO001], [ABO004], [ABO007], [ABO012], [ABO015], [ABO016]	6
Victor R. Basili *University of Maryland*	[ABO001], [ABO004], [ABO007], [ABO015], [ABO016]	5
José C. Maldonado *University of São Paulo at São Carlos*	[ABO001], [ABO007], [ABO015], [ABO016]	4
Guilherme Horta Travassos *Federal University of Rio de Janeiro*	[ABO001], [ABO007], [ABO016]	3
Omar S. Gómez *Universidad Politécnica de Madrid*	[ABO010], [ABO021], [ABO027]	3
Sandra Fabbri *Federal University of São Carlos*	[ABO001], [ABO007]	2
Per Runeson *Lund University*	[ABO006], [ABO035]	2
James Miller *University of Alberta*	[ABO017], [ABO036]	2
Dag I.K. Sjøberg *University of Oslo*	[ABO024], [ABO026]	2
Maria C.F. de Oliveira *University of São Paulo at São Carlos*	[ABO001], [ABO007]	2
Martín Solari *Universidad ORT Uruguay*	[ABO003], [ABO034]	2
Gregorio Robles *Universidad Rey Juan Carlos*	[ABO008], [ABO025]	2
Barbara Kitchenham *Keele University*	[ABO005], [ABO013]	2

Finally, we found 11 papers that do not present or use any definition of replication: [ABO002], [ABO005], [ABO008], [ABO018], [ABO019], [AB0025], [ABO026], [ABO028], [ABO031], [ABO032], and [ABO033].

RQ4: What topics or themes have been addressed by the ABO studies?

As mentioned in Section 3.5, the information used to answer this question was extracted from the ABO papers using thematic analysis and coding techniques. Eight categories emerged from this process: Recommendations, Replication Types, Process, Framework, Tools, Guidelines, Result Combinations and Miscellaneous. These categories were given a description, as follows:

Recommendations – papers in this category discuss experiences and lessons learned from performing replications. They also present challenges and solutions proposed to solve frequent problems. These recommendations are given at different levels of formality or systematization in each study.
Replication Types – papers in this category discuss the problem of how to distinguish different types of replications and their roles in the generation of scientific knowledge.
Process – papers in this category discuss general aspects of the process to conduct replication or aspects related with specific steps of this process, e.g., communication between experimenters, gain and transmission of knowledge, and experimental design issues specific to replications.
Framework – papers in this category propose conceptual frameworks, or abstract representations, to help understanding and improving the development of replications. Proposed frameworks are not theories, but offer an organization of concepts to assist the research in collecting and analyzing data in an experimental replication.
Tools – papers in this category propose tools to support some aspect of designing and performing replications.
Guidelines – these papers propose guidelines to support development or to report replications of experiments in Software Engineering. In these papers, the level of formality of the

Table 4
Researchers involved in ABO studies and replication papers.

Author/organization	Number of ABO	Number of REP
Natalia Juristo *Universidad Politécnica de Madrid*	9	2
Sira Vegas *Universidad Politécnica de Madrid*	6	1
Jeffrey C. Carver *University of Alabama*	6	4
Forrest Shull *Fraunhofer Centre for Experimental Software Engineering*	6	4
Victor R. Basili *University of Maryland*	5	4
José C. Maldonado *University of São Paulo at São Carlos*	4	1
Guilherme Horta Travassos *Federal University of Rio de Janeiro*	3	1
Sandra Fabbri *Federal University of São Carlos*	2	1
Per Runeson *Lund University*	2	7
James Miller *University of Alberta*	2	5
Dag I.K. Sjøberg *University of Oslo*	2	4

guidelines and the practical evaluation of their adequacy or applicability also vary significantly among the papers.

Result Combination – papers in this category address two issues. First, how to combine the results of several replications to increase statistical power of the results. Second, how variations between the original study and its replications make combination of results more complex.

Miscellaneous – we grouped together studies that do not fall in previous categories and, therefore, would form a new category with only one single paper.

A given ABO study may present results or propositions related to more than one category. For instance, Basili et al. [ABO004] and Brooks et al. [ABO036] discuss the characterization of replication types and also propose a conceptual framework for experimental replications. In such cases, we classified the studies according to its central goal or focus, and discussed its results in all categories it was related to. Thus, [ABO036] is classified in the Framework category and it is also discussed in the Replication types category.

Table 7 presents the distribution of papers in each category.

In the remaining of this section, we present a summary of the papers of each category. We tried to synthesize and integrate the results whenever possible. However, papers in each category almost never address the same research problem, making it very difficult to integrate their results. Nevertheless, we built summaries at the end of most topics, trying to provide synthesis of the studies that could help the reader to make sense of the central or more important issues addressed by the papers.

4.1.1. Recommendations

The eight papers in this category discuss challenges found and lessons learned during the development of replications and provide recommendations based on them. ABO studies providing recommendations were grouped according to the type of research method primarily addressed by the studies. Four groups were found, reporting recommendations for replicating: experiments or quasi-experiments (2); case studies (3), survey research (1); and replications of mining software repository studies (2).

4.1.1.1. Replicating experiments or quasi-experiments. Mende et al. [ABO009] studied replications in the domain of defect prediction. Their key recommendation is related to the availability of information about the original study: "The replication of studies is only possible when all details of the original study are known (or at least easy to guess)" [ABO009].

Shull et al. [ABO016] describe the development of a family of experiments about software reading techniques. Among their recommendation we highlight: (a) performing pilot studies by the replication researchers for understanding about the original study; (b) having a local expert who understand the technology being studied; and (c) the terminology used must be clearly defined and explained to subjects.

4.1.1.2. Replicating case studies. Ohlsson and Runeson [ABO006] studied replications of case studies to validate the definition of different levels of replication and to characterize the similarities and differences between an original study and a replication. Their most significant recommendation is related to the level of detail and consistence of the information provided in the original studies to facilitate replication: "original studies should be reported with much more detail and openness, ultimately including publication of raw data" [ABO006];

Ferrari et al. [ABO020] describe their experience in transitioning from the laboratory study to the large-scale case study performed in industry. They highlight five key lessons learned, and corresponding recommendations, for this type of replication: (a) integrate domain knowledge acquisition as part of the industrial case study; (b) the researcher performing the replication must adapt all investigative procedures of the original lab study design to align with the large-case study context; (c) due to the large difference between the two types of studies, it may be possible to extend the original lab study with new research questions by considering the case study environment; (d) data collection in the large-scale case study should not put excessive burden on participants, but should instead focus on non-obtrusive and flexible schemes; and (e) new threats to validity may occur in the large-scale study that have not been anticipated by analyzing the laboratory study threats.

Mockus et al. [ABO024] studied four replications of case studies on reproducibility of software development. The three key lessons learned from their studied are summarized as follows: (a) the main challenge in replicating industrial case studies is the opportunity for replication; (b) motivating other researchers to perform replications of case studies is difficult; and (c) measuring the same dependent and independent variable in different contexts may be a challenge, as "each case may lack some of the data sources or the data may be less reliable".

4.1.1.3. Replicating survey research. Cater-Steel et al. [ABO002] studied the replication of surveys in software engineering research. The authors summarize their experience with some recommendations mainly related to the survey instruments and other aspects of the research:

- Conduct research to verify if the survey instruments are up-to-date with current practice and, if needed, add questions to bring the instruments up-to-date.
- The addition of more questions requires care in reporting comparisons between original and replicated surveys.
- Consistently with general advice on reporting empirical studies, researchers should provide full detail of their surveys to facilitate replications.
- Communication with the researcher that performed de original survey is encouraged to provide a clearer understanding of the motivation, context and limitations of the original study.

4.1.1.4. Replicating studies in mining software repository. We found two articles that investigated replications in the domain of mining software repositories (MSR), which are based on the same study [ABO008] [ABO025]. Based on a literature review, the authors conclude that MSR articles published at the Mining Software Repositories Workshop/Working Conference do not satisfy certain requirements for easy replication. They present some recommendations to improve on this problem, all related to clearer and consistent definitions in the design and execution of the original study, including: (a) to present detailed description of the data being studied (including the exact time span or the versions of the software); (b) to indicate the specific version of the research tool used; and (c) to reach an agreement on a standardized way to refer to a location where additional data can be obtained.

4.1.1.5. Summarizing recommendations. We synthesized the main recommendations from all papers in this topic using the following technique based on the translation process of meta-ethnography [19]:

1. We built a translation table (Appendix C, Table 16) in which the first column groups the ABO studies according to the research methods as described above and the second column shows the study identifier.

Table 5
Definitions of replication from other sources.

ABO study	Definition of replication	Reference
[ABO001]	Partial replication as defined in [ABO017]: "to repeat the study, often while changing some of the parameters, to see if the original result is stable with regard to repetition and alteration of some of its components"	[ABO017]
[ABO020]	Literal replication as defined by Yin: "where the same constructs and basic measurements are used but different cases are selected that, based on the similarity of the cases used, the results are directly comparable to the original study"	Yin [15]
[ABO029]	"Replication means that other researchers try to reproduce the original experiment as closely as possible in other contexts and using different samples"	Judd and Kidder [16]
[ABO035]	"Replication of an experiment refers to repeating an experiment very closely following the method used in the baseline experiment"	Cartwright [17]

Table 6
Papers that provided their own definition.

ABO study	Definition of replication
[ABO003]	"Replication is usually construed as other researchers attempting to reproduce the research in other settings with different samples"
[ABO006]	"The replication involves application of a study under conditions as similar as possible to the original ones, but in another context or population"
[ABO007]	"In this work we will consider a replication to be a study that is run, based on the design and results of a previous study, whose goal is to either verify or broaden the applicability of the results of the initial study"
[ABO015]	
[ABO016]	
[ABO009]	"... the repetition of an experiment without any changes"
[ABO011]	"Intuitively, replication means the repetition of an experiment to double-check [verify] its results"
[ABO014]	
[ABO034]	
[ABO017]	"... to repeat the study, often while changing some of the parameters, to see if the original result is stable with regard to repetition and alteration of some of its components"
[ABO037]	"Replication, in the context of this thesis, is the repetition of an experiment, either as closely following the original experiment as possible, or with a deliberate change to one or several of the original experiment's parameters"

Table 7
Topics addressed by the ABO studies.

Topic	Amount	Paper ID
Recommendations	8	[ABO002] [ABO006] [ABO009] [ABO008] [ABO016] [ABO020] [ABO024] [ABO025]
Replication Types	7	[ABO010] [ABO012] [ABO013] [ABO021] [ABO027] [ABO030] [ABO037]
Processes	6	[ABO003] [ABO007] [ABO011] [ABO014] [ABO015] [ABO034]
Frameworks	6	[ABO001] [ABO004] [ABO017] [ABO022] [ABO031] [ABO036]
Tools	3	[ABO028] [ABO029] [ABO033]
Guidelines	2	[ABO005] [ABO018]
Result Combination	2	[ABO019] [ABO035]
Miscellaneous	3	[ABO023] [ABO026] [ABO032]

2. We, then, listed all central recommendations from each paper in the third column (Study recommendations).
3. We looked for recommendations that could be useful for any type of research method and created the fourth column with only the generic recommendations (First level generalization).
4. We grouped the generic recommendations in more abstract concepts in the fifth column (Second level generalization).
5. Finally, we integrated the recommendations by synthesizing them in three categories that summarizes the key points from the eight studies (Synthesis).

These three key categories are discussed below.

Understanding the original study is a central issue in performing a replication. The recommendations regarding this understanding fall into three groups: communications with the researchers that performed the original studies; the development of pilot studies by the researchers performing the replication; and the availability of precise and detailed information about the original studies in publications or research packages.

Variations between original study and replication are almost inevitable due to intended or unintended reasons. These variations must be fully assessed and their possible effects on the results of the replication and the comparison of these results and those of the original study must be evaluated. Intended variations are related to the need or desire of changing certain aspects of the original study, such as updating or extending research instruments or procedures, or even extending the research goals and questions. As these variations are either under direct control of the researcher performing the replications or at least known to her, they effect may be easier to assess and control. Unintended variations are related to the new context in which the replication will occur and pose more challenges to researchers because they are more difficult to be found and controlled.

Precise and unambiguous design and execution of the studies (original and replication) is a major issue in the successful development of any empirical study. Empirical studies must all be carefully designed and executed to increase the validity and, therefore, the usefulness of their results. In the case of replications, this becomes even more critical because imprecisions of design or execution of original or replication can lead to unintended (and often not explicitly addressed) variations among studies that could ultimately make the comparison of results between studies very difficult or even impossible.

4.1.2. Replication Types

We found eight studies addressing the issue of types of replication: seven of them are primarily concerned with this topic and one addresses this issue as a secondary theme. The studies in this topic fall into three groups: papers presenting similar results from the same literature review, that investigated replication types in other scientific disciplines (2); papers presenting replication types specific for SE replications (5); one paper describing a research proposal for the construction of a taxonomy of replications (1).

4.1.2.1. Replication types in other scientific disciplines. Gómez, Juristo and Vegas [ABO010] and Juristo and Gómez [ABO027] describe the results of a literature review about replication types in other scientific disciplines and also describe the replication types that have been used in SE literature by [ABO012], [ABO036], and [ABO037] (discussed below).

This literature review found eighteen different classification schemas and the results demonstrate a lack of uniformity or standardization in the classification of replication types at both inter and intra-disciplinary levels. The papers that define replication types in SE present the same lack of standardization, as will be discussed below.

4.1.2.2. Replication Types in SE. We found seven papers that primarily address issues related to replication types ([ABO010], [ABO012], [ABO013], [ABO021], [ABO027], [ABO030], and [ABO037]) and one paper that propose types of replication as a secondary goal [ABO004]. Recently, Baldassarre et al. [20] studied[1] the research about types and classifications of replications in SE research and identified that this research broadly describes types of replication based two dimensions: "(1) procedure, i.e., the steps followed in the study and (2) people, i.e., the experimenters conducting the replication". Following this line of reasoning, the ABO studies in this topic fall into three groups: 1 – procedures; 2 – people; 3 – procedures and people.

1. Types and classifications related to procedures

Basili et al. [ABO004] present a framework for organizing related studies in families of experiments. In this framework, replications are classified according to the following types:

(a) Strict replications (as close as possible to the original experiment).
(b) Replications that vary variables intrinsic to the object of study.
(c) Replications that vary variables intrinsic to the focus of the evaluation.
(d) Replications that vary context variables in the environment in which the solution is evaluated.
(e) Replications that vary the manner in which the experiment is run.
(f) Replications that extend the theory.

Shull et al. [ABO012] propose the classification of replications in two types: exact replications, in which the procedures of an experiment are followed as closely as possible; and conceptual replications, in which the same research question is evaluated by using a different experimental procedure by a different group of researchers. The category of exact replications is further refined into two sub-categories: "those replications in which researchers attempt to keep all the conditions of the experiment the same or very similar" (exact-dependent) and "those replications in which researchers deliberately vary one or more major aspects of the conditions of the experiment to address a specific research question" (exact-dissimilar). Shull et al. [ABO012] also discuss the role of these different types of replication and the use of lab packages as a way of communication between the teams performing the replication and the researchers that performed the original study. Kitchenham [ABO013] criticizes the role of lab packages and also the arguments in favor of exact replications. In the latter case, Kitchenham views exact replication adding too little to the understanding of software engineering phenomena and points out that frameworks or theories are essential in building this understanding.

Gómez, Juristo and Vegas [ABO021] propose to use the terms replication, reproduction, and re-analysis as replication types. This classification presents a series of conceptual problems and has not been used further. In particular, re-analysis of the same data set of an experiment should not be considered a replication because no new experiment is actually performed.

Krein and Knutson [ABO022] propose a framework for organizing research methods in SE, including experimental replications. Their framework adopts the following types and definitions:

(a) *Strict replication:* which is meant to replicate a prior study as precisely as possible.
(b) *Differentiated replication:* which intentionally alters aspects of the prior study in order to test the limits of that study's conclusions.
(c) *Dependent replication:* which is a study that is specifically designed with reference to one or more previous studies, and is, therefore, intended to be a replication study.
(d) *Independent replication:* which addresses the same questions and/or hypotheses of a previous study, but is conducted without knowledge of, or reference to, that prior study either because the researchers are unaware of the prior work, or because they want to avoid bias.

In their mapping study, da Silva et al. [5] criticize the concept of Independent Replications, as defined by Krein and Knutson [ABO022]. In their criticism, the authors of the mapping study observe that definitions of replication found in scientific literature make explicit the reference to an original or baseline study for the new experiment to be considered a replication.

Baldassarre et al. [20] propose a classification of replications according to its similarity to original experiment: Close (as similar as possible), Differentiate (some changes intentionally made), and Conceptual (only research question or hypothesis are the same). Similarly to Krein and Knutson's notion of Independent Replication, Conceptual Replications do not necessarily refer to an original experiment. Baldassarre et al. [20] conclude from the results of their study that "In spite of the agreement on the meaning, the definition of conceptual was only reported on 65% of the forms. Several comments questioned on the need for this type of replication and whether it applies to software engineering". This reinforces the criticism made by da Silva et al. [5] with respect to Independent Replications.

2. Types and classification related to people

Brooks et al. [ABO026] classify replications with respect to the involvement of the researchers that performed the original study in the development of the replication. Therefore, a replication is considered to be internal if the same researchers, or a sub-group of them, performed the original study and the replication. A replication is considered external if the researchers performing it are different from those that have performed the original study.

[1] This article is not included as an ABO study because it was published in 2014 and our review only covered articles published until 2013.

Regarding the people dimension, Baldassarre et al. [20] also classify replications as Internal and External, following a similar definition of Brooks et al. [ABO036].

da Silva [5] and Bezerra and da Silva [6] use the classification of internal and external replications in the two mapping studies. In both cases, the authors use an operational definition based on authorship of the papers reporting original study and replication. Therefore, a replication is classified as external if there are no common authors between the replication and its original study. Conversely, an internal replication has one or more common authors.

This definition is clear-cut and makes the classification precise and non-ambiguous. However, there is some controversy as to whether this definition is too strict or syntactical. Baldassarre et al. [20] uses a different definition in which a replication is considered external if "different experimenters (or most of them are different from the original group of experimenters) carry out the replication". Goméz et al. [24] does not use the internal/external terminology. Instead, they consider the experimenters (researchers performing several roles in the experiment) as one dimension that may vary between original study and its replication. However, they are not precise as to how many different experimenters are sufficient for a replication to be considered as Changed-experimenters.

3. Types and classifications related to procedures and people

Almqvist [ABO037] combined variations in the procedures with the classification of Internal and External replications proposed by Brooks et al. [ABO026] to define a classification that mix procedures and people. His classification defines the following types of replication:

(a) *Similar-external replications:* same experiment developed by other researchers.
(b) *Improved-internal replications:* variations in the experiment performed by the same researchers.
(c) *Similar-internal replications:* same experiment developed by the same researchers.
(d) *Differentiated-external replications:* variations in the experiment performed by other researchers.

4.1.2.3. Research proposal. As can be observed above, there is no agreement among the studies about replication types. Although the work of Baldassarre et al. [20] and Goméz et al. [24] provide good steps toward a taxonomy of replications, both have not been evaluated in practice. In this context, Magalhães et al. [ABO030] present a research proposal aimed to collect, analyze, and synthesize data toward the construction of a taxonomy of replications in empirical Software Engineering research. Based on the work of Hendrick [21] and Schmidt [2], the goal of this research proposal is to construct a taxonomy in which replication types are tied to variations between the original experiment and the replication, and also to the function or purpose of the replication. The propositions of Baldassarre et al. [20] and Goméz et al. [24] are consistent with this goal and could be used as a starting point.

4.1.2.4. Summarizing replication types. Any attempt to synthesize the classifications or types provided in the literature in general and in SE specifically must be performed with caution. As pointed out by Juristo et al. [ABO010], "Some authors use the same replication name, although they each define the replication differently. We also came across the opposite case, where authors used different replication names to refer to equivalent replications". This is the case both in other disciplines and also in SE research.

In Table 8, we present a summary of the five ABO studies that propose classifications and types of replications related to procedures. Definitions in the same row are similar, although not necessarily equivalent. For instance, as discussed above, Independent replications are explicitly not related to a prior study whereas the other types in the same row either explicitly mention an original or baseline study (Differentiate-external and Reproduction) or are not clear about this (Vary the manner in which the experiment is run and Conceptual).

Independent and Differentiate-external can only be external replications. All other types of replication can be performed by the same researchers or by different researchers, and therefore can be internal or external, according to the classification of Brooks et al. [ABO026].

4.1.3. Processes

The papers in this category are divided in two groups. Two papers, reporting versions of the same study, propose a full process to support the development of non-exact replications. Four papers, referring to two distinct studies, discuss the process of communication between the team that developed the previous (original) experiment and the team performing the replication.

4.1.3.1. Full process for non-exact replications. The study presented by Juristo and Vegas [ABO011], and then extended by the same researchers [ABO014], proposes a process whereby experimenters can run non-exact (dissimilar) replications and generate relevant knowledge in doing so. Both papers start by discussing the role of non-exact replication in creating knowledge in empirical studies in software engineering. According to the authors, "Researchers enacting this process will be able to identify new variables that are possibly having an effect on experiment results" and use the variations between the experiments to gain scientific knowledge. The proposed process consists of four phases, as explained by the authors: replication definition and planning, replication operation and analysis, replication interpretation, and analysis of the replication's contribution. The researchers evaluated the process in a set of case studies, "revealing the variables learned from two different replications of an experiment".

4.1.3.2. Communication between researchers. Communication between the researchers who ran the original or baseline experiment (called here the earlier team) and those that will run the replication (the replication team) has been recognized as a key factor to make the replication feasible. Four ABO studies ([ABO003], [ABO034], [ABO007], and [ABO015]) address the issues related to how much and what type of communication could or should happen. They also discuss the role of replication packages or lab packages in transferring knowledge between the teams. In this set of studies, we found two proposals that originated from two research groups. The studies in Vegas et al. [ABO003] and Juristo et al. [ABO034] address the types of meetings that could or should happen between the research teams, whereas the work of Shull et al. [ABO007] and Shull et al. [ABO015] proposes a knowledge-sharing model to enhance communication.

Vegas et al. [ABO003] discuss how different communication mechanisms (CM) affect the results of a replication, by analyzing three replications in which different CM were used. Their contention is that a minimum level of communication is necessary, in particular before the beginning of the replication, to prevent unwanted changes in the design that could make the aggregation of results difficult or even impossible. They concluded that:

> *"... we have found that there should be a minimum amount of communication among experimenters for replications to be successful. Nevertheless, this communication does not necessarily have*

Table 8
Summary of types and classifications of replications.

Basili et al. [ABO004]	Shull et al. [ABO012]	Krein and Knutson [ABO022]	Almqvist [ABO037]	Juristo et al. [ABO021]
Strict [replications that duplicate as accurately as possible the original experiment]	**Exact-dependent** [researchers attempt to keep all the conditions of the experiment the same or very similar]	**Dependent-strict** [to replicate a prior study as precisely as possible]	**Similar-internal** [replications that are performed by the same experimenters than the original experiment and that can be classified as close replications*] **Similar-external** [replications that are performed by other experimenters than the original experiment and that can be classified as close replications*]	
Vary variables intrinsic to the object of study [vary independent variables] Vary variables intrinsic to the focus of the evaluation [vary dependent variables] **Vary context variables** [identify potentially important environmental factors that affect the results of the process under investigation]	**Exact-dissimilar** [researchers deliberately vary one or more major aspects of the conditions of the experiment]	**Dependent-differentiated** [intentionally alters aspects of the prior study in order to test the limits of that study's conclusions]	**Improved-internal** [replications by the same experimenters under different conditions, in different settings or with modified tasks.]	**Replication** [verifies that the observed findings are stable enough to be discovered more than once. Replication uses the same method as in the baseline experiment]
Vary the manner in which the experiment is run [testing the same hypotheses as previous experiments have done, but altering the details of the experiment]	**Conceptual** [the same research question is evaluated by using a different experimental procedure]	**Independent** [addresses the same questions and/or hypotheses of a previous study, but is conducted without knowledge of, or deference to, that prior study]	**Differentiated-external** [involves deliberate, or at least known, variations in fairly major aspects of the conditions of the study and conducted by other researchers than the original]	**Reproduction** [in reproduction a new experiment is run (using different experimental methods) to test the same hypotheses as the baseline experiment]
Replications that extend the theory [making large changes to the process, product, and/or context models to see if basic principles still hold]				
				Re-analysis [the data of a previously run experiment are used to verify the results rather than re-running the experiment, using the same or different analysis techniques]

to be prolonged over years, but can be based on a couple of meetings held in a short period of time just before and just after the experiment."

Juristo et al. [ABO034], using the results Vegas et al. [ABO003], propose a communication process consisting of two meetings between the earlier team and the replication team: an adaptation meeting and a combination meeting. The former has the objective of helping the replication team to "tailor the experiment to the new setting", while during the latter, the two teams would discuss and combine their results. The process was evaluated in three replications and the authors concluded that the process is effective in the support of similar replications in software engineering.

Shull et al. [ABO015] address the issue of knowledge transfer between research teams. They argue that experimentation know-how and tacit knowledge are difficult to communicate and that replication packages are not enough to ensure effective knowledge sharing for a replication to be successful. Shull et al. [ABO007], based on the conclusions of [ABO015], propose a more concrete communication and knowledge sharing mechanism for replications. Toward that goal, they propose an adaptation of Nonaka-Takeuchi knowledge sharing model in the context of empirical software engineering, which is called Experimentation Knowledge-Sharing Model (EKSM) in their paper. They adapted the four phases Nonaka-Takeuchi's model into phases related to experimentation: "sharing tacit knowledge through socialization among replicators, making tacit knowledge explicit through lab packages, improving explicit knowledge through lab package evolution, and internalizing experimental knowledge by using the lab packages".

As seen above, the discussion on communication is often complemented by discussions on the availability, use, and contents of replication packaged or lab packages, because the two issues are inevitably intertwined. The ABO studies that address both issues seem to acknowledge that the amount of communication needed is somehow related to the amount of information (or lack thereof) in the replication packages. They also agree that stand-alone replication packages are not enough to assure successful replications, in particular because experimentation know-how and other forms of tacit knowledge are very difficult or even impossible to code in such instruments.

4.1.3.3. Summarizing processes. The process of communication and interaction between researchers is addressed by most of the ABO studies in this topic. This communication is essential for the replication team to understand the original study in enough detail. This communication issue has also been addressed by some ABO studies that we analyzed in the Recommendation topic, thus demonstrating that this process plays an important role in replication.

This is especially true in software engineering, in which experiments commonly involve the participation of human subjects. As demonstrated by Lung et al. [18], studies with human subjects are prone to have more unwanted variations between original study and the replication, which can potentially affect the comparison of the results of the two experiments. Thus, the process of

replication can be improved when the experimenters conducting the replication are able to use different types and levels of communication, as well as different tools and mechanisms to communicate and interact with the team of the original experiment, in order to understand issues of the original study, such as procedures, design, operational definitions of variables, threats to validity, data collection, and analysis.

However, communication between teams of researchers can propagate flaws or biases from the earlier team to the replication team, invalidating the verification purposes of an external replication. Therefore, the central question is how much information should be exchanged and how to prevent the propagation of flaws or biases between the research teams.

4.1.4. Frameworks

Papers within this theme comment on the difficulties of conducting replications and propose conceptual frameworks as a better alternative for understanding, organizing, and performing replications of SE experiments. Each paper in this category proposes a distinct framework, briefly described below.

4.1.4.1. FIRE (Framework for Improving the Replication of Experiments). Mendonça et al. [ABO001] propose FIRE as a framework to address knowledge sharing issues both at the intra-group (internal replications) and inter-group (external replications) levels. FIRE encourages coordination of replications in order to facilitate knowledge transfer for lower cost, higher quality replications and more generalizable results.

As can be viewed in Fig. 6, FIRE contains two cycles. An internal cycle (experiment execution and intra-group learning), in which experimenters focus on independently and successfully planning, executing, learning and packaging the experiment within their own context. There is also an external (inter-group learning) cycle (EC), in which experimenters are concerned with collaborative package standardization, experimental knowledge evolution, and knowledge sharing.

4.1.4.2. Family of experiments. A framework for organizing sets of related studies is introduced by Basili et al. [ABO004]. This study proposes to apply the goal-question-metric template (GQM) to organize sets of similar experiments. The researchers demonstrate the use of their framework on a set of experiments in Software Reading Techniques. The study argues that the use of this framework facilitates incremental knowledge building and is important to "(a) investigate the effects of alternative values for important attributes of the experimental models; (b) vary the strategy with which detailed hypotheses are investigated; and (c) make up for certain threats to validity that often arise in realistically designed experiments" [ABO004].

4.1.4.3. Miller's replication framework. Miller [ABO017] proposes a replication framework that identifies "four key dimensions of replication effectiveness" that, according to the author, should be balanced in a replication, or series of replications: (a) existential realism, which is related to the gaps between the experimental situation and the real world (subject, tasks, artefacts or situational gaps); (b) robustness, which is related to the stability of the study results across a range of minor changes in the study design; (c) impact of findings, which is related to convincing power of findings of an experiment, which could make practitioners in real world settings to change, adapt or adopt new practices based upon these findings; and (d) resource, which is related to the costs associated with undertaking of experimental inquiries.

4.1.4.4. Cycle of maturing knowledge. Krein and Knutson [ABO022] describe a unified framework for research methodologies in empirical software engineering, which is called Cycle of Maturing Knowledge (CMK). Two of the central objectives in building this framework are related to replications: (a) "to refine the concept of replication as it applies to SE and place it appropriately within the unified framework of research methodologies"; and (b) "to identify the role that replication plays in the knowledge building process". According to the authors, CMK identifies three areas for advancement in SE empirical research: (1) in addition to strict replication, we must explore additional methods for internal and external validation of observations; (2) we need to continue developing methods for conducting differentiated replications, such that results can be synthesized; and (3) we must spend more time synthesizing, that is, iterating the knowledge-building cycle, conducting "families" of "differentiated replications".

4.1.4.5. SOFAS applied to mining studies. Ghezzi and Gall [ABO031] propose a framework called SOFAS (SOFtware Analysis Services) as a platform to assist the replication of Mining Software Repository studies. According to the authors, "SOFAS is a platform that enables a systematic and repeatable analysis of software projects by providing extensible and composable analysis workflows. These workflows can be applied on a multitude of software projects, facilitating the replication and scaling of mining studies".

4.1.4.6. Brooks' replication framework. Brooks et al. [ABO036] extended the framework for software engineering experimentation proposed by Basili et al. [22] to differentiate between various types of internal and external replications with respect to their confirmation power. The framework is based a characterization of internal replications (series of experiments carried out by the same researchers) and external replications (carried out by experimenters that are independent of the original researchers). A second component of the framework is the classification of the variations between original study and replication. According to Brooks' framework, the elements from Basili's framework [22] (method, task, and subjects) can vary in a replication in three ways: similar (as close as possible as the original), alternative (different from the original), and improved (based on the original with enhancements).

4.1.4.7. Summarizing frameworks. This topic was difficult to summarize due to the diverse nature of the proposed frameworks. In Table 9, we present each framework with its central goal, the main recommendations addressed, and the replication processes that are somehow related to the framework.

4.1.5. Tools

In the context that involves tools to support replications, Scatalon et al. [ABO028] describe a workflow to generate lab packages based on ontology of controlled experiments domain, which can be used to share understanding about the original experiment. Replicators can use this tool while performing a replication. This ontology deals with different types of lab packages, which are important tools for supporting and facilitating replications by reducing the amount of effort required from independent researchers to understand a prior experiment. The input of this workflow is any data set describing an experiment. For example, the information to generate lab packages can be extracted from the description of an experiment found in the literature.

Gallardo [ABO029] presents a research proposal to build a web environment to support the process of replication based on configuration management and product line ideas. When completely developed,[2] this tool shall archive and manage experimental materials to allow replications with massive experimental data storage.

[2] This tool was under development until the date of the publication of [ABO029].

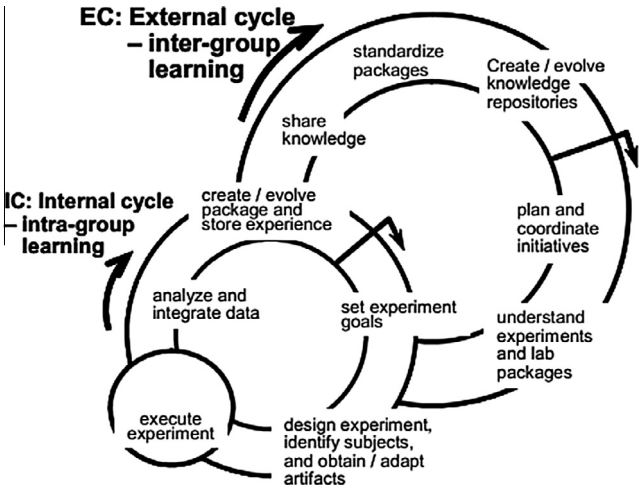

Fig. 6. The framework FIRE (extracted from Mendonça et al. [ABO001]).

According to the author, the platform will be accessible to several research groups working together on the same families of experiments and it shall help to transfer information to experimenters on how to correctly apply instruments and materials. Further, it shall provide an infrastructure for storing information and identifying replications in order to guarantee their integrity, reliability, and traceability for reuse within the experimental research cycle.

Finally, Squire [ABO033] addresses problems related the replicability of empirical software engineering studies of mailing list archives. Currently, mailing lists are archived in several places online and the research teams wishing to perform empirical studies on their contents must design their own solution for collecting, storing, and cleaning data. Consequently, such studies are prone to be more difficult to replicate. The tool proposed by Squire [ABO033] is supposed to improve replicability of email archive-based software engineering research by standardizing some of the choices about how emails are processed in the empirical studies. It also uses a common method to provide a familiar and standardized vocabulary for communicating how the email messages were collected and stored.

4.1.5.1. Summarizing tools. Similarly to the previous topic, this topic was also difficult to summarize due to the diverse nature of the tools proposed. In Table 10, we present each tool with its central goal, the main recommendations addressed, and the replication processes that are somehow related to the tool.

4.1.6. Guidelines

Guidelines to report original experiments and replications are equally important to achieve successful replications. The former should emphasize the right amount and structure of information to make replication easier. Complementary, specific guidelines to report replications are important to allow consistent comparison between the original study and its replications. In both cases, such guidelines should assist original experimenters and replicators in identifying the necessary information and to assure that this information is published in a consistent and standardized way. However, only 2 papers addressed this topic.

4.1.6.1. Guidelines to report experiments. Kitchenham et al. [ABO005] used perspective-based reading to evaluate the reporting guidelines for controlled experiments in software engineering proposed by Jedlitschka and Pfahl [23]. One of the perspectives used in the evaluation was the Replicator, that is, the researcher reading a report of an experiment with the goal of replicating the study. The ABO study argues that the reporting guidelines should receive amendments to support the needs of researchers aiming at performing replications. However, the study does not provided specific directions about which amendments should be built into the existing guidelines.

4.1.6.2. Guidelines to report replications. Carver [ABO018] presents a preliminary proposal of reporting guidelines specifically aimed at replications, with the goal of standardizing how replications are reported in the literature. The author points out the existence of guidelines for reporting controlled experiments and case studies, but none specifically for reporting replications. The proposal considers some elements that should be included in any report describing an experimental replication:

(a) *Information about the original study:* research question(s); participants; design; artefacts; context variables; summary of the results.
(b) *Information about the replication:* motivation for conducting the replication; level of interaction with original experimenters; changes to the original experiment.
(c) *Comparison of results to original:* consistent results, differences in results.
(d) *Drawing conclusions across studies:* combining conclusions from the original study with conclusions from the replication.

Carver et al. [4] conducted a survey with authors and reviewers that participated in the 2014 Special Issue about replications in software engineering, in the Empirical Software Engineering Journal. In this survey, they found that the vast majority of the authors attempted to follow the guidelines (17/18) and 75% (16/24) of the reviewers whose papers followed the guidelines reported that the guidelines made the papers easier to review.

4.1.6.3. Summarizing guidelines. Reporting guidelines for empirical studies in general should explicitly deal with the issues that must be addressed during design, development, and report of the studies to increase their replicability. Issues related to the detail and precision of the information have already been discussed in the Recommendations and Processes topics. However, existing guidelines in software engineering do not satisfactorily address replication issues explicitly.

As for the report of replications, the work of Carver [ABO018] provides a good starting point and has been tested in practice. We believe that these guidelines should be further evaluated and refined, and its use more broadly disseminated.

4.1.7. Result combination

Two studies address the problems related to combination of the results of the original study and its replications. Dieste et al. [ABO019] discuss how to used meta-analysis to aggregate results from "useless" replications, that is, experiments run with few subject and, therefore, with low statistical power. The authors argue that meta-analysis can be used to aggregate results of these small-scale replications and achieve results equivalent to experiments run on large samples of subjects.

Runeson et al. [ABO035] discuss the problems of unwanted variations among experiments and how they make it difficult or even impossible to aggregate or compare results. The authors argue that "reducing the complexity of software engineering experiments should be considered by researchers who want to obtain reliable and repeatable empirical measures".

Table 9
Summary information about frameworks.

ABO study	Framework	Objective	Recommendations addressed	Related processes
[ABO001]	FIRE	A framework to address knowledge sharing issues both at the intra-group (internal replications) and inter-group (external replications) levels	Understanding the original study	Communication
[ABO004]	Family of experiments	A framework based on the goal-question-metric template (GQM) to organize sets of similar experiments	Variations	Non-exact replications
[ABO017]	Miller's replication framework	A replication framework that identifies "four key dimensions of replication effectiveness" that should be balanced in a replication, or series of replications: (a) Existential realism (b) Robustness (c) Impact of findings (d) Resources	Variations	
[ABO022]	Cycle of maturing knowledge	Unified framework for research methodologies in empirical software engineering	Variations	Non-exact replications
[ABO031]	SOFAS applied to mining studies	A framework to assist the replication of Mining Software Repository studies	Specific recommendations presented by [ABO008] and [ABO025]	
[ABO036]	Brook's replication framework	Framework based on a characterization of internal and external replications that presents classification of the variations between original and replication	Variations	Communication

4.1.8. Miscellaneous

The three studies in this category present distinct issues related to replications that does not fall in any other category. We summarized their main contributions:

- Mäntylä et al. [ABO023] discuss that the concept of replication in software engineering should be more broadly understood, in particular in the replications of empirical studies other than experiments, such as case studies and surveys. The authors performed a literature review on articles published in SE journal and argue that various published studies should be considered replications although their authors did not classify them as such. The authors contend that with a broader view the number of replications in software engineering would be considerably larger.
- Sjøberg et al. [ABO026] discuss the importance of replications to the progress and future of empirical software engineering as science. They argue that in the future, as the empirical SE field matures, replication would become more frequent and systematic.
- Callele et al. [ABO032] address the problem of replication in industrial practice. In particular, they emphasize that industrial practice focus on distinct issues than academics when design an experiment, and that such distinctions must be addressed when performing a replication in a different context. The authors also discuss the problems of justifying the application of results obtained in academic setting in the industrial settings.

4.2. The use of ABO studies in the papers reporting replications

RQ5: Which ABO studies are cited by the papers that reported replications?

To answer this question, we verified the list of references of 135 papers reporting replications, published between 1994 and 2012, which were identified by da Silva et al. [5] and Bezerra and da Silva [6]. We found that fewer than 39% (52/135) cite ABO studies. Overall, these 52 replication papers make 92 citations of 18 ABO studies.

In Table 11, we map the replication papers and the cited ABO studies. We organized the information according to the topics of the ABO studies to facilitate the references to RQ4, and in each topic, ABO studies are presented in decreasing order of the total number of citations. In Appendix B, we present the complete list of replication reports that referenced ABO studies.

ABO studies in the topic Tools have not been cited by the replication papers. The most cited topic was Framework. Four ABO studies classified in this topic have been cited 51 times, representing 55% (51/92) of the citations. The paper by Basili et al. [ABO004], published in 1999, has been cited by 25% (35/135) of the papers reporting replications. As presented in the answer to RQ4, this ABO study proposes a framework to conduct families of experiments. The references to this study are spread over the years, from 2001 until 2012.

Brooks et al. [ABO036] present the second most cited ABO study. This study also proposes a framework to conduct

Table 10
Summary information about tools.

ABO Study	Tool	Objective	Recommendations addressed	Related processes
[ABO028]	Workflow to generate lab packages	To share understanding about the original experiment	Understanding the original study	Communication
[ABO029]	Web environment to support replication based on configuration management and software product line	To archive and manage experimental materials to allow replications with massive experimental data storage	Understanding the original study Variations	Communication
[ABO033]	Tool for archive-based software engineering research	To improve replicability by standardizing some of the choices about how emails are processed in the empirical studies	Understanding the original study Variations	

replications based on an extension to the framework for experimentation in software engineering proposed by Basili et al. [22]. This research has been cited in 5% (8/135) of the replication papers.

The two most cited ABO studies propose conceptual frameworks for performing replications. This is an indicative that the researchers performing replications acknowledge the importance of conceptual frameworks for their research work.

The category of ABO studies that propose or discuss processes to support replication is the second most cited. In this category, both Shull et al. [ABO015] and Juristo and Vegas [ABO014] have been cited in 5% (7/135) of the replication papers. Thus, issues related to the process (in particular, the communication between researchers) also appear to be a concern for researchers performing replications. The next most cited ABO study [ABO017] also proposes a conceptual framework and 6 replication papers have cited it. The remaining ABO studies have been cited in three or less replication papers. In particular, replication papers have cited only two ABO studies in the category of recommendations ([ABO008] cited three times and [ABO016] cited only once).

Juristo and Vegas [REP006FE] cite 10 ABO studies, of which five are self-references ([ABO014], [ABO010], [ABO011], [ABO019], [ABO003]) and two of them were the single references to the ABO studies [ABO019] and [ABO003]. Similarly, Laukkanen and Mäntylä [REP003FE] make the single self-reference to the study [ABO023]. Considering that these self-references add very little to disseminating knowledge outside the same research group, we could say that these three papers did not have a broader impact in the replication work so far. Overall, we argue that the impact of the ABO studies on the replication papers is small, with only a few studies being cited and with the vast majority of the ABO studies not being used or used by very few replications.

RQ6: How the results or propositions presented in the cited ABO studies have been used in papers that report replications?

We analyzed how the replications have used the ABO studies. We interrogated the replication papers looking for evidence that allowed the classification of this use in three categories:

- *Level 2 (fully-used):* the main contribution of the ABO study, related to its central topic as classified in RQ4, are indeed used in the replication work.
- *Level 1 (used):* secondary results of the ABO study, those results not related to the main topic, are used in the replication paper, but the main contributions of the ABO study (related to the topic) has not been used.
- *Level 0 (cited):* this level refers to a superficial use, in which the ABO study is cited (mainly as a corroboration for an argument or as part of a justification or motivation in the replication paper) but the results or propositions of the study are not neither applied nor used in the replication work.

In Table 12, we present examples of the uses of [ABO004] in each level to show how we coded the answers to this research question. The first and the second authors performed this coding, double-checking the results for higher consistency.

The majority of the uses are of Level 0 (cited), with 70% (64/92) of the citations. Uses of Level 1 (used) represent 19% (17/92) of the citations, and the remaining 13% (11/92) are uses of Level 2 (fully-used). Tables 13 and 14 show the Level 2 (fully-used) and Level 1 (used) references to the ABO studies, respectively. Regarding uses of Level 2 (fully-used), Replication Types is the most cited topic (4) followed by Frameworks (3), and Guidelines (2). ABO studies in the categories of Recommendations and Processes received one reference each of Level 2 (fully-used).

Based on the answers to this research question, we argue that the level of use of ABO studies is superficial, with only a minority of the replications really applying the contributions related to the main topic of the cited studies. It seems that the effort of investigating issues and proposing contributions to enhance the quality of the replication research in software engineering, materialized by the ABO studies, has not made a strong impact in the replication work so far.

5. Discussions

Our goal, in this review of research about replications in empirical software engineering, is to plot the general landscape of the

Table 11
References of ABO in replications performed between 1994 and 2012.

Topic	ABO study	Replication papers	Total of citations
Frameworks	[ABO004]	[REP007], [REP012], [REP016], [REP104], [REP002FE], [REP004FE], [REP006FE], [REP013FE], [REP019FE], [REP022FE], [REP023FE], [REP025FE], [REP026FE], [REP027FE], [REP033FE], [REP035FE], [REP003], [REP019], [REP024], [REP026], [REP027], [REP035], [REP038], [REP039], [REP041], [REP103], [REP118], [REP122], [REP123], [REP124], [REP125], [REP126], [REP129], [REP130], [REP131].	35
	[ABO036]	[REP003], [REP019], [REP033], [REP048], [REP051], [REP123], [REP124], [REP126]	8
	[ABO017]	[REP009], [REP020], [REP021], [REP024], [REP039], [REP030FE]	6
	[ABO001]	[REP009], [REP006FE]	2
Processes	[ABO015]	[REP019], [REP036], [REP098], [REP118], [REP120], [REP015FE], [REP036FE]	7
	[ABO014]	[REP012], [REP120], [REP003FE], [REP006FE], [REP024FE], [REP036FE], [REP038FE]	7
	[ABO007]	[REP006FE]	1
	[ABO003]	[REP006FE]	1
	[ABO011]	[REP006FE], [REP030FE	2
Replication Types	[ABO012]	[REP003], [REP003FE], [REP004FE], [REP015FE], [REP027FE], [REP030FE], [REP036FE]	7
	[ABO010]	[REP003FE], [REP006FE], [REP030FE]	3
	[ABO021]	[REP003FE], [REP006FE]	2
	[ABO013]	[REP030FE], [REP036FE]	2
Recommendations	[ABO008]	[REP015FE], [REP024FE], [REP004FE]	3
	[ABO016]	[REP016]	1
Guidelines	[ABO018]	[REP006FE], [REP028FE], [REP038FE]	3
Result Combination	[ABO019]	[REP006FE]	1
Miscellaneous	[ABO023]	[REP003FE]	1

Table 12
Examples of uses of [ABO004].

Replication paper	Level of use	Justification	Excerpt from paper
[REP019]	Level 0 (cited)	Only refers to the importance of family of experiments	"Basili and Lanubile expand the idea of replicated experiments into the concept of a 'family of experiments'..."
[REP024]	Level 1 (used)	Uses the classification of strict replication from [ABO004], but does not apply the full proposed framework	"This replication can be classified as what is known in literature as 'strict replication' in that it does not vary any of the research hypotheses..."
[REP035FE]	Level 2 (fully-used)	The GQM framework is used as proposed in [ABO004]	"Our research objectives are outlined using the Goal/Question/Metric (GQM) framework defined by..."

body of work about replications and to complement the review of replications produced by da Silva [5] and Bezerra and da Silva [6]. In this section, we discuss our results, their implications for software engineering research, and the limitations of our work. We also briefly discuss the results of the RESER workshop with respect to the published studies about replication.

5.1. Strengths and limitations of research about replication in SE

Our results show that the research about replication of empirical studies in software engineering has evolved, but several limitations have yet to be addressed. We summarize the main strengths and limitations of the set of ABO studies in this section, starting with the strengths:

- The number of ABO studies is increasing, in particular since 2010, after the first edition of the RESER workshop. The average number of publications per year grew from one publication between 1996 and 2009 to seven between 2010 and 2013.
- Important topics, such as replication types, replication processes, and conceptual frameworks, have been studied.
- Although the number of researchers and, in particular, research groups interested in studies about replication could be considered small, they include experts in empirical software engineering research. The identification of researchers and organizations, interested in the topic of replications, could foster collaborations and increase the number of researchers interested in this area of investigation.
- Just fewer than 38% (51/135) of the replications published between 1994 and 2012 cite ABO studies. Eighteen ABO studies in total (approximately 50%) are cited in these 51 replications. This shows that the ABO studies were consulted by almost half of the research groups performing replications. This also indicates that half of the ABO studies were not used in the development of replications so far.

However, important limitations still exist that raises some concerns:

- The absolute number (37) is small considering the breadth of issues and research topics that are related to replication of empirical studies in general and in SE in particular.
- The ABO studies are spread over a large range of research topics (see the answer to RQ4) and even the studies classified in the same topic do not address the same research problem, making it difficult to synthesize results toward a better understanding of problems related to replication.
- Almost 30% (11/37) of the ABO studies do not provide a definition of what constitute a replication and those that do provide, use different and sometimes non-rigorous and incompatible definitions (see answer to RQ3).
- Only one ABO study researched the issue of replication types and roles explicitly. Other studies propose replication types as part of their secondary goals. Overall, the types and roles proposed are not uniform. We found a lack of standardization about replication types in the ABO studies.

Table 13
Uses of ABO studies of Level 2 (fully-used).

ABO study	Replication paper	Justification	Excerpt from paper
[ABO004] Framework	[REP035FE]	Framework used	"Our research objectives are outlined using the Goal/Question/Metric (GQM) framework defined by Basili et al.1999."
[ABO012] Replication Types	[REP015FE]	Replication Types used	"Our study can be further defined as an exact replication..."
[ABO012] Replication Types	[REP027FE]	Replication Types used	"...this replication study can be classified ... as a dependent replication."
[ABO012] Replication Types	[REP036FE]	Replication Types used	"Our goal was to conduct a dependent replication by following the procedures of the original experiment as closely as possible..."
[ABO014] Processes	[REP038FE]	Process (partially) used (replication type used)	"we performed a ... non-identical replication" "We observed the recommendations of Juristo and Vegas for the inclusion of variations in the survey design and to compare the results...."
[ABO016] Recommendations	[REP016]	Recommendations were followed	"...suggested, simple studies rarely provide definite answers. Following these suggestions, we have carried out a family of experiments."
[ABO017] Frameworks	[REP030FE]	Framework (partially) used	"We followed recommendations by Miller to change some elements compared to the previous studies to check the stability of the results. ... we also stick to the same report structure as much as possible..."
[ABO018] Guidelines	[REP006FE]	Guidelines followed	"The replication is described along the lines in..."
[ABO018] Guidelines	[REP038FE]	Guidelines followed	"We also followed the guidelines of Carver to write the report of the replication"
[ABO021] Replication Types	[REP003FE]	Replication Types used	"In the field of experimental software engineering our work should be viewed as reproduction"
[ABO036] Frameworks	[REP033]	Framework used	"The replication framework of Brooks et al. provides a classification. Accordingly, we would classify this internal" replication as (similar, improved, improved)."

5.2. Implications for research

We believe that our findings and also the gaps we identified in the literature have important implications for empirical research in software engineering. We summarize the five questions our implications raise, which require attention from the research community.

Question 1: What should be considered a replication?

As can be seen from the answer to RQ3, there is no commonly accepted and widely used definition of replication. In fact, only 38% of the ABO studies provide some statement that the authors called a definition and four of them use some definition from the scientific literature. Just above 32% (12/37) of the ABO studies provide definitions through classifications or taxonomy of replications, not using a direct and intentional definition. Some of the definitions are not mutually consistent. For instance, Krein and Knutson's notion of independent replication [ABO022] implies no reference to an original study. This notion is inconsistent with most of the other definitions, for instance from [ABO011], [ABO014], and [ABO024].

We need an intentional or connotative definition of replication, i.e., a definition that specifies necessary and sufficient conditions for a study B to be considered a replication of a study A. In this direction, we advocate that for some study B to be considered a replication is must make an explicit reference to another study A (often called the original or baseline study), which B intends to be replication of. This is a necessary condition, but is not sufficient. Schmidt [2] uses the concept of primary information focus defined by Hendrick [21] to propose an intentional definition of replication as follows: "*B is a replication of A if A's primary information focus is re-established in B*".

Possible directions for further research

In software engineering research, we must study what are the conditions under which we consider that the primary information focus of experiment A is successfully re-established in another experiment B. However, this is not an easy research problem. For instance, Runeson et al. [ABO035] conducted a comparative analysis of three dissimilar experiments on unit test versus inspections. Two experiments are similar and use the same inspection technique, whereas the third studied uses a different inspection technique. The authors consider inspection (in general) to be the primary information focus, thus considering the experiment that used a distinct inspection technique as a replication. Another view would be to consider the technique of inspection as the primary information focus, thus considering the third experiment not to be a replication of the other two.

Question 2: What should be considered a successful replication?

This issue has been superficially addressed by only a few ABO studies. We believe this is an important issue and a gap in the research that requires attention.

One could be tempted to understand the success of a replication in terms of its relationship with the findings from the original study. In this sense, a simple definition of success could be: B is a successful replication of A if B confirms A's results if and only if A's results are actually true, and does not confirm A's results if and only A's results are actually not true. However, this is a useless definition because we run B to check whether results from A are actually true or not true. Therefore, we need a definition of success that does not rely on knowing upfront whether the original study is valid or not.

Success, in general, is related to goal or purpose. The starting point to answer this question, therefore, is the establishment of the goal of the replication and criteria to define to what extent this goal was achieved. According to Schmidt [2] and Hendrick [21], different types of replications are more adequate to achieve certain goals or purposes. Therefore, one of the primary success criteria is the conformity between type and purpose. For instance, if the goal of a replication is to control for fraud, then we must perform an external replication in which the team performing the replication is independent from the team the performed the original experiment. An internal replication (type) is not adequate to control for fraud (purpose), and thus no internal replication can successfully replicate an experiment if the goal is to control for fraud.

Few ABO studies that address the topic of replication type also discuss replication function or purpose. Goméz et al. [ABO010] discuss both replication types and replication function. Further, they also discuss the elements that can vary in a replication. However, they do not relate type, variation, and function. Therefore, no conformity criterion was proposed in their study.

Schmidt [2] proposes a framework that relates replication type, variations, and function of replication. We believe that this framework could be further investigated and adapted to the specific needs of SE research.

Possible directions for further research

In the direction of a framework that relates replication types and functions, Goméz et al. [24] recently presented a classification of replications. According to the authors, "the proposed classification can be used to identify changes in a replication and, based on these, understand the level of verification" that a replication provides of the baseline (original) experiment. This study is a starting point toward a suitable framework to understand the conditions under which a replication is successful. Its adequacy still requires further empirical evaluation by applying the classification to existing and new replications.

Question 3: What are the types of replications and their functions?

This question is addressed by five of the ABO studies. However, only one of them has this question as its central goal [ABO010]. The other studies that propose classifications or types of replications do this as part of other goals such as the definition of frameworks [ABO004] or to classify articles analyzed in a literature review [ABO037].

As we noted in the answer to RQ4 (in the topic Replication Types), there is no consensus on terminology and classification scheme for replications in software engineering. Baldassarre et al. [20] address this problem and propose a two dimensional taxonomy that relates procedures with the experimenters (people) performing the procedures. However, this proposition does not address the relationship between type and function. The taxonomy proposed by Baldassarre et al. [20] is consistent with the framework proposed by Schmidt [2], and it could be augmented to address the type-function relationship problem.

As mentioned above, Goméz et al. [24] present a classification of replications based on a deep analysis of the literature in software engineering and other fields. This classification relates types of replications with the different verification goals or functions. It is a comprehensive classification scheme that still needs to be tested in practice.

Possible directions for further research

As discussed above, the relationship between type and function of a replication is important in the definition of success criteria to

Table 14
Uses of ABO studies of Level 1 (used).

ABO Study	Replication Paper	Justification	Excerpt from Paper
[ABO004] Framework	[REP007]	Use ABO004 to define what is an experiment and Replication Types used	"An experiment may be a part of a common family of studies..." "We have carried out a strict identical replica ..."
[ABO004] Framework	[REP002FE]	Replication Types used	"The family consisted of a controlled experiment (...) and two strict replications of it (...), in which none of the dependent or independent variables vary"
[ABO004] Framework	[REP019FE]	Replication Types used	"The second experiment was a differentiated replication"..."introduces variations in essential aspects of the experimental conditions"
[ABO004] Framework	[REP033FE]	Replication Types used	"In this kind of replication, variations in essential aspects (e.g., different kinds of participants) of the original experimental conditions are introduced"
[ABO004] Framework	[REP003]	Replication Types used	"In order to confirm the results obtained in the first experiment, we replicated this experiment under the same conditions (strict replication), changing only the subjects ..."
[ABO004] Framework	[REP024]	Replication Types used	"This replication can be classified as what is known in literature as strict replication"..."in that it does not vary any of the research hypotheses and it reuses instrumentation of the original experiment."
[ABO004] Framework	[REP027]	Discusses the use of students as subjects.	"...students can play a very important role in experimentation in the field of software engineering"
[ABO004] Framework	[REP035]	Importance of families of experiments Replication Types used	"...a family of experiments permits the accumulation of the knowledge needed to extract significant conclusions" "we replicated this experiment under the same conditions (strict replication)"
[ABO004] Framework	[REP103]	Replication Types used	"According to the terminology on replications introduced by Basili et al. (1999, p. 469), the second experiment is a replication that does not vary any research hypothesis."
[ABO004] Framework	[REP124]	Replication Types used	"does not vary the hypotheses of the basic experiment"
[ABO004] Framework	[REP126]	Used to justify the replication used	"tested exactly the same hypotheses as the basic Experiment"
[ABO036] Framework	[REP003]	Used the ABO to corroborate the importance of sharing material to allow external replications.	"As the diffusion of the experimental data is important to external replication of the experiments (Brooks et al., 1996), we have put all the material of this experiment on a web site"
[ABO036] Framework	[REP123]	Used the ABO to corroborate the importance of sharing material to allow external replications.	"As the diffusion of the experimental data is important to the external replication"... "of the experiments we have put all the material of this experiment on our web site"
[ABO036] Framework	[REP126]	Justifies which were the conditions of the replication.	"The experiment we conducted in Italy tested exactly the same hypotheses as the basic experiment (Brooks et al.1996; Basili et al. 1999)".
[ABO017] Framework	[REP039]	Replication Types used	"This is what we intend to do, we will test the same hypothesis as Jørgenson et al., using different datasets. Miller suggests an exact replication, using the same data will set up a correlation between the studies..."
[ABO001] Frameworks	[REP009]	Replication Types used	"... a replication can be internal (conducted by the same research group of the original study) or external (in a different context). This paper presents an external replication..."

evaluate replications. We believe that the works of Baldassarre et al. [20] and Goméz et al. [24] provide good starting points to study the issues related to success criteria of replications.

Question 4: How should replications be performed?

This question can be addressed from at least two complementary perspectives. First, there is the issue of how to perform a single replication of an original study successfully. Guidelines and rigorous process definitions, consistent with the success criteria related to types and functions of replications discussed above, are important elements to address issues in this perspective.

A second perspective is related to knowledge building from several replications or families of experiments. This is related to follow a strategy or systematic procedure to define the sequence of replications based on the intended variations from the original study. One goal in following such procedure is to generate knowledge while also reducing the risk of running an unsuccessful replication.

Based on the framework proposed by Schmidt [2], replications can be designed to achieve three central purposes: (1) verify various aspects of the validity or truth of the original results, e.g., control for sampling error, control for artefacts, and control for fraud, (2) to generalize the original results to different populations extending the knowledge provided by the original study, and (3) to verify hypothesis through different procedures to create deeper understanding about the investigated phenomena.

Hendrick [21] defines the concept of systematic replications as a strategy that starts from the original study and through intended variations on the elements of the experiment, moves through each of the three purposes described by Schmidt [2]. Hendrick [21] and Schmidt [2] argue that the risk of running an unsuccessful replication increases as we move from the purposes related to item (1)–(3) due to the increase in the variations. In their point of view, a systematic procedure of replication must start with lower risk replications and move to higher risk ones as our understanding of the problem increases.

Possible directions for further research

Software engineering research would benefit from a systematic approach to perform replications. Further, as discussed by Kitchenham [ABO013] we should base replications on well-founded theories to decrease risk and improve our understanding of the phenomena under study. Goméz et al. [24] discuss the role of systematic replications in the context of their classification proposal and argue that "the replication types proposed for different verification purposes gain power if they are used through systematic replication". However, the authors do not explicitly propose a systematic replication process. Future research could investigate how to use their classification of replications to propose such process.

Question 5: How should replications be reported?

Finally, successful replications should be successfully reported. In fact, a good report of a replication could also be considered a criterion for a successful replication. It is not enough to do a good job performing the study; it is necessary to perform a good job also producing a quality full report of the replication. What it means for a report to be of quality is still a problem that requires research.

Carver [ABO018] presents a preliminary proposal of reporting guidelines for replications. da Silva et al. [5] use this proposal to create a set of quality criteria to evaluate replications. The result of the quality evaluation performed by da Silva et al. [5] shows that, in general, the replication papers do not comply with the prescriptions of the reporting guidelines.

Possible directions for further research

Reporting guidelines should emphasize the need of the consistent use of terminology related to replication. In particular, replication type, replication function or purpose, and types of variation of design elements should be standardized and used consistently in replication reports. As discussed above, this standardization will require an effort from the research community to agree on taxonomy and consistent terminology for experiments and their replication. Considering the positive evaluation of Carver's guidelines reported by Carver et al. [4], we believe that these guidelines should be extended to include standardized terminology and its use should be stimulated by conference program committees and journal editors.

5.3. The results of the RESER workshop

The International Workshop on Replication in Empirical Software Engineering Research (RESER) is a venue for publication and discussion of issues related to replication in software engineering. According to the workshop website "the primary goal of this workshop is to raise the quality and amount of replication work performed in software engineering research". The first edition of the workshop was held as a joint event of the ICSE'2010 conference, in Cape Town, South Africa. After this, two other editions were held jointly with the ESEM symposium in 2011 in Banff, Canada, and in 2013, in Maryland, US.

A full analysis of the impact of journal and events from which the ABO studies were selected is outside the scope of our mapping study. This includes an impact analysis of the RESER workshop. However, because RESER is the only event that specifically targets replication in software engineering research, we believe that adding a summary of its results is relevant.

The first two editions of RESER (2010 and 2011) published five replications that were analyzed by da Silva et al. [5] (one replication from RESER 2010) and Bezerra and da Silva [6] (four replications from RESER 2011). RESER 2013 published 9 replications that were not analyzed in the two mapping studies because they were published more recently.

As mentioned in Section 4.1 (RQ1), RESER is the venue in which almost 27% (10/37) of the ABO studies are published (8 in RESER 2010 and 2 in RESER 2013). In Table 15, we present the list of ABO studies published at the RESER workshops, together with information about their citations and use in the replication papers.

Four ABO studies published at RESER are cited seven times in four distinct replication papers. Three of these citations are self-references. One replication paper (REP006FE) cite three ABO studies, including the unique self-reference to [ABO019]. This seems to indicate that RESER has had a small impact on the replication research and that the impact is restricted to a small set of researchers. However, this conclusion can be biased because of the timeframe of the analysis. ABO studies were published at RESER in 2010 and 2013, and the replication papers were selected until 2012. Therefore, the small number of references to RESER papers could be explained by the short timeframe between the publication of the ABO studies and the REP papers. This could also explain the apparently large number of self-references.

Regarding the coverage of research topics, ABO studies published at RESER address seven out of the eight topics identified in our study. This shows a very comprehensive coverage of research topics.

We can also observe that 70% (7/10) of the RESER's ABO studies are authored by at least one researcher that also performed a replication. This is significantly different from the numbers found when considering the entire set of ABO studies, as discussed in Section 4.1 (RQ2). This may indicate that participants at RESER form a community of researchers that are actively involved in performing replications and also studying issues about replications.

As mentioned above, the guidelines developed by Carver [ABO018], presented at RESER 2010, have been positively evaluated in practice. Carver et al. [4] performed a survey with authors and reviewers of replication papers, and found that the vast majority of the authors attempted to follow the guidelines (17/18) and 75% (16/24) of the reviewers whose papers followed the guidelines reported that the guidelines made the papers easier to review.

This brief account of the results of the RESER workshop is not intended to be a full analysis of the impact of the workshop, as mentioned above. Nevertheless, we belief that RESER has been an important venue for discussions about replication in software engineering and that some of the ABO studies published there provided important starting points and conceptual underpinnings for further research, [ABO018] being a case in point.

5.4. Limitations of this study

The most common limitations that may occur in a mapping study are limited coverage, possible biases introduced in the selection process, and inaccuracies during data extraction, analysis, and synthesis. Thus, these are also the main possible limitations of this study. However, some measures were taken to try to minimize these limitations.

The combination of automatic search in several engines and manual search on relevant publications improves the coverage of the review. The manual search reduced the possibility of missing ABO studies that did not use "replication" or any of the synonyms we used in our automatic search.

The first and third authors performed all data extractions independently and the results were compared and combined. The disagreements that emerged in the combination of the data extraction were resolved in a consensus meeting with the second author, which is a specialist in systematic literature reviews. The same

Table 15
Summary of ABO studies published at RESER workshops.

ABO Study	Topic	Citations	Level of use
[ABO018]	Guidelines	[REP006FE]	Level 2
		[REP028FE]	Level 0
		[REP038FE]	Level 2
[ABO019]	Result Combination	[REP006FE]	Level 0
[ABO020]	Recommendations		
[ABO021]	Replication Types	[REP003FE]	Level 2
		[REP006FE]	Level 0
[ABO022]	Frameworks		
[ABO023]	Miscellaneous	[REP003FE]	Level 0
[ABO024]	Recommendations		
[ABO025]	Recommendations		
[ABO030]	Replication Types		
[ABO033]	Tools		

process was used in the thematic analysis that classified the ABO studies into the categories and to produce the summaries used to answer RQ4 and in the other data analysis and synthesis.

6. Conclusions

In this article, we presented a review of 37 papers reporting studies on concepts, classifications, guidelines, frameworks, and other topics about replication in Software Engineering published between 1996 and 2013. We used the papers selected from two mapping studies that covered the period between 1996 and 2012, and from a search procedure performed by the authors to cover the year 2013. Over 67% (25/37) of the papers are published in conferences and workshops (19 full and 6 short papers). Just under 25% (9/37) are journal papers.

Considering the importance of replication for empirical software engineering research and the breadth of topics related to replication work, we argue that the number of ABO studies is very small. Further, the studies are spread over eight different topics and the papers in each topic almost never address the same research problem, making it very difficult to integrate their results and to build a more solid knowledge body to support the development of replications in SE.

It seems that the knowledge presented in the ABO studies has practical value. Over 75% (28/37) of the ABO studies were performed by at least one researcher that had been also involved in the development of a replication. We argue, therefore, that research groups with some practical experience in performing replications developed the majority of ABO studies and this could increase the applicability their results and propositions in practice.

However, the dissemination of the knowledge generated by the ABO studies has been limited, so far. Just under 50% (18/37) of ABO studies are cited by 51 replication papers, and three of them are cited only once by the same authors that published the ABO study (self-reference). Further, over 60% of the replication papers do not cite any ABO study.

Regarding the authorship of the replication papers that cite ABO studies, 40% (21/52) of them are authored by at least one researcher that also published an ABO study. Thus, the replication papers that cite ABO studies and have been published by researchers that did not performed an ABO study represent only 23% (31/135) of the total. We argued before that the number of researchers actively engaged in producing and publishing ABO studies is small. This would not be a problem if their knowledge had been successfully spread to other researchers engaged in performing replications. The above figures seem to indicate that this is not the case. The potential impact of this issue is that replication work is being conducted without a solid conceptual background about replications, with could result in low quality results. Conference program committees and journal editorial boards could provide guidance to potential authors without practical or theoretical background about replications by pointing to the ABO studies for reference.

Further, based on the answers to RQ6, we argued above that the use of the contributions from the ABO studies is superficial, with only a minority of the replications really applying the contributions of the cited studies. It seems that the effort related to the study of topics that could improve the role of replications in empirical software engineering research, materialized by the ABO studies, has not made a strong impact in the replication work so far.

Sjøberg et al. [25], toward the realization of their vision about the future of empirical methods in software engineering research, propose a target related to improving the quality of empirical studies: by the year 2025 "Replications and triangulation of research designs are [should be] frequently used means for achieving robust results". From the results of our study, complemented by the results of the two mapping studies [5,6] we argue that we are still a long way from this target. Further, we are not consistently, as a research community, moving toward it.

The five questions discussed in Section 5.2 contribute to the definition of a common research agenda that could be used by the SE research community to work in the direction of a consistent body of knowledge about replications that could contribute to achieve that target. We believe that the development of this common agenda should be a collaborative effort, in the same spirit of the common research agendas proposed by Sjøberg et al. [25]. Communities like the International Software Engineering Network (ISERN) and the Workshops on Replications in Software Engineering Research (RESER) are natural venues to foster this collaboration. We hope that this article also contributed to this collaborative agenda.

Acknowledgments

Professor Fabio Q. B. da Silva holds a research grant from the Brazilian National Research Council (CNPq), process #314523/2009-0. Cleyton V. C. de Magalhães and Ronnie E. S. Santos are both master students at the Center of Informatics of UFPE where they receive a scholarship from CAPES and FACEPE (process #IBPG-0651-1.03/12), respectively. We would like to thank the anonymous reviewers of this article for their comments and constructive criticisms that lead to important improvements in the content and structure of this version of the paper.

Appendix A. Reference list of ABO studies

[ABO001] M. G. Mendonça, J. C. Maldonado, M. C. F. de Oliveira, J. Carver, S. C. P. F. Fabbri, F. Shull, G. H. Travassos, E. Hohn, V. R. Basili, A framework for software engineering experimental replications, in: 13th IEEE International Conference on Engineering of Complex Computer Systems, 2008. ICECCS 2008. IEEE, 2008, pp. 203–212.

[ABO002] A. Cater-Steel, M. Toleman, T. Rout, Addressing the challenges of replications of surveys in software engineering research, in: 2005 International Symposium on Empirical Software Engineering, 2005. IEEE, 2005, pp. 10-pp.

[ABO003] S. Vegas, N. Juristo, A. Moreno, M. Solari, P. Letelier, Analysis of the influence of communication between researchers on experiment replication, in: Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering, ACM, 2006, pp. 28–37. doi: 10.1145/1159733.1159741

[ABO004] V. R. Basili, F. Shull, F. Lanubile, Building knowledge through families of experiments, Software Engineering, IEEE Transactions on 25 (4) (1999) 456–473.

[ABO005] B. Kitchenham, H. Al-Khilidar, M. A. Babar, M. Berry, K. Cox, J. Keung, F. Kurniawati, M. Staples, H. Zhang, L. Zhu, Evaluating guidelines for empirical software engineering studies, in: Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering, ACM, 2006, pp. 38–47.

[ABO006] M. C. Ohlsson, P. Runeson, Experience from replicating empirical studies on prediction models, In: Proceedings of the 8th International Software Metrics Symposium, Ottawa, Ontario, Canada, IEEE Computer Society, 2002, pp. 217–226.

[ABO007] F. Shull, M. G. Mendonça, V. Basili, J. Carver, J. C. Maldonado, S. Fabbri, G. H. Travassos, M. C. Ferreira, Knowledge-sharing issues in experimental software engineering, Empirical Software Engineering 9 (1–2) (2004) 111–137.

[ABO008] G. Robles, Replicating MSR: A study of the potential replicability of papers published in the mining software repositories proceedings, in: Mining Software Repositories (MSR), 2010 7th IEEE Working Conference on, IEEE, 2010, pp. 171–180

Table 16
Translation table of the synthesis of recommendations.

Research method being replicated	ABO study	Study recommendations (individual study summarization)	Generic recommendation (first-level generalization)	Concepts (second-level generalization)
Experiments or quasi-experiments	[ABO009]	The replication of studies is only possible when all details of the original study are known (or at least easy to guess)	Details of the original study are necessary to make replications possible	Understanding the original study (details of original study available)
	[ABO016]	To perform pilot studies by the replication researchers to understand about the original study	Perform pilot studies to understand original study in details	Understanding the original study (pilot study)
		To have a local expert who understand the technology being studied;		
		The terminology used must be clearly defined and explained to subjects	Precise and unambiguous definitions of study details (terminology, tools, instruments, etc.) are necessary	Precise and unambiguous design and execution of the studies (original and replication)
Case studies	[ABO006]	Original studies should be reported with much more detail and openness, ultimately including publication of raw data	Details of the original study, including raw data, are necessary	Understanding the original study (details of original study available)
	[ABO020]	To integrate domain knowledge acquisition as part of the case;		
		The researcher performing the replication must adapt all investigative procedures of the original lab study design to align with the large-case study context;	Adapt the instruments and procedures to the new context of study	Variations (instrument and procedures)
		Due to the large difference between the two types of studies, it may be possible to extend the original lab study with new research questions by considering the case study environment;	Extend research goals and questions	Variations (research goals)
		Data collection in the large-scale case study should not put excessive effort on participants, it should focus on flexible schemes;		
		New threats to validity may occur in the large-scale study that has not been anticipated by analyzing the lab study threats	Evaluate the possibility of new threats to validity not investigated in the original study	Variations (context)
	[ABO024]	The main challenge in replicating industrial case studies is the opportunity for replication		
		Motivating other researchers to perform replications of case studies is difficult		
		Measuring the same dependent and independent variable in different contexts may be a challenge, as ''each case may lack some of the data sources or the data may be less reliable''	Different contexts can create measurement challenges	Variations (context)
Survey research	[ABO002]	Conduct research to verify if the survey instruments are up-to-date with current practice and if needed add questions to bring the instrument up-to-date	Update and possibly extend research instruments	Variations (instrument and procedures)
		The addition of more questions requires care in reporting comparisons between original and replicated surveys	Evaluate whether the change in the research instruments will impact the comparison between original and replicated study	Variations (instrument and procedures)
		Communication with the researcher that performed de original survey is encouraged to provide a clearer understanding of the motivation, context and limitations of the original study	Communicate with the original researcher to understand the original study in details	Understanding the original study (communication)
		Researchers should provide full detail of their surveys to facilitate replications	Provide full detail of study to facilitate replication	Understanding the original study (details of original study available)
Replications in mining software repositories (MSR)	[ABO008] and [ABO025]	To present detailed description of the data being studied (including the exact time span or the versions of the software);	Details of the original study, including details about the date being studies, are necessary	Understanding the original study (details of original study available)
		To indicate the specific version of the research tool used;	Precise and unambiguous definitions of study details (terminology, tools, instruments, etc.) are necessary	Precise and unambiguous design and execution of the studies (original and replication)
		To reach an agreement on a standardized way to refer to a location where additional data can be obtained	Precise and unambiguous definitions of study details (terminology, tools, instruments, etc.) are necessary	Precise and unambiguous design and execution of the studies (original and replication)

Table 17
Synthesis of recommendations.

Synthesis		
Understanding the original study is a central issue in performing a replication. The recommendations regarding this understanding fall in three categories: communications with the researchers that performed the original studies; the development of pilot studies by the researchers performing the replication; and the availability of precise and detailed information about the original studies in publications or research packages	**Variations** between original study and replication are almost inevitable due to intended or unintended reasons. These variations must be fully assessed and their possible effects on the results of the replication and the comparison of these results and those of the original study must be evaluated. Intended variations are related to the need or desire of updating or extending research instruments or procedures, or even extending the research goals and questions. As these variations are under direct control of the researcher performing the replications, they effect may be easier to assess and control Unintended variations are related to the new context in which the replication will occur and pose more challenges to researchers because they are more difficult to be found and controlled	**Precise and unambiguous design and execution of the studies (original and replication)** is a major issue in the successful development of any empirical study. Empirical studies must all be carefully designed and executed to increase the validity and, therefore, the usefulness of their results. In the case of replications, this becomes even more critical because imprecisions of design or execution of original or replication can lead to unintended (and often not explicitly addressed) variations among studies that could ultimately make the comparison of results between studies very difficult or even impossible

[ABO009] T. Mende, Replication of defect prediction studies: problems, pitfalls and recommendations, in: Proceedings of the 6th International Conference on Predictive Models in Software Engineering, ACM, 2010, p. 5.

[ABO010] O. S. Gómez, N. Juristo, S. Vegas, Replications types in experimental disciplines, in: Proceedings of the 2010 ACM-IEEE international symposium on empirical software engineering and measurement, ACM, 2010, p. 3. doi: 10.1145/1852786.1852790.

[ABO011] N. Juristo, S. Vegas, The role of non-exact replications in software engineering experiments, Empirical Software Engineering 16 (3) (2011) 295–324.

[ABO012] F. J. Shull, J. C. Carver, S. Vegas, N. Juristo, The role of replications in empirical software engineering, Empirical Software Engineering 13 (2) (2008) 211–218. doi: 10.1007/s10664-008-9060-1.

[ABO013] B. Kitchenham, The role of replications in empirical software engineering – a word of warning, Empirical software engineering 13 (2) (2008) 219–221. doi: 10.1007/s10664-008-9061-0.

[ABO014] N. Juristo, S. Vegas, Using differences among replications of software engineering experiments to gain knowledge, in: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, 2009, pp. 356–366. doi: 10.1109/ESEM.2009.5314236.

[ABO015] F. Shull, V. Basili, J. Carver, J. C. Maldonado, G. H. Travassos, M. Mendonça, S. Fabbri, Replicating Software Engineering Experiments: Addressing the Tacit Knowledge Problem. In: Proceedings of the 2002 International Symposium on Empirical Software Engineering, p. 7–16. Nara, Japan, October 3–4 2002.

[ABO016] F. Shull, J. Carver, G. H. Travassos, J. C. Maldonado, R. Conradi, V. R. Basili, Replicated studies: building a body of knowledge about software reading techniques, Series On Software Engineering And Knowledge Engineering, 12 (2003) 39–84.

[ABO017] J. Miller, Replicating software engineering experiments: a poisoned chalice or the holy grail, Information and Software Technology 47 (4) (2005) 233–244.

[ABO018] J. C. Carver, Towards reporting guidelines for experimental replications: A proposal, in: International Workshop on Replication in Empirical Software Engineering Research, Cape Town, South Africa, 2010.

[ABO019] O. Dieste, E. Fernandez, R. García, N. Juristo, Hidden evidence behind useless replications, in: Proceedings of the International Workshop on Replication in Empirical Software Engineering Research, 2010.

[ABO020] R. Ferrari, O. Sudmann, C. Henke, J. Geisler, W. Schafer, N. H. Madhavji, Transitioning from lab studies to large-scale studies: Emerging results from a literal replication, in: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research, 2010.

[ABO021] O. Gómez, N. Juristo, S. Vegas, Replication, reproduction and re-analysis: Three ways for verifying experimental Findings, in: Proceedings of the 1st international workshop on replication in empirical software engineering research (RESER 2010), Cape Town, South Africa, 2010.

[ABO022] J. L. Krein, C. D. Knutson, A Case for Replication: Synthesizing Research Methodologies in Software Engineering. In: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research (RESER) May 4, 2010. Cape Town, South Africa.

[ABO023] M. V. Mäntylä, C. Lassenius, J. Vanhanen, Rethinking replication in software engineering: Can we see the forest for the trees, in: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research, 2010.

[ABO024] A. Mockus, B. Anda, D. I. Sjøberg, Experiences from replicating a case study to investigate reproducibility of software development, in: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research, 2010

[ABO025] G. Robles, D. M. Germán, Beyond replication: an example of the potential benefits of replicability in the mining of software repositories community, in: Proceedings of the 1st international workshop on replication in empirical software engineering Research (RESER 2010), 2010.

[ABO026] D. I. K. Sjøberg, T. Dybå, M. Jørgensen. The Future of Empirical Methods in Software Engineering Research. Future of Software Engineering (FOSE'07).

[ABO027] N. Juristo, O. S. Gómez, Replication of software engineering experiments, in: Empirical software engineering and verification, Springer, 2012, pp. 60–88. doi: 10.1007/978-3-642-25231-0_2.

[ABO028] L. P. Scatalon, R. E. Garcia, R. C. M. Correia, Packaging controlled experiments using an evolutionary approach based on ontology (s), in: SEKE, 2011, pp. 408–413.

[ABO029] E. G. E. Gallardo, Using configuration management and product line software paradigms to support the experimentation process in software engineering, in: Research Challenges in Information Science (RCIS), 2012 Sixth International Conference on, IEEE, 2012, pp. 1–6.

[ABO030] C. V. de Magalhães, F. Q. da Silva, Towards a taxonomy of replications in empirical software engineering research: A research proposal, in: 3rd International Workshop on Replication in Empirical Software Engineering Research (RESER), IEEE, 2013, pp. 50–55. doi: 10.1109/RESER.2013.10

[ABO031] G. Ghezzi, H. C. Gall, Replicating mining studies with sofas, in: Proceedings of the 10th Working Conference on Mining Software Repositories, IEEE Press, 2013, pp. 363–372.

[ABO032] D. Callele, K. Wnuk, M. Borg, Confounding factors when conducting industrial replications in requirements engineering, in: 1st International Workshop on Conducting Empirical Studies in Industry (CESI), IEEE, 2013, pp. 55–58.

[ABO033] M. Squire, A replicable infrastructure for empirical studies of email archives, in: 3rd International Workshop on Replication in Empirical Software Engineering Research (RESER), 2013, IEEE, 2013, pp. 43–49. doi: 10.1109/RESER.2013.11.

[ABO034] N. Juristo, S. Vegas, M. Solari, S. Abrahão, I. Ramos, A process for managing interaction between experimenters to get useful similar replications, Information and Software Technology 55 (2) (2013) 215–225.

[ABO035] P. Runeson, A. Ste k, A. Andrews, Variation factors in the design and analysis of replicated controlled experiments, Empirical Software Engineering (2013) 1–28. doi: 10.1007/s10664-013-9262-z.

[ABO036] A. Brooks, J. Daly, J. Miller, M. Roper, M. Wood, Replication of experimental results in software engineering, International Software Engineering Research Network (ISERN) Technical Report ISERN-96-10, University of Strathclyde. 1996.

[ABO037] J. P. F. Almqvist, Replication of controlled experiments in empirical software engineering-a survey, Department of Computer Science, Faculty of Science, Lund University. 2006.

Appendix B. Reference list of replications

[REP003] S. Abrahão, G. Poels, A family of experiments to evaluate a functional size measurement procedure for web applications, Journal of Systems and Software 82 (2) (2009) 253–269.

[REP007] L. Reynoso, E. Manso, M. Genero, M. Piattini, Assessing the influence of import-coupling on OCL expression maintainability: A cognitive theory-based perspective, Information Sciences 180 (20) (2010) 3837–3862. doi: 10.1016/j.ins.2010.06.028

[REP009] A. C. Dias-Neto, G. H. Travassos, Evaluation of {model-based} testing techniques selection approaches: An external replication, in: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society, 2009, pp. 269–278.

[REP012] F. Ricca, G. Scanniello, M. Torchiano, G. Reggio, E. Astesiano, On the effectiveness of screen mockups in requirements engineering: results from an internal replication, in: Proceedings of the 2010 ACMIEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2010, p. 17.

[REP016] J. A. Cruz-Lemus, A. Maes, M. Genero, G. Poels, M. Piattini, The impact of structural complexity on the understandability of UML statechart diagrams, Information Sciences 180 (11) (2010) 2209–2220. doi: 10.1016/j.ins.2010.01.026

[REP019] G. Du, J. McElroy, G. Ruhe, A family of empirical studies to compare informal and optimization based planning of software releases, in: Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering, ACM, 2006, pp. 212–221.

[REP020] C. Andersson, A replicated empirical study of a selection method for software reliability growth models, Empirical Software Engineering 12 (2) (2007) 161–182.

[REP024] P. Ardimento, M. T. Baldassarre, D. Caivano, G. Visaggio, Assessing multiview framework (MF) comprehensibility and efficiency: A replicated experiment, Information and Software Technology 48 (5) (2006) 313–322.

[REP026] M. Staron, L. Kuzniarz, C. Wohlin, Empirical assessment of using stereotypes to improve comprehension of UML models: A set of experiments, Journal of Systems and Software 79 (5) (2006) 727–742.

[REP027] G. Canfora, A. Cimitile, F. Garcia, M. Piattini, C. A. Visaggio, Evaluating performances of pair designing in industry, Journal of Systems and Software 80 (8) (2007) 1317–1327.

[REP033] L. C. Briand, C. Bunse, J. W. Daly, A controlled experiment for evaluating quality guidelines on the maintainability of object-oriented designs, Software Engineering, IEEE Transactions on 27 (6) (2001) 513–530.

[REP035] G. Canfora, F. García, M. Piattini, F. Ruiz, C. A. Visaggio, A family of experiments to validate metrics for software process models, Journal of Systems and Software 77 (2) (2005) 113–129.

[REP036] E. Mendes, N. Mosley, S. Counsell, A replicated assessment of the use of adaptation rules to improve web cost estimation, in: Proceedings of International Symposium on Empirical Software Engineering. IEEE, 2003, pp. 100–109.

[REP038] T. Thelin, C. Andersson, P. Runeson, N. Dzamashvili-Fogelstrom, A replicated experiment of usage based and checklist-based reading, in: Software Metrics, 2004. Proceedings of 10th International Symposium on, IEEE, 2004, pp. 246–256.

[REP039] M. Shepperd, M. Cartwright, A replication of the use of regression towards the mean (r2m) as an adjustment to effort estimation models, in: 11th IEEE International Symposium Software Metrics, 2005. IEEE, 2005, pp. 10-pp.

[REP041] F. Lanubile, T. Mallardo, F. Calefato, C. Denger, M. Ciolkowski, Assessing the impact of active guidance for defect detection: a replicated experiment, in: Software Metrics, 2004. Proceedings. 10th International Symposium on, IEEE, 2004, pp. 269–278.

[REP048] M. Roper, M. Wood, J. Miller, An empirical evaluation of defect detection techniques, Information and Software Technology 39 (11) (1997) 763–775.

[REP051] O. Laitenberger, K. El Emam, T. G. Harbich, An internally replicated quasi-experimental comparison of checklist and perspective based reading of code documents, Software Engineering, IEEE Transactions on 27 (5) (2001) 387–421.

[REP098] J. C. Maldonado, J. Carver, F. Shull, S. Fabbri, E. Dória, L. Martimiano, M. Mendonça, V. Basili, Perspective-based reading: a replicated experiment focused on individual reviewer effectiveness, Empirical Software Engineering 11 (1) (2006) 119–142.

[REP103] M. M. Müller, Two controlled experiments concerning the comparison of pair programming to peer review, Journal of Systems and Software 78 (2) (2005) 166–179.

[REP104] F. Calefato, D. Gendarmi, F. Lanubile, Investigating the use of tags in collaborative development environments: a replicated study, in: Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement, ACM, 2010, p. 24.

[REP118] J. Lung, J. Aranda J, S. Easterbrook, G; Wilson, On the difficulty of replicating human subjects studies in software engineering. International Conference on Software Engineering, 2008, pp. 191–200.

[REP120] A. C. C. França, P. R. da Cunha, F. Q. da Silva, The effect of reasoning strategies on success in early earning of programming: lessons learned from an external experiment replication, in: Proceedings of the 14th international conference on Evaluation and Assessment in Software Engineering, British Computer Society, 2010, pp. 81–90.

[REP021] C. Andersson, P. Runeson, Replicated Quantitative Analysis of Fault Distributions in Complex Software Systems. IEEE Transactions on Software Engineering, 33(5), 2007, pp. 273–286.

[REP122] M. Genero, M. Piattini, L. Jiménez, Empirical validation of class diagram complexity metrics, in: Proceedings. XXI International Conference of the Chilean Computer Science Society, 2001. SCCC'01. IEEE, 2001, pp. 95–104.

[REP123] M. Genero, L. Jiménez, M. Piattini, A controlled experiment for validating class diagram structural complexity metrics, in: Object-Oriented Information Systems, Springer, 2002, pp. 372383.

[REP124] M. Genero, M. Piattini, E. Manso, G. Cantone, Building UML class diagram maintainability prediction models based on early metrics, in: Software Metrics Symposium, 2003. Proceedings. Ninth International, IEEE, 2003, pp. 263–275.

[REP125] M. Genero, M. Piatini, E. Manso, Finding "early" indicators of uml class diagrams understandability and modifiability, in: Empirical Software Engineering, 2004. ISESE'04. Proceedings. 2004 International Symposium on, IEEE, 2004, pp. 207–216.

[REP126] M. Genero, E. Manso, A. Visaggio, G. Canfora, M. Piattini, Building measure-based prediction models for UML class diagram maintainability, Empirical Software Engineering 12 (5) (2007) 517–549. doi: 10.1007/s10664-007-9038-4.

[REP129] J. A. Cruz-Lemus, M. Genero, M. E. Manso, M. Piattini, Evaluating the effect of composite states on the understandability of UML statechart diagrams, in: Model Driven Engineering Languages and Systems, Springer, 2005, pp. 113–125.

[REP130] J. Cruz-Lemus, M. Genero, M. Piattini, S. Morasca, Improving the experimentation for evaluating the effect of composite states on the understandability of UML statechart diagrams, in: Proceedings of 5th ACM-IEEE International Symposium on Empirical Software Engineering, Rio de Janeiro, Brazil, 2006, pp. 9–11.

[REP131] J. A. Cruz-Lemus, M. Genero, M. E. Manso, S. Morasca, M. Piattini, Assessing the understandability of UML statechart diagrams with composite statesa family of empirical studies, Empirical Software Engineering 14 (6) (2009) 685–719. doi: 10.1007/s10664-009-9106-z.

[REP002FE] J. A. Cruz-Lemus, M. Genero, D. Caivano, S. Abrahão, E. Insfrán, J. A. Carsí, Assessing the influence of stereotypes on the comprehension of UML sequence diagrams: A family of experiments, Information and Software Technology 53 (12) (2011) 1391–1403.

[REP003FE] E. I. Laukkanen, M. V. Mäntylä, Survey reproduction of defect reporting in industrial software development, in: 2011 International Symposium on Empirical Software Engineering and Measurement (ESEM), IEEE, 2011, pp. 197–206. doi: 10.1109/ESEM.2011.28.

[REP004FE] R. Premraj, K. Herzig, Network versus code metrics to predict defects: A replication study, in: Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on, IEEE, 2011, pp. 215–224. doi: 10.1109/ESEM.2011.30

[REP006FE] N. Juristo, S. Vegas, Design patterns in software maintenance: An experiment replication at UPM – Experiences with the RESER'11 joint replication project, in: Replication in Empirical Software Engineering Research (RESER), 2011 Second International Workshop on, IEEE, 2011, pp. 7–14. doi: 10.1109/RESER.2011.8.

[REP013FE] C. Gravino, M. Risi, G Scanniello, G. Tortora, Does the Documentation of Design Pattern Instances Impact on Source Code Comprehension? Results from Two Controlled Experiments. 2011, pp. 449–462. doi: 10.1109/TSMCA.2010.2087017.

[REP015FE] D. Bowes, T. Hall, A. Kerr, Program slicing-based cohesion measurement: the challenges of replicating studies using metrics, in: Proceedings of the 2nd International Workshop on Emerging Trends in Software Metrics, ACM, 2011, pp. 75–80. doi: 10.1145/1985374.1985392

[REP019FE] G. Reggio, F. Ricca, G. Scanniello, F. Di Cerbo, G. Dodero, A precise style for business process modeling: Results from two controlled experiments, in: Model Driven Engineering Languages and Systems, Springer, 2011, pp. 138–152.

[REP022FE] S. Di Martino, F. Ferrucci, C. Gravino, F. Sarro, Using web objects for development effort estimation of web applications: a replicated study, in: Product-Focused Software Process Improvement, Springer, 2011, pp. 186–201.

[REP023FE] V. de Castro, M. Genero, E. Marcos, M. Piattni, Empirical study to assess whether the use of routes facilitates the navigability of web information systems, Software, IET 5 (6) (2011) 528–542. doi: 10.1049/iet-sen.2010.0062.

[REP024FE] B. Biegel, Q. D. Soetens, W. Hornig, S. Diehl, S. Demeyer, Comparison of similarity metrics for refactoring detection, in: Proceedings of the 8th Working Conference on Mining Software Repositories, ACM, 2011, pp. 53–62. doi: 10.1145/1985441.1985452.

[REP025FE] O. A. L. Lemos, F. C. Ferrari, F. F. Silveira, A. Garcia, Development of auxiliary functions: should you be agile? An empirical assessment of pair programming and test-first programming, in: Proceedings of the 2012 International Conference on Software Engineering, IEEE Press, 2012, pp. 529–539.

[REP026FE] G. R. Bergersen, D. I. Sjøberg, Evaluating methods and technologies in software engineering with respect to developers' skill level, 2012, p.101–110.

[REP027FE] K. Wnuk, M. Höst, B. Regnell, Replication of an experiment on linguistic tool support for consolidation of requirements from multiple sources, Empirical Software Engineering 17 (3) (2012) 305–344.

[REP028FE] D. S. Kusumo, M. Staples, L. Zhu, H. Zhang, R. Jeffery, Risks of off-the-shelf-based software acquisition and development: a systematic mapping study and a survey.

[REP030FE] T. G. Grbac, P. Runeson, D. Huljenić, A Second Replicated Quantitative Analysis of Fault Distributions in Complex Software Systems, IEEE Transactions on Software Engineering, V. 39, No.4, 2012, pp. 462–476

[REP033FE] C. Gravino, M. Risi, G. Scanniello, G. Tortora, Do professional developers benefit from design pattern documentation? a replication in the context of source code comprehension, Springer, 2012. doi: 10.1007/978-3-642-33666-9_13.

[REP035FE] N. Salleh, E. Mendes, J. Grundy, Investigating the effects of personality traits on pair programming in a higher education setting through a family of experiments, Empirical Software Engineering 19 (3) (2014) 714–752.

[REP036FE] Ö. Albayrak, J. C. Carver, Investigation of individual factors impacting the effectiveness of requirements inspections: a replicated experiment, Empirical Software Engineering 19 (1) (2014) 241–266. doi: 10.1007/s10664-012-9221-0

[REP038FE] F. Q. da Silva, A. C. C. França, M. Suassuna, L. M. de Sousa Mariz, I. Rossiley, R. C. de Miranda, T. B. Gouveia, C. V. Monteiro, E. Lucena, E. S. Cardozo, et al., Team building criteria in software projects: A mix-method replicated study, Information and Software Technology 55 (7) (2013) 1316–1340.

Appendix C. Summary of recommendations

See Tables 16 and 17.

References

[1] C.V. de Magalhães, F.Q.B. da Silva, R.E. Santos, Investigations about replication of empirical studies in software engineering: preliminary findings from a mapping study, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, ACM, 2014, p. 37, http://dx.doi.org/10.1145/2601248.2601289.

[2] S. Schmidt, Shall we really do it again? The powerful concept of replication is neglected in the social sciences, Rev. Gen. Psychol. 13 (2) (2009) 90, http://dx.doi.org/10.1037/a0015108.

[3] R.M. Lindsay, A.S. Ehrenberg, The design of replicated studies, Am. Stat. 47 (3) (1993) 217–228.

[4] J.C. Carver, N. Juristo, M.T. Baldassarre, S. Vegas, Replications of software engineering experiments, Empirical Softw. Eng. 19 (2) (2014) 267–276, http://dx.doi.org/10.1007/s10664-013-9290-8.

[5] F.Q. da Silva, M. Suassuna, A.C.C. França, A.M. Grubb, T.B. Gouveia, C.V. Monteiro, I.E. dos Santos, Replication of empirical studies in software engineering research: a systematic mapping study, Empirical Softw. Eng. (2012) 501–557, http://dx.doi.org/10.1007/s10664-012-9227-7.

[6] R. Bezerra, F.Q. da Silva, Replication of Empirical Studies in Software Engineering: A Systematic Mapping Study Extension (Official title in Portugeuse: Replicação de Estudos Empíricos em Engenharia de Software: Extensão de um Mapeamento Sistemático). Master Thesis, Centre for Informatics, Federal University of Pernambuco, 2014.
[7] M.A. La Sorte, Replication as a verification technique in survey research: a paradigm, Sociol. Quart. 13 (2) (1972) 218–227.
[8] J. Gould, W.L. Kolb, Dictionary of the Social Sciences, Tavistock, London, 1964.
[9] J. Daly, A. Brooks, J. Miller, M. Roper, M. Wood, Veri cation of results in software maintenance through external replication, in: International Conference on Software Maintenance, Proceedings, IEEE, 1994, pp. 50–57.
[10] C.D. Knutson, J.L. Krein, L. Prechelt, N. Juristo, Report from the 1st international workshop on replication in empirical software engineering research (RESER 2010), ACM SIGSOFT Softw. Eng. Notes 35 (5) (2010) 42–44.
[11] J.L. Krein, C.D. Knutson, L. Prechelt, N. Juristo, Report from the 2nd international workshop on replication in empirical software engineering research (RESER 2011), ACM SIGSOFT Softw. Eng. Notes 37 (1) (2012) 27–30, http://dx.doi.org/10.1145/2088883.2088889.
[12] J.L. Krein, C.D. Knutson, C. Bird, Report from the 3rd international workshop on replication in empirical software engineering research (RESER 2013), ACM SIGSOFT Softw. Eng. Notes 39 (1) (2014) 31–35, http://dx.doi.org/10.1145/2557833.2557858.
[13] M. Petticrew, H. Roberts, Systematic Reviews in the Social Sciences: A Practical Guide, John Wiley & Sons, 2008.
[14] B. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical Report, EBSE Technical Report EBSE-2007-01, 2007.
[15] R.K. Yin, Case Study Research: Design and Methods, Sage Publications, 2002.
[16] C.M. Judd, E.R. Smith, L.H. Kidder, Research Methods in Social Relations, Rinehart and Winston Inc., Orlando, FL, 1991.
[17] N. Cartwright, Replicability, reproducibility, and robustness: comments on harry collins, Hist. Polit. Econ. 23 (1) (1991) 143–155.
[18] J. Lung, J. Aranda, S. Easterbrook, G. Wilson, On the difficulty of replicating human subjects studies in software engineering, in: Proceedings of the 30th International Conference on Software Engineering, ICSE '08, 2008, pp. 191–200.
[19] G.W. Noblit, R.D. Hare, Meta-Ethnography: Synthesizing Qualitative Studies (Qualitative Research Methods), [S.l.]: Sage Publications Inc., 1988.
[20] M.T. Baldassarre, J. Carver, O. Dieste, N. Juristo, Replication types: towards a shared taxonomy, in: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, ACM, 2014, p. 18.
[21] C. Hendrick, Replications, strict replications, and conceptual replications: are they important?, J Soc. Behav. Pers. (1991) 41–49.
[22] V.R. Basili, R.W. Selby, D.H. Hutchens, Experimentation in software engineering, IEEE Trans. Softw. Eng. (7) (1986) 733–743.
[23] A. Jedlitschka, D. Pfahl, Reporting guidelines for controlled experiments in software engineering, in: 2005 International Symposium on Empirical Software Engineering, IEEE, 2005, p. 10.
[24] O.S. Gómez, N. Juristo, S. Vegas, Understanding replication of experiments in software engineering: a classification, Inf. Softw. Technol. 56 (8) (2014) 1033–1048.
[25] D.I. Sjøberg, T. Dyba, M. Jorgensen, The future of empirical methods in software engineering research, in: Future of Software Engineering, FOSE'07, IEEE, 2007, pp. 358–378.