IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 28, 2019, accepted November 2, 2019, date of publication November 7, 2019,
date of current version February 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2952191

Replication of Studies in Empirical Software Engineering: A Systematic Mapping Study, From 2013 to 2018

MARGARITA CRUZ, BEATRIZ BERNÁRDEZ, AMADOR DURÁN, JOSÉ A. GALINDO, AND ANTONIO RUIZ-CORTÉS
Department of Computer Languages and Systems, Universidad de Sevilla, 41004 Seville, Spain

Corresponding author: Margarita Cruz (cruz@us.es)

This work was supported in part by the European Commission (FEDER) and the Spanish Government under projects BELI under Grant TIN2015-70560-R, in part by OPHELIA under Grant RTI2018-101204-B-C22, and in part by the Juan de la Cierva Postdoctoral Program.

ABSTRACT ***Context***: In any discipline, replications of empirical studies are necessary to consolidate the acquired knowledge. In Software Engineering, replications have been reported since the 1990s, although their number is still small. The difficulty in publishing, the lack of guidelines, and the unavailability of replication packages are pointed out by the community as some of the main causes. ***Objective***: Understanding the current state of replications in Software Engineering studies by evaluating current trends and evolution during the last 6 years. ***Method***: A Systematic Mapping Study including articles published in the 2013–2018 period that report at least one replication of an empirical study in Software Engineering. ***Results***: 137 studies were selected and analysed, identifying: *i*) forums; *ii*) authors, co-authorships and institutions; *iii*) most cited studies; *iv*) research topics addressed; *v*) empirical methods used; *vi*) temporal distribution of publications; and *vii*) distribution of studies according to research topics and empirical methods. ***Conclusions***: According to our results, the most relevant forums are the *Empirical Software Engineering* and *Information and Software Technology* journals, and the *Empirical Software Engineering and Measurement* conference. We observed that, as in previous reviews by other researchers, most of the studies were carried out by European institutions, especially Italian, Spanish, and German researchers and institutions. The studies attracting more citations were published mainly in journals and in the *International Conference on Software Engineering*. Testing, requirements, and software construction were the most frequent topics of replication studies, whereas the usual empirical method was the controlled experiment. On the other hand, we identified research gaps in areas such as software engineering process, software configuration management, and software engineering economics. When analysed together with previous reviews, there is a clear increasing trend in the number of published replications in the 2013–2018 period.

INDEX TERMS Empirical software engineering, replications, systematic mapping study.

I. INTRODUCTION

Replication of empirical studies in Software Engineering (SE) is an essential activity for achieving greater validity and reliability in research results [2], [13]. Most definitions consider replications to be repetitions of research procedures already performed in a so–called *original* or *baseline* studies [6]. As many other processes in SE, replications can be classified according to several criteria. Depending on whether or not they are carried out by the same experimenters as in the original study, they can be classified as *internal* or *external* [14]. According to the degree to which the original experiment procedure is followed, they can be also classified as *exact* and *conceptual* [15] or *closed* and *differentiated* [16]. Basili *et al*. refer to *strict* replications when the original study is duplicated as accurately as possible [17], whereas Gómez *et al*. classify them as *literal*, *operational* and *conceptual* depending on the changes carried out and their purpose [18].

Regardless of their classification, the importance of replications is twofold: *i*) to confirm or extend results; and *ii*) to know the influence of new variables, some of them due to changes imposed by the environment [19]. Although the first article reporting a replication in SE was published

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Anwar Hossain.

TABLE 1. Comparison of related reviews in chronological order.

Review work	Review type	Review context	Review period	Primary studies[a]	Review focus
[1]	SLR	Experiments	1993–2002	103	Current state of experiments
[2]	SLR	Replications	1993–2002	41	Current state of replications
[3]	SLR	Replications	2010	15	Reporting guidelines
[4]	SMS	Replications	1994–2010	96	Current state of replications
[5]	SLR	Experiments	2013–2014	43	Deviating data points
[6]	SMS	Replications	1996–2013	37	Current state of replications
[7]	SMS	Replications	2011–2012	39	Current state of replications
[8]	SMS	Experiments	1999–2012	13	Context characterization
[9]	SMS	Experiments	2003–2013	110	Human subjects
[10]	SLR	Experiments	2011–2013	405	Internal and external validity
[11]	SMS	Replications	1998–2016	39	Families of experiments
[12]	SLR	Replications	1999–2017	28	Replications results
This article	SMS	Replications	2013–2018	137	Current state of replications

[a]Number of primary studies analysed in the review

in 1994 [20], the number of reported replications since then remains low, mainly because of the difficulty in publishing, lack of guidelines and the unavailability of replication packages [10]. Although some guidelines have been proposed for reporting controlled experiments in SE [21]–[23], to the best of our knowledge, Carver's guidelines [3] are the only high–level proposal for reporting replications, although they are very general and their application is not always straightforward.

On the other hand, the main techniques to inquire into literature are Systematic Mapping Study (SMS) and Systematic Literature Review (SLR). SMSs take *primary sources*, i.e. papers under study, that are heterogeneous in terms of comparability but are related to a broader area, and provide a mapping and categorization of the different facets detected in the studies [24]. Researchers read the title, abstract, and optionally, another part of the paper stepwise. The idea is to get the whole picture of a broad research area and identify research gaps using visual representations. On the contrary, SLRs take primary sources that are homogeneous in terms of comparability and compare them to get to conclusions [25]. SLRs provide a synthesis of the knowledge existing in a specific field. To do so, researchers have to review in–depth, understand and classify the whole content of the studies. In [24], these techniques are compared pointing out that they should be used in a complementary way. An SMS can be conducted first to get the whole picture of a broad research area and then, a specific area of interest is investigated with an SLR.

In our case, after having to report two internal replications [26], [135] of a baseline experiment [27] and finding some problems applying Carver's guidelines [3]—especially about reporting changes between replications—we decided to carry out an SMS on replications in SE as a first step before performing an SLR. In order to know the trends and emerging topics, and also to identify research niches that facilitate and foster replications, we have reviewed studies that report at least one replication in SE in recent years. As a result, we have identified the main trends, leading venues, institutions, authors and related literature. Our goal is to highlight the potential research opportunities regarding both research topics and empirical methods used by experimenters. Since there were reviews on replications covering previous years [4], [6], [7], only studies published since 2013 were considered.

The remainder of this article is organized as follows: Section II discusses previous related reviews; Section III covers the methodology used in the performed SMS; Section IV presents results after analysing primary studies; Section V analyses the threats to validity; and Section VI presents the concluding remarks and future work.

II. RELATED REVIEWS

This section presents previous reviews related to replications in SE. Table 1 summarizes chronologically these studies showing their type (SMS or SLR), context, covered period, number of primary studies reviewed and focus.

A. REVIEWS ABOUT CURRENT STATE OF REPLICATIONS

Sjøberg *et al*. performed in 2005 a manual search of controlled experiments in SE published between 1993 and 2002 [1]. Although not focused on replications, they found that 18% of the surveyed studies were replications. They also identified a lack of terminological consistency and the need of guidelines for the conduction and reporting of experiments facilitating their replication and meta–analysis. For the same interval of years, Almqvist performed in 2006 a manual search in SE journals and conference proceedings focusing on replications [2]. He grouped the studies into 20 series of experiments, classified replications as internal and external and analysed whether replications confirmed the results of the original studies. His main recommendations were to improve experimental reporting, to develop

replication guidelines, and to have laboratory packages available to promote replicability.

In 2014–2015, three SMSs analysed the current state of replications in SE [4], [6], [7]. Da Silva *et al.* [4] and Cartaxo *et al.* [7] selected articles published until 2012 reporting replications and extracted information such as the percentage and evolution of internal and external replications and the main topics addressed. Magalhães *et al.* [6] explored articles published until 2013 that did not report any replication but addressed different topics about replications such as conceptual frameworks, guidelines, processes and recommendations about how to perform and report replications. These studies agree on the lack of taxonomies, guidelines, and standardization on how to report empirical studies and replications.

B. REVIEWS REGARDING SPECIFIC ASPECTS OF REPLICATIONS

As commented in the introductory section, Carver performed in 2010 an SLR focused on replications in which an initial proposal of reporting guidelines for publishing experimental replications was presented [3]. In 2014, Larsson *et al.* performed another SLR about data accessibility for replications and the presence of deviating data points, also known as outliers. Their review concluded that only 37% of the reviewed studies had available data and that there was a need to improve the replication of data analysis [5].

In 2015, Cartaxo *et al.* carried out an SMS to identify the mechanisms to support context characterization and therefore to facilitate the replication of empirical studies in SE. Their research concluded that there were few studies supporting context characterization [8]. Also in 2015, Falcão *et al.* performed an SMS to analyse the experiments published in the *Empirical Software Engineering Journal* (EMSE), the *International Conference on Empirical Software Engineering and Measurement* (ESEM), and the *International Conference on Empirical Assessment & Evaluation in Software Engineering* (EASE). Their analysis included, among other features, information on whether the experiments were replications, the type of replications, researchers and institutions publishing experiments and categories of subjects involved in the experiments [9]. In the same year, Siegmund *et al.* carried out an SLR about the status of empirical research in SE, searching at three main venues: the *International Conference on Software Engineering* (ICSE), the *European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (ESEC/FSE), and the ESEM conference [10]. The main contribution of their review was how to address the trade–off between internal and external validity. Their analysis included, among other features, whether the experiments were replications, the empirical methods used and the types of participants.

In 2018, Santos *et al.* published an SMS identifying the techniques used to analyse *families of experiments*, considering them as a group of replications in which: *i*) there is access to raw data; *ii*) the changes introduced are known; and *iii*) the effects of at least two technologies are evaluated in three experiments on the same response variable [11]. In addition, the changes introduced in the analysed replications were classified according to the dimensions proposed in [18]. Shepperd *et al.* conducted also in 2018 an SLR in the areas of software project effort prediction and peer programming. Their analysis included the level of confirmation between replication and original studies according to the type of replication. They concluded that internal replications find confirmatory evidence eight times more than external replications. They also stressed the need to improve reporting on the original, replication and expected results [12].

III. SYSTEMATIC MAPPING STUDY PROCESS

To perform our SMS on replications in SE, we followed the process proposed in [24], similarly to a recent SMS in the area of Software Product Lines [28]. This procedure proposes three main phases in an SMS, namely:

1) Planning the review, which includes the definition of both process protocol and research questions.
2) Identification of studies, which covers the selection of primary sources.
3) Data extraction and classification, where the mapping is developed and conclusions are obtained.

For a better understanding of the followed process, we have modelled it using BPMN [29], as shown in Fig. 1.

FIGURE 1. High–Level process model of the systematic mapping study.

A. PHASE 1: PLANNING THE REVIEW

In this first phase, the directives for carrying out the review and the research questions are stated. As shown in Fig. 2, we distinguish two main tasks, namely *protocol definition* and *definition of research questions*.

1) PROTOCOL DEFINITION

To minimize the threats to validity, the search strategy, the inclusion criteria (IC_i) and the exclusion criteria (EC_i) must be specified. The search strategy starts by querying the SCOPUS and Web of Science (WoS) repositories with a custom search string taking into account the inclusion criteria. Then, the results are filtered applying the exclusion criteria and eliminating duplicated studies. The inclusion criteria were the following:

- IC_1: Studies reporting at least one replication of an empirical study in SE.
- IC_2: Studies published between 2013 and 2018.

Our search starts in 2013 because, as commented in Section II, there are two SMSs on replications in SE including studies performed until 2012 [4], [7]. On the other hand, the exclusion criteria were the following:

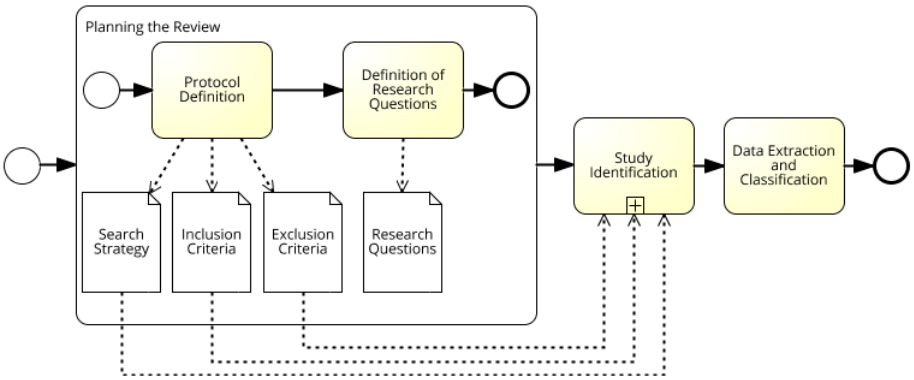


FIGURE 2. Process model of the *planning the review* phase.

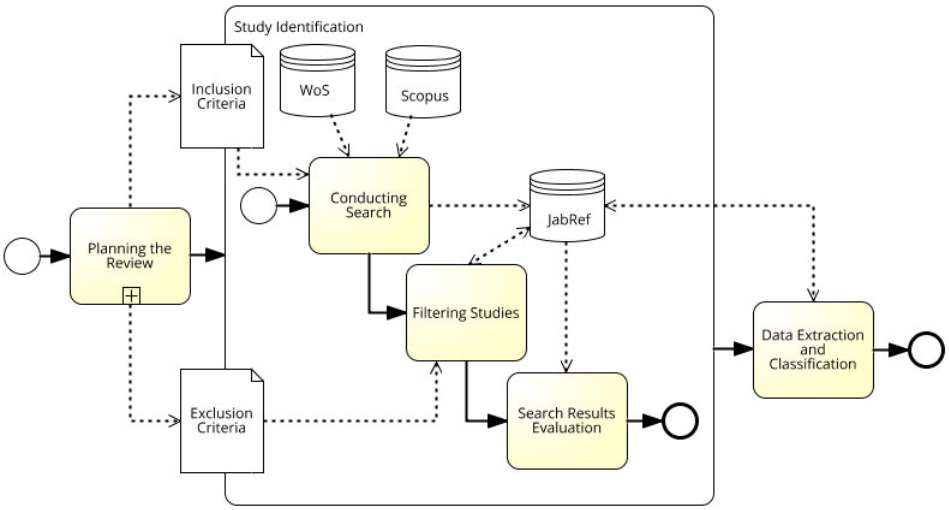


FIGURE 3. Process Model of the *identification of studies* phase.

- EC_1: The term *replication* is used referring to another context rather than experiment replication (e.g. data replication in databases or networks).
- EC_2: The term *replication* is used only as future work.
- EC_3: The document is not a research paper but an extended abstract, a tutorial, a summary of conferences, etc.
- EC_4: The study is not written in English.

2) DEFINITION OF RESEARCH QUESTIONS

In order to understand the current state of replications in SE, the research questions for our SMS were the following:

- RQ_1: Which forums are used to publish replications?
- RQ_2: Who are the authors and institutions publishing replications?
- RQ_3: Which are the most cited studies reporting replications?
- RQ_4: Which research topics have been most/less replicated?
- RQ_5: What empirical methods have been most used in replications?
- RQ_6: How is the number of studies reporting replications evolving?
- RQ_7: How are the studies distributed according to research topics and empirical methods?

B. PHASE 2: STUDY IDENTIFICATION

In this second phase, the studies to be included in the SMS are identified following the process model shown in Fig. 3.

1) CONDUCT SEARCH FOR PRIMARY SOURCES

The automatic search was conducted querying the SCOPUS repository containing primary studies, i.e. scientific journals, books, and conference proceedings. The query string was developed considering the inclusion criteria and usual synonyms such as *experiment*, *empirical study* or *controlled experiment*. After several iterations analysing retrieved studies and adjusting query terms, the final query string was the following:

```
"software engineering" and
title-abs-key( "experiment*" or "case stud*"
or "observational stud*" or "pilot stud*"
```

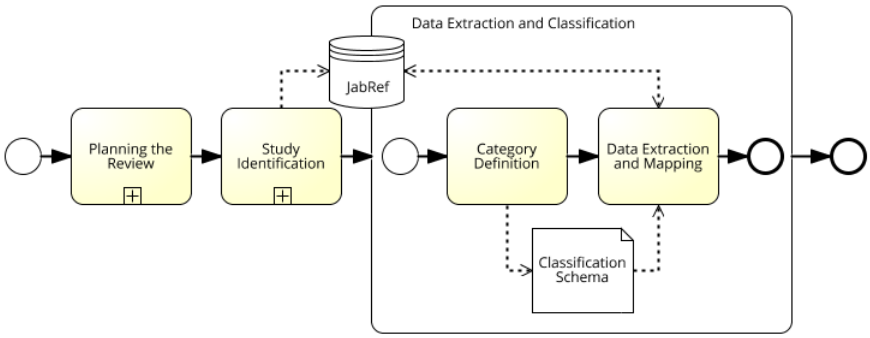


FIGURE 4. Process model of *data Extraction and classification* phase.

```
or "survey" ) and
title-abs-key("repli*" or "family of*") and
pubyear > 2012 and pubyear < 2019
```

where `"software engineering"` restricts the search to studies within that area by searching in the title, abstract, keywords, and references, whereas `title-abs-key` searches only in the title, abstract and keywords of studies, but not in their references.[1] The query was executed on May 2019 and returned 909 studies which were stored in a local JabRef[2] file. For the sake of completeness, a similar search was also performed in the WoS repository with the following query string:

```
ALL="software engineering" and
TS=("experiment*" or "case stud*" or
"observational stud*" or "pilot stud*" or
"survey") and
TS=("repli*" or "family of*") and
PY > 2012 and PY < 2019
```

which was also executed on May 2019 but returned only 152 studies. This remarkable difference between the number of retrieved studies from each repository is mainly due to the WoS field label `ALL`, which does not search in references, as is the case for SCOPUS. Nevertheless, we decided not to change the SCOPUS query string because we were aware that some relevant studies were discarded in the case the `"software engineering"` string was not looked up in the references. Therefore, after removing repeated studies, the final number of studies was 921.

2) FILTERING STUDIES

The exclusion criteria were applied mainly by one researcher to ensure that all the studies were reviewed uniformly. After reading the title and abstract of each study, other sections such as the introduction or conclusions were also reviewed before making the decision of excluding a study or not. Only in case of doubt, a second researcher was consulted. After applying this manual process, 149 studies were initially included. When we detected that very similar versions of the same study had been published in more than one journal or conference, only the most recent journal publication was included, discarding the others. As a result, 12 duplicate studies were excluded. The final list of 137 primary sources is provided in the first appendix.

3) EVALUATION OF SEARCH RESULTS

The search accuracy was evaluated in two ways. Firstly, by modifying the final query strings and checking that some studies of interest were not included. As a double check, we also verified that the number of studies of the best known authors in the area was lower when using the modified search strings. Secondly, the year interval was changed in order to compare the results with those from [7], obtaining about 80% of their results.

C. PHASE 3: DATA EXTRACTION AND CLASSIFICATION

In this last phase, the mapping itself is performed after defining classification categories, as shown in Fig. 4.

1) DEFINITION OF CATEGORIES

In our mapping, publications were classified according to two facets, the addressed *research topic area* and the *type of empirical study* carried out. With respect to the former facet, instead of keyword–oriented approaches such as [24], we decided to use some of the 15 well–known SWEBOK [30] *knowledge areas*[3] (KAs), namely *requirements* (REQS), *design* (DESG), *construction* (CONS), *testing* (TEST), *maintenance* (MAIN), *configuration management* (CONF), *management* (MNGT), *process* (PROC), *models and methods* (METH), *quality* (QUAL), *professional practice* (PROF), and *economics* (ECON). Note that some of the studies were classified in more than one research topic area.

For the latter facet, primary sources were classified according to the experimental method used, i.e. *experiment* (EX), *quasi–experiment* (QE), *case study* (CS) and *survey* (SV). Note that *quasi–experiments* are considered by some authors

[1] Some relevant studies presented the string `"software engineering"` only in their references, this is why `"software engineering"` was not restricted to `title-abs-key`.

[2] JabRef is an open source bibliography reference manager using BibTeX as its native file format. See www.jabref.org for details.

[3] We have used all SWEBOK KAs except the three *foundations* areas, i.e. computing, mathematical, and engineering foundations, which are too general for our purpose.

TABLE 2. Journals in which replication studies have been published in 2013–2018 ordered by number of publications.

Published studies	Journal name	Acronym	SCImago	JCR	Q_{2018}[a]
21	Empirical Software Engineering	EMSE	✓	✓	Q1
14	Information and Software Technology	IST	✓	✓	Q1
6	IEEE Transactions on Software Engineering	TSE	✓	✓	Q1
5	ACM Transactions on Software Engineering and Methodology	TOSEM	✓	✓	Q1
5	Journal of Systems and Software	JSS	✓	✓	Q1
3	Journal of Visual Languages and Computing	JVLC	✓	✓	Q3
3	Software Quality Journal	SQJ	✓	✓	Q2
3	Software and Systems Modeling	SoSyM	✓	✓	Q1
2	Advanced Materials Research	AMR	✓	✗	–
2	Science of Computer Programming	SCP	✓	✓	Q3
2	Journal of Software — Evolution and Process	JSEP	✓	✓	Q3
1	Applied Soft Computing	ASC	✓	✓	Q1
1	Future Generation Computer Systems	FGCS	✓	✓	Q1
1	Requirements Engineering	RE	✓	✓	Q2
1	Computer Journal	CJ	✓	✓	Q4
1	Advances in Intelligent Systems and Computing	AISC	✗	✗	–
1	Journal of Theoretical and Applied Information Technology	JTAIT	✓	✗	–
1	International Journal of Software Engineering and Knowledge Engineering	IJSEKE	✓	✓	Q4
1	e–Informatica Software Engineering Journal	EISEJ	✓	✗	–
1	Expert Systems with Applications	ESA	✓	✓	Q1

[a] 2018 quartile in JCR index.

such as [23] as *controlled experiments* where the assignment of treatments to subjects is not random, whereas other authors consider them as *experiments* in which researchers do not control every factor, e.g. when they cannot obtain a satisfactory sample [1], [31]. In our SMS, only primary studies classified as *quasi–experiments* by their authors have been considered as such.

2) DATA EXTRACTION AND MAPPING

At the end of the process, the mapping was performed by applying the classification scheme to the 137 primary sources following the steps below.

1) The classification was carried out separately by two researchers according to the two aforementioned facets. After reading the title, abstract and keywords—and other sections in case of doubt—each researcher classified each primary source.
2) The two classifications were compared and, in case of disagreement, both researchers examined the document again until consensus was reached.

IV. SYSTEMATIC MAPPING STUDY RESULTS

In this section, we address each of the research questions defined in Section III-A2 using graphs to visualize collected data and providing subsequent interpretations.

A. WHICH FORUMS ARE USED TO PUBLISH REPLICATIONS?

Of the 137 primary studies identified, 75 (55%) were published in journals and 62 (45%) in conference proceedings. Table 2 presents journals according to the number of published replication studies, including also whether they are indexed in journal quality rankings such as JCR or SCImago. The two leading journals, EMSE and IST, include almost half of the articles published in journals, i.e. they represent 25% of all the studies included in our review. It is worth noting the presence of the *Advanced Materials Research* (AMR) journal whose scope is very different to SE. We verified manually that the two articles published in this journal [108], [109] were correctly selected, i.e. that they complied with the inclusion and exclusion criteria.

For each journal, Table 3 presents the number of studies classified according to the two facets of our mapping, i.e. empirical method used and research topic area. As can be seen, the *controlled experiment* (EX) is the preferred empirical method and there is a great variety of research topics, although *requirements* (REQS) and *testing* (TEST) stand out from the rest. Note that neither *configuration management* (CONF) nor *economics* (ECON) topics, for which no replication studies were found, are included in Table 3.

Table 4 shows conferences in which at least one replication study was presented in 2013–2018, with the ESEM conference standing out from the rest with more than 15% of all the studies included in our review. Despite being a workshop on replications, only 3 replication studies in the same period were presented at the *International Workshop on Replication in Empirical Software Engineering Research at* ICSE (RESER). This discontinued workshop was held in 2010, 2011 and 2013, so due to the range of years covered in our review, only the 2013 edition was included. For the sake of brevity, conferences where only one replication study has been presented are listed in Table 12 in Appendix B.

Similarly to Table 3, Table 5 depicts for each conference with more than one replication study, the number of studies

TABLE 3. Mapping of journals in which replication studies have been published in 2013–2018 across facets.

	Empirical method				Research topic area									
Journal	SV	QE	EX	CS	REQS	DESG	CONS	TEST	MAIN	MNGT	PROC	METH	QUAL	PROF
EMSE	1	2	17	1	4	1	10	5	6	0	0	1	4	2
IST	2	1	9	2	6	2	1	2	1	0	1	2	1	3
TSE	0	0	5	1	3	2	0	0	1	0	0	1	2	0
TOSEM	0	0	5	0	2	1	2	2	1	0	0	0	0	0
JSS	0	0	5	0	2	2	0	2	1	0	0	1	0	2
JVLC	0	0	3	0	1	2	1	0	0	1	0	0	0	0
SQJ	1	0	1	1	0	1	0	3	0	0	0	0	1	0
SoSyM	0	0	3	0	1	2	0	0	1	0	0	2	0	0
AMR	0	0	1	1	0	1	0	1	0	1	0	0	0	0
SCP	0	0	2	0	1	0	0	1	0	0	0	1	0	0
JSEP	0	0	2	0	0	0	1	1	1	0	0	0	1	0
ASCJ	0	0	0	1	0	0	1	1	0	0	0	0	1	0
FGCS	0	1	0	0	0	0	0	0	0	1	0	0	1	0
RE	0	0	0	1	1	0	0	0	0	0	0	0	0	0
CJ	0	0	1	0	0	0	0	1	0	0	0	0	0	0
AISC	0	0	1	0	1	1	0	0	0	0	0	0	0	0
JTAIT	0	0	1	0	1	0	0	0	0	0	0	0	0	0
IJSEKE	0	0	1	0	0	0	0	0	0	0	0	1	0	0
EISEJ	0	0	1	0	0	0	0	1	0	0	0	0	0	0
ESA	0	0	0	1	0	0	0	1	0	0	0	0	1	0
Total	4	4	58	9	23	15	16	21	12	3	1	9	12	7

TABLE 4. Conferences in which more than one replication study has been published in 2013–2018 ordered by number of publications.

Published studies	Conference name	Acronym
9	International Symposium on Empirical Software Engineering and Measurement	ESEM
6	International Conference on Software Engineering	ICSE
4	International Conference on Evaluation and Assessment in Software Engineering	EASE
4	International Conference on Software Engineering and Knowledge Engineering	SEKE
3	International Workshop on Replication in Empirical Software Engineering Research	RESER
3	Requirements Engineering Conference	RE
3	Ibero–American Conference on Software Engineering	CIBSE
2	Americas Conference on Information Systems	AMCIS
2	ACM Symposium on Applied Computing	SAC
2	IEEE International Conference on Program Comprehension	ICPC

with respect to the categories of the two facets of our mapping. In the case of conferences, *testing* (TEST) and *software construction* (CONS) are the most frequently addressed topics, although *software quality* (QUAL) and *requirements* (REQS) have close numbers. As for journals, no replication studies were found for the *configuration management* (CONF) and *economics* (ECON) topics, which were therefore not included in Table 3. With respect to the most widely used empirical methods, the *controlled experiment* (EX), as in journals, clearly stands out from the rest.

B. WHO ARE THE AUTHORS AND INSTITUTIONS PUBLISHING REPLICATIONS?

Authors with at least four replications published in the review period are mapped across SMS facets in Table 6, where it can be seen that Giuseppe Scanniello from the University of Basilicata (Italy) and Natalia Juristo and Oscar Dieste from the Technical University of Madrid (Spain) are the most prolific authors. It can also be seen that the most used empirical method is the *controlled experiment* (EX) and the most addressed research topic is *requirements* (REQS). Other well–known authors such as Claes Wohlin from the Blekinge Institute of Technology (Sweden), have relevant publications on empirical SE in the review period, e.g. [32] with 88 citations in SCOPUS, but we have not included them in Table 6 because, although they contribute significantly to the field, they have not reported any replication explicitly in the 2013–2018 period.

To analyse co–authoring networks, we have used the VOSviewer tool [33]. Fig. 5 shows a co–authorship map generated with VOSviewer in which the authors in Table 6, i.e. authors with at least 4 publications, are depicted as clustered

TABLE 5. Mapping of conferences in which more than one replication study has been published in 2013–2018 across facets.

	Empirical method				Research topic area									
Conference	SV	QE	EX	CS	REQS	DESG	CONS	TEST	MAIN	MNGT	PROC	METH	QUAL	PROF
ESEM	2	0	6	1	0	2	2	3	0	3	0	2	4	0
ICSE	0	0	5	1	0	1	3	2	1	0	0	0	1	1
EASE	0	0	4	0	2	1	0	2	0	0	0	0	1	0
SEKE	2	0	2	0	1	0	1	1	0	0	0	2	1	0
RESER	0	0	2	1	1	0	1	1	0	0	0	1	0	0
RE	1	0	1	1	3	0	0	0	0	0	0	0	0	0
CIBSE	0	0	3	0	0	0	1	1	0	2	0	0	0	1
AMCIS	1	0	1	0	0	0	1	0	0	0	1	0	1	0
SAC	0	0	1	1	0	0	1	1	0	0	0	0	0	0
ICPC	0	0	2	0	0	1	0	0	0	0	0	1	0	0
Total	6	0	27	5	7	5	10	11	1	5	1	6	8	2

TABLE 6. Mapping of authors with at least four replication studies published in 2013–2018 across facets.

	Empirical method				Research topic areas									
Author	SV	QE	EX	CS	REQS	DESG	CONS	TEST	MAIN	MNGT	PROC	METH	QUAL	PROF
Scanniello, G.	0	0	15	0	7	7	5	3	2	1	0	2	1	0
Juristo, N.	0	1	7	0	2	1	0	2	1	0	0	3	3	2
Dieste, O.	1	0	6	0	3	0	1	3	0	0	0	2	2	2
Fernández, D.M.	5	0	0	1	6	0	0	0	0	0	0	0	0	0
Ricca, F.	0	0	6	0	2	2	1	0	1	0	0	2	0	0
Abrahão, S.	0	0	6	0	2	3	0	1	0	0	1	3	0	0
Gravino, C.	0	0	6	0	4	4	2	0	2	0	0	0	0	0
Tortora, G.	0	0	6	0	4	4	2	0	2	0	0	0	0	0
Genero, M.	0	0	5	0	5	1	1	0	3	0	0	0	0	0
Spínola, R.O.	5	0	0	0	3	0	1	0	0	1	0	1	0	0
Insfran, E.	0	0	5	0	2	3	0	1	0	0	1	2	0	0
Torchiano, M.	1	0	4	0	1	2	1	0	1	0	0	2	0	0
Carver, J.C.	1	0	4	0	2	1	2	0	2	0	0	0	1	0
Conte, T.	3	0	2	0	3	1	0	1	0	0	1	0	1	0
Vegas, S.	0	0	4	0	0	1	0	0	1	0	0	3	1	1
Risi, M.	0	0	4	0	1	2	2	2	1	0	0	0	0	0
Fucci, D.	0	0	4	0	0	0	1	3	0	0	0	0	3	1
Wagner, S.	4	0	0	0	4	0	0	0	0	0	0	0	0	0
Prikladnicki, R.	3	0	1	0	3	0	0	0	0	0	1	0	1	0
Reggio, G.	0	0	4	0	1	2	0	0	1	0	0	2	0	0
Total	23	1	87	1	55	34	19	16	17	2	4	22	13	6

linked bubbles. The size of the bubbles depends on the number of publications whereas the width of the links depends on the number of common publications between authors. Groups of authors with common publications are displayed clustered. Coherently with the data in Table 6, three main clusters are depicted in Fig. 5 around the most prolific authors, i.e. Scanniello, Juristo, and Daniel Méndez Fernández, who is with the Technical University of Munich (Germany). Other two smaller clusters are also identified around Silvia Abrahão, from the Technical University of Valencia (Spain), and Filippo Ricca, from the University of Genoa (Italy). When all authors are considered, i.e. not only those with at least four publications, the resulting co–authorship map is shown in Fig. 6.

Table 7 shows those institutions which currently house the most active authors and therefore lead the advances in replications in empirical SE. To obtain this data, the institutions of the authors of each primary study were counted. In the case several authors in the same study belonged to the same institution, the institution was counted only once.

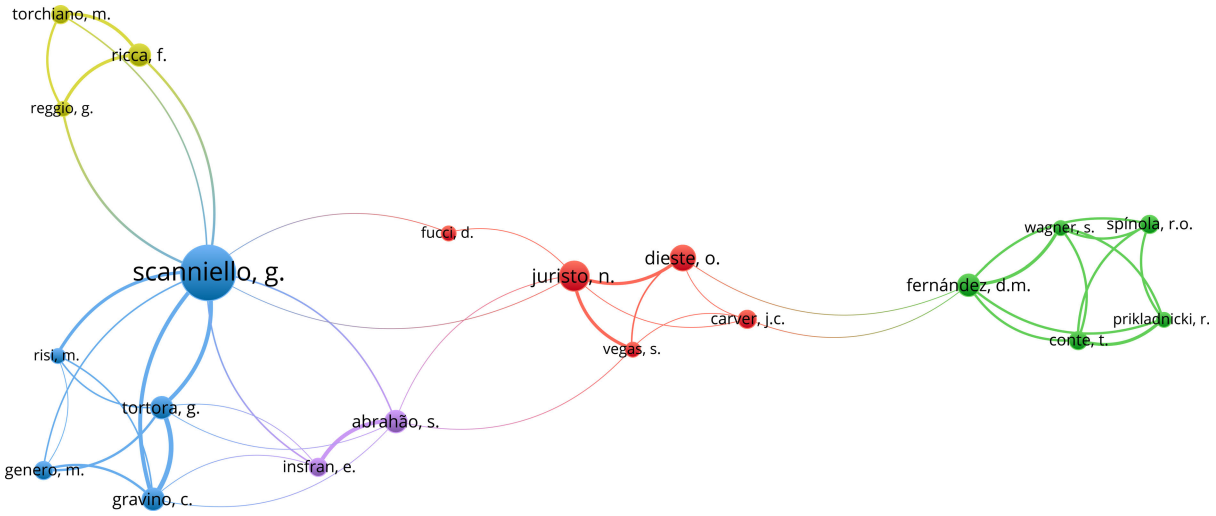


FIGURE 5. Co–authorship map of authors with at least four replication studies in 2013–18.

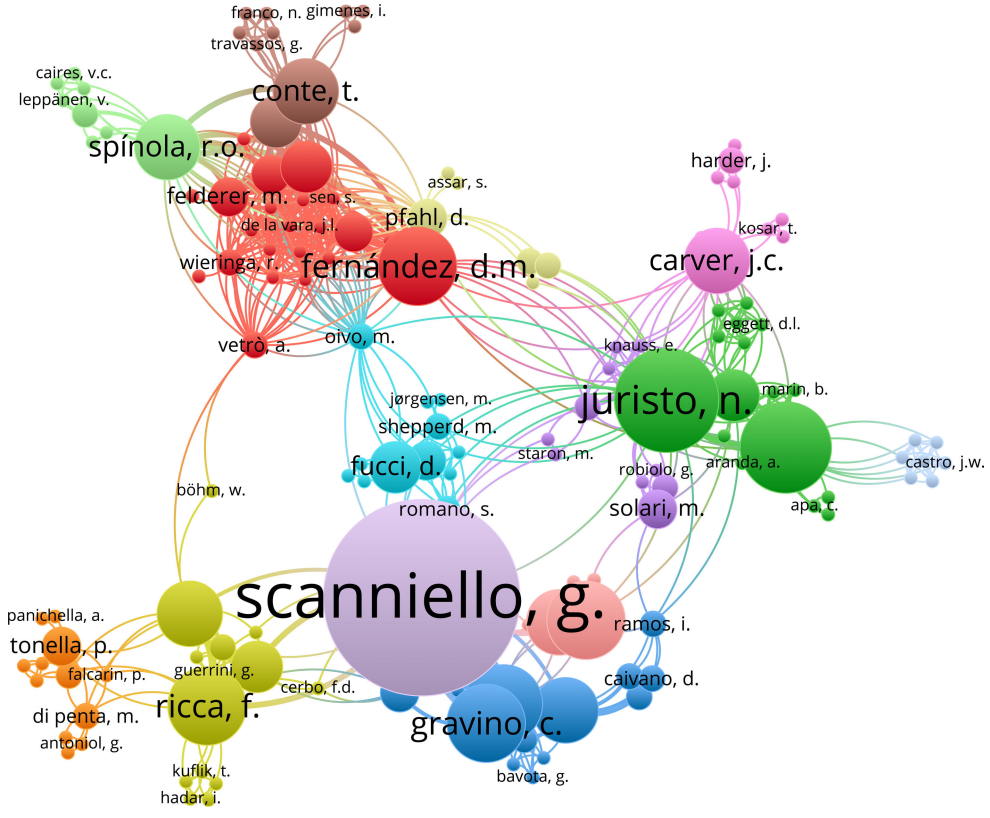


FIGURE 6. Co–authorship map of authors with at least one replication study in 2013–18.

European universities clearly lead the ranking, with Italy and Spain as the most active countries in the field. This fact is a clear continuation of the results in [4], where the most active institutions were the University of Castilla–La Mancha (Spain), the University of Sannio (Italy), the Simula Research Laboratory (Norway), and the University of Valladolid (Spain). In our survey, the five leading institutions are Italian (University of Basilicata and University of Salerno) and Spanish universities (University of Castilla–La Mancha, Technical University of Madrid, and Technical University of Valencia).

C. WHICH ARE THE MOST CITED STUDIES REPORTING REPLICATIONS?

Table 8 lists the most cited studies, including the forum where they were published, their year of publication, their number of citations according to SCOPUS as of May 2019, as well as the empirical method applied and the corresponding research

TABLE 7. Institutions with at least four replication studies published in 2013–2018 ordered by the number of publications.

Published studies	Institution	Official website	Country
15	University of Basilicata	http://portale.unibas.it	Italy
8	University of Salerno	https://www.unisa.it/	Italy
7	University of Castilla-La Mancha	https://www.uclm.es/	Spain
6	Technical University of Madrid	http://www.fi.upm.es	Spain
6	Technical University of Valencia	http://www.upv.es/	Spain
5	Technical University of Munich	https://www.tum.de/	Germany
5	Universita di Genova	https://unige.it/	Italy
5	Universidade Federal do Amazonas	https://ufam.edu.br/	Brazil
5	University of Saõ Paulo	https://icmc.usp.br/	Brazil
4	University of Oulu	https://www.oulu.fi/	Finland
4	University of Lund	https://lunduniversity.lu.se/	Sweden
4	Fondazione Bruno Kessler	https://www.fbk.eu/en/	Italy
4	Politecnico di Torino	https://www.polito.it/	Italy

TABLE 8. Most cited studies reporting at least one replication in 2013–2018 ordered by the number of cites according to SCOPUS as of May 2019.

Study	Forum	Year	#Cites	#Cites/Year	Emp. Method	Topic Areas
Gothra *et al.* [66]	ICSE	2015	109	27,3	CS	QUAL
Pearson *et al.* [131]	ICSE	2017	49	24,5	EX	TEST
Binkley *et al.* [161]	EMSE	2013	44	7,3	EX	TEST
Ceccato *et al.* [162]	EMSE	2014	41	8,2	EX	CONS
Abrahão *et al.* [159]	TSE	2013	40	6,7	EX	REQS, DESG
Fernández D.M. *et al.* [58]	IST	2015	34	8,5	SV	REQS
Fernández A. *et al.* [163]	JSS	2013	34	5,7	EX	TEST, METH
Fernández D.M. *et al.* [116]	EMSE	2017	28	14	SV	REQS
Scanniello *et al.* [102]	TOSEM	2014	27	5,4	EX	REQS, CONS
Ribeiro *et al.* [92]	ICSE	2014	24	4,8	EX	MAIN
Hadar *et al.* [71]	IST	2013	24	4	EX	REQS

topic areas. Note that the subjects of the two most cited studies [65], [130] were not humans but automated methods for defect prediction and fault location. Although our main interest is in replications of experiments with human subjects, we did not exclude these studies following similar criteria than [4]. Note also that the two most cited articles by D. M. Fernández are about a family of surveys on the state of the practice of Requirements Engineering in Germany. Regarding forums where the most cited contributions have been published, the only conference is ICSE, with 3 studies, and the rest are journals, highlighting EMSE (3 studies) and IST (2 studies).

D. WHICH RESEARCH TOPICS HAVE BEEN MOST/LESS REPLICATED?

Table 9 shows primary studies grouped by related research topic areas. Notice that the number of studies in Table 9 (215) is greater than the number of primary studies (137) because some studies were classified in more than one topic area. Coherently with the data presented in previous sections and in previous reviews [4], [7], the research topic areas that have attracted most attention are *testing* (TEST), *requirements* (REQS) and *software construction* (CONS) with more than 25% of primary studies addressing each of them.

Some research gaps have been identified related with the *economics* (ECON), *configuration management* (CONF), and *process* (PROC) research topic areas, i.e. only 3 replications studies were published in the PROC area and none in ECON and CONF areas in the SMS period. Although there are some empirical studies in the SMS period related to the CONF area, such as [34], and the PROC area, such as [35], we have not been able to find any related to the ECON area. Sometimes, the empirical studies are recent, such as [36], so it seems reasonable that their replications were not published yet. Nevertheless, other not so recent studies such as [37], have not been replicated, which is a very common situation in most SE experiments, as described in [18].

E. WHAT EMPIRICAL METHODS HAVE BEEN MOST USED IN REPLICATIONS?

Table 10 shows the number of primary studies grouped by the empirical method used. As it can be seen, the *controlled experiment* (EX) is used in almost 75% of the studies. It is worth noting the low number of *quasi–experiments* (QE),

TABLE 9. Research topic areas and related primary studies published in 2013–2018 ordered by the number of publications.

Topic Area	Primary studies	#Studies
TEST	[42], [44], [46], [48], [52], [53], [62], [63], [64], [68], [74], [82], [84], [93], [95], [102], [103], [106], [108], [113], [114], [116], [121], [130], [132], [133], [134], [142], [144], [146], [151], [155], [156], [159], [162], [166], [172], [173]	38
REQS	[40], [43], [45], [51], [57], [58], [59], [70], [72], [78], [79], [83], [85], [86], [87], [92], [98], [99], [100], [101], [111], [115], [118], [120], [123], [125], [126], [135], [153], [158], [159], [163], [165], [167], [168]	35
CONS	[46], [49], [66], [67], [69], [71], [72], [75], [84], [94], [96], [101], [105], [106], [119], [121], [122], [127], [128], [129], [131], [136], [137], [139], [140], [141], [145], [146], [149], [152], [155], [156], [160], [161], [164]	35
DESG	[41], [47], [53], [67], [69], [75], [80], [81], [82], [87], [93], [97], [98], [107], [109], [111], [113], [124], [135], [158], [163], [167], [169], [171], [172], [174], [175]	27
QUAL	[39], [40], [42], [44], [50], [52], [62], [63], [64], [65], [68], [75], [76], [84], [90], [103], [111], [114], [117], [138], [143], [147], [155], [170]	24
METH	[41], [54], [55], [56], [60], [76], [77], [107], [110], [136], [143], [148], [162], [163], [165], [169], [174]	17
MAIN	[45], [49], [58], [59], [71], [80], [91], [96], [97], [133], [140], [141], [148], [149]	14
MNGT	[47], [72], [73], [81], [88], [89], [90], [100], [109], [147], [157]	11
PROF	[39], [61], [66], [94], [104], [110], [112], [135], [137], [146], [150]	11
PROC	[129], [154], [170]	3
	Total	215

TABLE 10. Empirical methods and number of primary studies published in 2013–2018 using them ordered by the number of publications.

Method	Primary studies	#Studies
EX	[40], [41], [42], [43], [45], [46], [47], [48], [49], [52], [53], [54], [55], [56], [58], [59], [60], [62], [63], [64], [67], [69], [70], [71], [73], [74], [75], [76], [77], [80], [82], [87], [88], [89], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100], [101], [103], [106], [109], [110], [111], [114], [116], [119], [120], [121], [123], [124], [126], [127], [128], [130], [131], [132], [133], [134], [135], [137], [138], [139], [140], [141], [142], [143], [144], [145], [146], [148], [150], [151], [153], [154], [155], [156], [158], [159], [160], [161], [162], [163], [164], [165], [166], [167], [168], [169], [170], [171], [172], [173], [174], [175]	101 (73,72%)
CS	[44], [50], [61], [65], [68], [72], [81], [83], [84], [85], [86], [90], [102], [108], [112], [117], [118], [129], [152]	19 (13,87%)
SV	[57], [78], [79], [104], [105], [107], [113], [115], [122], [125], [136], [157]	12 (8,76%)
QE	[39], [51], [66], [147], [149]	5 (3,65%)
	Total	137 (100%)

probably due to the fact that some researches present them as controlled experiments in the abstract and even in the introduction of theirs studies, but as a quasi–experiment in inner sections. As a matter of fact, some authors consider quasi–experiments as a specific type of controlled experiment, as commented in Section III-C1 [1], [31].

F. HOW IS THE NUMBER OF STUDIES REPORTING REPLICATIONS EVOLVING?

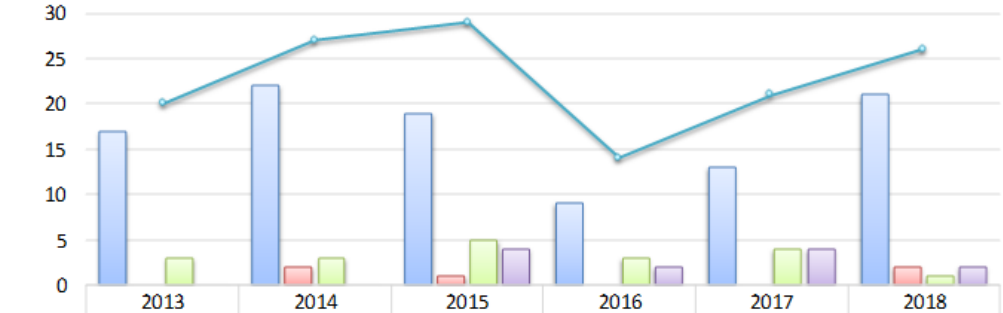
Fig. 7 shows the number of replications published in the review period. With an average of 22.8 studies per year, strong variations are observed due to the small number of publications. With regard to empirical methods, the *controlled experiment* (EX)—with an average of 16.8 per year—present similar numbers each year except for 2016. The number of *quasi–experiments* (QE) is very low, as commented in the previous section. On the other hand, the number of *case studies* (CS) each year is very similar and *surveys* (SV) present a growing trend since 2015.

G. HOW ARE THE STUDIES DISTRIBUTED ACCORDING TO RESEARCH TOPICS AND EMPIRICAL METHODS?

Fig. 8 shows the number of studies across the two facets of our mapping. Consistently with the results commented in previous sections, *testing* (TEST), *software construction* (CONS), *requirements* (REQS), and *design* (DESG) are the research topics with more replications using *controlled experiments* (EX). On the other hand, *quality* (QUAL)—with 7 case studies in 24 contributions—, REQS and TEST are the topic areas where *case study* (CS) has been applied more frequently as empirical method.

V. THREATS TO VALIDITY

According to [4], the most common limitations in a systematic review are limited coverage, possible biases introduced in



	2013	2014	2015	2016	2017	2018
EX	17	22	19	9	13	21
QE	0	2	1	0	0	2
CS	3	3	5	3	4	1
SV	0	0	4	2	4	2
ALL	20	27	29	14	21	26

FIGURE 7. Number of replication studies published in 2013–2018 grouped by the empirical method used.

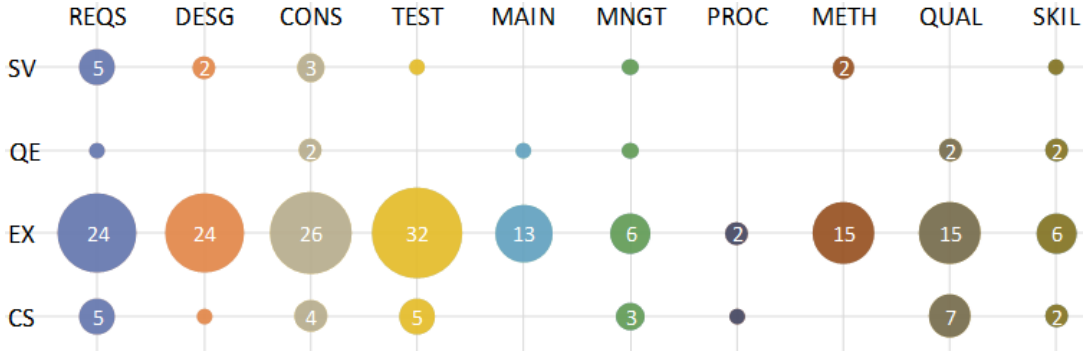


FIGURE 8. Mapping of replication studies published in 2013-2018 across facets.

the selection process, and inaccuracies during data extraction. With respect to coverage, the initial search was performed in the SCOPUS repository verifying that the main SE journals and conferences, especially those related to empirical SE, were indexed in SCOPUS. To extend coverage, the WoS repository was included in the search to check whether new results were obtained. The second action was ensuring that the queries returned as many valid results as possible by diversifying search terms using synonyms. The search strings were obtained after several iterations and results were carefully checked, as described in Section III-B.

With respect to the selection process and data extraction, we relied on automated mechanism when possible instead of manual methods to prevent any bias or errors. The only two manual processes were the filtering of primary studies and their classification into defined categories. The filtering was carried out by one researcher only to ensure that all studies were reviewed with the same criteria. The classification process, due to its difficulty, was carried out by two researchers and studies were examined until consensus was reached.

VI. CONCLUSION AND FUTURE WORK

In this article, we have presented an SMS on replications of empirical studies in SE from 2013 to 2018. We have identified forums used to publish contributions as well as authors and institutions holding the current know–how. We have also identified the studies with the greatest impact and the temporal distribution of studies. To identify current trends as well as research gaps in the field, we have performed a mapping to classify studies according to two facets, namely the addressed research topic areas and the used empirical methods. When possible, Table 11 summarises and compares our results with those of previous reviews [4], [7].

Although previous reviews did not addressed the identification of the forums used to publish replications in SE, in our review we have found that in 2013–2018, the number of studies published in journals (75) is higher than in the conferences (62) and that three forums—the EMSE and IST journals and the ESEM conference—concentrate more than 30% of published replications.

Regarding authors and institutions, M. Piattini, M. Genero (both with U. of Castilla–La Mancha) and E. Manso (U. Valladolid) were identified in [4] as the three authors publishing more replications, and the universities of Castilla–La Mancha (Spain), Sannio (Italy), Valladolid (Spain) and the Simula Research Laboratory (Norway) as the leading institutions. In our review, the most prolific authors are G. Scanniello (U. of Basilicata), N. Juristo and O. Dieste

TABLE 11. Comparison of results with Da Silva *et al.* [4] and Bezerra *et al.* [7].

RQ	Context	Da Silva *et al.* [4]	Bezerra *et al.* [7]	This mapping
RQ_1	Forums	Not applicable	Not applicable	The EMSE and IST journals and the ESEM conference are the most used forums for publishing replications in SE.
RQ_2	Authors and Institutions	M. Piattini, M. Genero and E. Manso. U. of Castilla–La Mancha, U. of Sannio, Simula Research Lab., and U. of Valladolid.	Not applicable	G. Scanniello, N. Juristo and O. Dieste. U. of Basilicata and U. of Salerno (Italy), and U. of Castilla-La Mancha, Tech. U. of Madrid and Tech. U. of Valencia (Spain).
RQ_3	Most cited studies	Not applicable	Not applicable	Two papers presented at ICSE about automated methods for defect prediction and fault location. Among the most cited, two surveys on requirements by D. M. Fernández.
RQ_4	Research Topics	REQS, CONS, QUAL	DESG, TEST, METH	Our results are similar to previous reviews, with TEST, REQS, and CONS as the main research topics studied in replications. We have also identified research gaps in PROC, CONF, and ECON.
RQ_5	Empirical methods	QE	QE	In the first two reviews, the most commonly used empirical method was QE whereas replications in our mapping use mostly EX.
RQ_6	Studies per year	Clear increasing trend	Clear increasing trend	In the period covered, the number of replication studies suffers strong variations. When considered together with previous reviews, there is a clear increasing trend in the number of published replications.
RQ_7	Methods vs Topics	Not applicable	Not applicable	In our review, the most common combination in replication studies is EX and TEST.

(T. U. of Madrid), D. M. Fernández (T. U. of Munich), S. Abrahão (T. U. of Valencia), F. Ricca (U. of Genoa) and C. Gravino and G. Tortora (U. of Salerno) and the leading institutions are the universities of Basilicata, Salerno, Castilla–La Mancha, T. U. of Madrid and T. U. of Valencia. Using the co–authorship maps (see Fig. 5 and 6), we have also identified clusters of cooperating authors such as the cluster headed by G. Scanniello and other Italian researchers, and the cluster of authors with the T. U. of Madrid (N. Juristo, O. Dieste and S. Vegas), which includes also cooperation with other researches such as J. C. Carver, D. M. Fernández, and S. Abrahão among others. Although the leading authors and institutions change over time, it seems that most replications in SE have been carried out by Italian and Spanish researchers and institutions, without underestimating the relevant contributions from Brazil and Germany.

With respect to the most cited studies, the first two ones are papers presented at ICSE on automated techniques for defect prediction models [65] and fault localization [130]. Remarkably, only three of the most prolific authors—G. Scanniello, D. M. Fernández and S. Abrahõo—are authors of four of the studies in the top 10 most cited, whereas none of the institutions housing the authors of [65], [130] are in the list of leading institutions.

In previous reviews, the most addressed research topics were requirements, construction, quality, design, testing, and methods. In our review, the interest of researchers have focused mainly on testing, requirements, and software construction, all of them already present in [4], [7]. We have also identified a clear research gap in replications in software engineering process, configuration management, and software engineering economics. About empirical methods, there is a clear trend from quasi–experiments in previous reviews to controlled experiments in our SMS, where the combination of testing and controlled experiment is the most frequently found.

Regarding the evolution of the number of studies reporting replications, previous studies distinguished three periods, 1994–2003, 2004–2009 and 2010–2012, with an average of 4.1, 11.7 and 24.3 replications per year. In 2013–2018, 137 studies were found in 6 years, i.e. an average of 22,8 per year. In addition, some studies include more than one replication, so the average number could be greater. This represents a clear increasing trend of the number of published replications in SE in the period 2013–2018 when previous reviews are considered.

As future work, we will study the impact of the only specific proposal, i.e. Carver's guidelines [3], on the reporting of replications. In particular, we will focus on how authors report changes between original experiments and replications in order to elaborate a proposal on how to specify such changes. In the longer term, our intention is to integrate the proposal with the experimental information repository eXemplar [38].

LABORATORY PACKAGE

The laboratory package is available at https://exemplar.us.es/demo/CruzSMS2019 and includes the following items: *i*) the search strings for the SCOPUS and WoS repositories, so the queries can be reproduced; *ii*) a RIS file containing the 137 primary studies; *iii*) instructions to generate the

co–authorship maps using VOSviewer and the aforementioned RIS file.

APPENDIX A
PRIMARY STUDIES

The 137 primary studies considered in the SMS presented in [40]–[176] listed below.

APPENDIX B
OTHER CONFERENCES

Table 12 lists conferences in which only one replication study has been presented in the 2013–2018 period.

TABLE 12. Conferences in which only one replication study has been published in 2013–2018.

Conference name	Acronym
International Conference on Product-Focused Software Process Improvement	PROFES
IEEE International Workshop on Machine Learning Techniques for Software Quality Evaluation	MaLTeSQuE
ACM Research in Adaptive and Convergent Systems	RACS
Brazilian Symposium on Software Components, Architectures and Reuse	SBCARs
IEEE International Conference on Software Engineering and Service Sciences	ICSESS
International Conference on Software Testing, Verification and Validation	ICST
CM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering	ESEC/FSE
Working Conference on Reverse Engineering	WCRE
Student Research Workshop at the Conference of the European Chapter of the Association for Computational Linguistics	EACLSRW
Working IEEE/IFIP Conference on Software Architecture	WICSA
International Workshop on CrowdSourcing in Software Engineering	CSI-SE
International Workshop on Emerging Trends in Software Metrics	WETSoM
International Workshop on Conducting Empirical Studies in Industry	CESI
Annual Conference on Innovation and Technology in Computer Science Education	ITiCSE
International Conference on Software Quality. The Future of Systems and Software Development	SWQD
CSI International Conference on Software Engineering	CONSEG
IEEE International Conference on Software Maintenance and Evolution	ICSME
International Conference on Agile Software and Systems Development	XP
IEEE International Working Conference on Mining Software Repositories	MSR
Winter Simulation Conference	WSC
International Conference on Augmented Cognition. Neurocognition and Machine Learning	AC
IEEE Symposium on Computational Intelligence and Data Mining	CIDM
ACM Southeast 2018 Conference	ACMSE
International Conference on Conceptual Modeling	ER

REFERENCES

[1] D. I. K. Sjoeberg, "A survey of controlled experiments in software engineering," *IEEE Trans. Softw. Eng.*, vol. 31, no. 9, pp. 733–753, Sep. 2005.
[2] J. P. F. Almqvist, "Replication of controlled experiments in empirical software engineering—A survey," M.S. thesis, Dept. Comput. Sci., Lund Univ., Lund, Sweden, 2006.
[3] J. C. Carver, "Towards reporting guidelines for experimental replications: A proposal," in *Proc. 1st Int. Workshop Replication Empirical Softw. Eng.*, 2010, pp. 1–4.
[4] F. Q. B. da Silva, M. Suassuna, A. C. C. França, A. M. Grubb, T. B. Gouveia, C. V. F. Monteiro, and I. E. dos Santos, "Replication of empirical studies in software engineering research: A systematic mapping study," *Empirical Softw. Eng.*, vol. 19, no. 3, pp. 501–557, Jun. 2014.
[5] H. Larsson, E. Lindqvist, and R. Torkar, "Outliers and replication in software engineering," in *Proc. Asia–Pacific Softw. Eng. Conf. (APSEC)*, vol. 1, Dec. 2014, pp. 207–214.
[6] C. V. C. de Magalhães, F. Q. B. da Silva, R. E. S. Santos, and M. Suassuna, "Investigations about replication of empirical studies in software engineering: A systematic mapping study," *Inf. Softw. Technol.*, vol. 64, pp. 76–101, Aug. 2015.
[7] R. M. M. Bezerra, F. Q. B. da Silva, A. M. Santana, C. V. C. Magalhaes, and R. E. S. Santos, "Replication of empirical studies in software engineering: An update of a systematic mapping study," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Oct. 2015, pp. 1–4.
[8] B. Cartaxo, A. Almeida, E. Barreiros, J. Saraiva, W. Ferreira, and S. Soares, "Mechanisms to characterize context of empirical studies in software engineering," in *Proc. Experim. Softw. Eng. Latin Amer. Workshop (ESELAW)*, 2015, pp. 1–14.
[9] L. Falcao, W. Ferreira, A. Borges, V. Nepomuceno, S. Soares, and M. T. Baldassare, "An analysis of software engineering experiments using human subjects," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Oct. 2015, pp. 1–4.
[10] J. Siegmund, N. Siegmund, and S. Apel, "Views on internal and external validity in empirical software engineering," in *Proc. IEEE Int. Conf. Softw. Eng. (ICSE)*, vol. 1, May 2015, pp. 9–19.
[11] A. Santos, O. S. Gómez, and N. Juristo, "Analyzing families of experiments in SE: A systematic mapping study," *IEEE Trans. Softw. Eng.*, to be published.
[12] M. Shepperd, N. Ajienka, and S. Counsell, "The role and value of replication in empirical software engineering results," *Inf. Softw. Technol.*, vol. 99, pp. 120–132, Jul. 2018.
[13] N. Juristo and O. S. Gómez, "Replication of software engineering experiments," in *Proc. LASER Summer School Softw. Eng. (LASER)*, 2012, pp. 60–88.
[14] A. Brooks, J. Daly, J. Miller, M. Roper, and M. Wood, "Replication of experimental results in software engineering," Univ. Strathclyde, Glasgow, U.K., Tech. Rep. ISERN-96-10, 1996.
[15] F. J. Shull, J. C. Carver, S. Vegas, and N. Juristo, "The role of replications in empirical software engineering," *Empirical Softw. Eng.*, vol. 13, no. 2, pp. 211–218, 2008.
[16] N. Juristo and S. Vegas, "The role of non-exact replications in software engineering experiments," *Empirical Softw. Eng.*, vol. 16, no. 3, pp. 295–324, 2011.
[17] V. R. Basili, F. Shull, and F. Lanubile, "Building knowledge through families of experiments," *IEEE Trans. Softw. Eng.*, vol. 25, no. 4, pp. 456–473, Jul. 1999.
[18] O. S. Gómez, N. Juristo, and S. Vegas, "Understanding replication of experiments in software engineering: A classification," *Inf. Softw. Technol.*, vol. 56, no. 8, pp. 1033–1048, Aug. 2014.
[19] B. Kitchenham, "The role of replications in empirical software engineering—A word of warning," *Empirical Softw. Eng.*, vol. 13, no. 2, pp. 219–221, Apr. 2008.
[20] J. W. Daly, A. Brooks, J. Miller, M. Roper, and M. Wood, "Verification of results in software maintenance through external replication," in *Proc. IEEE Int. Conf. Softw. Maintenance (ICSM)*, Sep. 1994, pp. 50–57.
[21] A. Jedlitschka and D. Pfahl, "Reporting guidelines for controlled experiments in software engineering," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Nov. 2005, pp. 95–104.
[22] N. Juristo and A. M. Moreno, *Basics of Software Engineering Experimentation*. New York, NY, USA: Springer, 2013.
[23] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering: An Introduction*. Berlin, Germany: Springer-Verlag, 2012.

[24] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," *Ease*, vol. 8, pp. 68–77, Jun. 2008.

[25] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering—A systematic literature review," *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, Jan. 2009.

[26] B. Bernárdez, A. Durán, J. A. Parejo, N. Juristo, and A. Ruiz-Cortés, "Effects of mindfulness on conceptual modeling performance: A series of experiments," *IEEE Trans. Softw. Eng.*, to be published.

[27] B. Bernárdez, A. Durán, J. A. Parejo, and A. Ruiz-Cortés, "A controlled experiment to evaluate the effects of mindfulness in software engineering," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Sep. 2014, pp. 17–27.

[28] L. Montalvillo and O. Díaz, "Requirement-driven evolution in software product lines: A systematic mapping study," *J. Syst. Softw.*, vol. 122, pp. 110–143, Dec. 2016.

[29] *Business Process Model and Notation (BPMN) Version 2.0.2*, Object Manage. Group, Needham, MA, USA, 2014.

[30] P. Bourque and R. E. Fairley, Eds., *Guide to the Software Engineering Body of Knowledge, Version 3.0*. Washington, DC, USA: IEEE Computer Society, 2014.

[31] R. Ramač, B. Turhan, I. Karac, N. Juristo, and V. Mandić, "Lessons learned from a partial replication of an experiment in the context of a software engineering course," in *Proc. Int. Sci. Conf. Ind. Syst. (IS)*, 2017, pp. 198–203.

[32] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. Eval. Assessment Softw. Eng. (EASE)*, 2014, pp. 38:1–38:10.

[33] N. J. van Eck and L. Waltman. (2019). VOSviewer version 1.6.11. Centre for Science and Technology Studies (CWTS) of Leiden University. [Online]. Available: https://www.vosviewer.com/

[34] L. Hattori, M. D'Ambros, M. Lanza, and M. Lungu, "Answering software evolution questions: An empirical evaluation," *Inf. Softw. Technol.*, vol. 55, no. 4, pp. 755–775, Apr. 2013.

[35] L. Jiang, K. M. Carley, and A. Eberlein, "Assessing team performance from a socio-technical congruence perspective," in *Proc. Int. Conf. Softw. Syst. Process*, Jun. 2012, pp. 160–169.

[36] W. Mehmood, N. Shah, M. J. Khan, M. Memon, and M. Ikramullah, "Fine-granular model merge solution for model-based version control system," *Malaysian J. Comput. Sci.*, vol. 29, no. 3, pp. 225–246, 2016.

[37] D. Dig, K. Manzoor, R. E. Johnson, and T. N. Nguyen, "Effective software merging in the presence of object-oriented refactorings," *IEEE Trans. Softw. Eng.*, vol. 34, no. 3, pp. 321–335, May 2008.

[38] J. A. Parejo, S. Segura, P. Fernandez, and A. Ruiz-Cortés, "Exemplar: An experimental information repository for software engineering research," in *Proc. Jornadas Ingeniería del Softw. Bases Datos*, 2014, pp. 155–159.

[39] S. T. Acuña, M. N. Gómez, J. E. Hannay, N. Juristo, and D. Pfahl, "Are team personality and climate related to satisfaction and software quality? aggregating results from a twice replicated experiment," *Inf. Soft. Technol.*, vol. 57, no. 1, pp. 141–156, Jan. 2015.

[40] Ö. Albayrak and J. C. Carver, "Investigation of individual factors impacting the effectiveness of requirements inspections: A replicated experiment," *Empirical Softw. Eng.*, vol. 19, no. 1, pp. 241–266, 2014.

[41] S. Ali, T. Yue, and L. Briand, "Does aspect-oriented modeling help improve the readability of uml state machines?" *Softw. Syst. Model.*, vol. 13, no. 3, pp. 1189–1221, 2014.

[42] C. Apa, O. Dieste, E. G. Espinosa and E. R. C. Fonseca, "Effectiveness for detecting faults within and outside the scope of testing techniques: An independent replication," *Empirical Softw. Eng.*, vol. 19, no. 2, pp. 378–417, 2014.

[43] A. M. Aranda, O. Dieste, and N. Juristo, "Effect of domain knowledge on elicitation effectiveness: An internally replicated controlled experiment," *IEEE Trans. Softw. Eng.*, vol. 42, no. 5, pp. 427–451, May 2016.

[44] Ö. F. Arar and K. Ayan, "Deriving thresholds of software metrics to predict faults on open source software: Replicated case studies," *Expert Syst. Appl.*, vol. 61, pp. 106–121, Nov. 2016.

[45] G. Bavota, C. Gravino, R. Oliveto, A. De Lucia, G. Tortora, M. Genero, and J. A. Cruz-Lemus, "A fine-grained analysis of the support provided by UML class diagrams and ER diagrams during data model maintenance," *Softw. Syst. Model.*, vol. 14, no. 1, pp. 287–306, Feb. 2015.

[46] A. Brooks, J. Chambers, C. N. Lee, and F. Mead, "A partial replication with a sample size of one: A smoke test for empirical software engineering," in *Proc. Int. Workshop Replication Empirical Softw. Eng. (RESER)*, Oct. 2013, pp. 56–65.

[47] G. Cavalcanti, P. Accioly, and P. Borba, "Assessing semistructured merge in version control systems: A replicated experiment," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Oct. 2015, pp. 267–276.

[48] M. Ceccato, A. Marchetto, L. Mariani, C. D. Nguyen, and P. Tonella, "Do automatically generated test cases make debugging easier? An experimental assessment of debugging effectiveness and efficiency," *ACM Trans. Softw. Eng. Methodol.*, vol. 25, no. 1, pp. 5:1–5:38, 2015.

[49] D. Chatterji, J. C. Carver, N. A. Kraft, and J. Harder, "Effects of cloned code on software maintainability: A replicated developer study," in *Proc. Work. Conf. Reverse Eng. (WCRE)*, Oct. 2013, pp. 112–121.

[50] S. Donadelli, Y. C. Zhu, and P. C. Rigby, "Organizational volatility and post-release defects: A replication case study using data from Google Chrome," in *Proc. IEEE Int. Work. Conf. Mining Softw. Repositories (MSR)*, May 2015, pp. 391–395.

[51] W. Engelsman and R. Wieringa, "Understandability of goal concepts by requirements engineering experts," in *Proc. Int. Conf. Conceptual Model. (ER)*, 2014, pp. 97–106.

[52] S. U. Farooq and S. Quadri, "An externally replicated experiment to evaluate software testing methods," in *Proc. Eval. Assessment Softw. Eng. (EASE)*, Apr. 2013, pp. 72–77.

[53] M. Felderer and A. Herrmann, "Manual test case derivation from UML activity diagrams and state machines: A controlled experiment," *Inf. Softw. Technol.*, vol. 61, pp. 1–15, May 2015.

[54] K. R. Felizardo, E. F. Barbosa, R. M. Martins, P. H. D. Valle, and J. C. Maldonado, "Visual text mining: Ensuring the presence of relevant studies in systematic literature reviews," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 25, no. 5, pp. 909–928, 2015.

[55] K. R. Felizardo, S. R. S. Souza, and J. C. Maldonado, "The use of visual text mining to support the study selection activity in systematic literature reviews: A replication study," in *Proc. Int. Workshop Replication Empirical Softw. Eng. (RESER)*, Oct. 2013, pp. 91–100.

[56] K. Felizardo, E. Barbosa, and J. C. Maldonado, "A visual approach to validate the selection review of primary studies in systematic reviews: A replication study," in *Proc. Int. Conf. Softw. Eng. Knowl. Eng. (SEKE)*, 2013, pp. 141–146.

[57] D. M. Fernández and S. Wagner, "Naming the pain in requirements engineering: A design for a global family of surveys and first results from Germany," *Inf. Softw. Technol.*, vol. 57, no. 1, pp. 616–643, Jan. 2015.

[58] A. Fernández-Sáez, M. Genero, D. Caivano, and M. R. V. Chaudron, "Does the level of detail of UML diagrams affect the maintainability of source code?: A family of experiments," *Empirical Softw. Eng.*, vol. 21, no. 1, pp. 212–259, Feb. 2016.

[59] A. Fernández-Sáez, M. Genero, M. R. V. Chaudron, D. Caivano, and I. Ramos, "Are forward designed or reverse-engineered UML diagrams more helpful for code maintenance?: A family of experiments," *Inf. Softw. Technol.*, vol. 57, no. 1, pp. 644–663, Jan. 2015.

[60] F. Fittkau, S. Finke, W. Hasselbring, and J. Waller, "Comparing trace visualizations for program comprehension through controlled experiments," in *Proc. IEEE Int. Conf. Program Comprehension (ICPC)*, May 2015, pp. 266–276.

[61] A. França, F. Q. B. da Silva, A. L. C. Felix, and D. E. S. Carneiro, "Motivation in software engineering industrial practice: A cross-case analysis of two software organisations," *Inf. Softw. Technol.*, vol. 56, no. 1, pp. 79–101, Jan. 2014.

[62] D. Fucci, G. Scanniello, S. Romano, M. Shepperd, B. Sigweni, F. Uyaguari, B. Turhan, N. Juristo, and M. Oivo, "An external replication on the effects of test-driven development using a multi-site blind analysis approach," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, 2016, pp. 3:1–3:10.

[63] D. Fucci and B. Turhan, "On the role of tests in test-driven development: A differentiated and partial replication," *Empirical Softw. Eng.*, vol. 19, no. 2, pp. 277–302, 2014.

[64] D. Fucci and B. Turhan, "A replicated experiment on the effectiveness of test-first development," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Oct. 2013, pp. 103–112.

[65] B. Ghotra, S. McIntosh, and A. E. Hassan, "Revisiting the impact of classification techniques on the performance of defect prediction models," in *Proc. Int. Conf. Softw. Eng. (ICSE)*, vol. 1, May 2015, pp. 789–800.

[66] M. N. Gómez and S. Acuña, "A replicated quasi-experimental study on the influence of personality and team climate in software development," *Empirical Softw. Eng.*, vol. 19, no. 2, pp. 343–377, Apr. 2014.
[67] C. Gravino, G. Scanniello, and G. Tortora, "Source-code comprehension tasks supported by UML design models: Results from a controlled experiment and a differentiated replication," *J. Vis. Lang. Comput.*, vol. 28, pp. 23–38, Jun. 2015.
[68] T. Grbac, P. Runeson, and D. Huljenić, "A quantitative analysis of the unit verification perspective on fault distributions in complex software systems: An operational replication," *Softw. Qual. J.*, vol. 24, no. 4, pp. 967–995, Dec. 2016.
[69] L. Haaranen, P. Ihantola, J. Sorva, and A. Vihavainen, "In search of the emotional design effect in programming," in *Proc. Int. Conf. Softw. Eng. (ICSE)*, vol. 2, May 2015, pp. 428–434.
[70] I. Hadar, I. Reinhartz-Berger, T. Kuflik, A. Perini, F. Ricca, and A. Susi, "Comparing the comprehensibility of requirements models expressed in Use Case and Tropos: Results from a family of experiments," *Inf. Softw. Technol.*, vol. 55, no. 10, pp. 1823–1843, Oct. 2013.
[71] J. Harder and N. Göde, "Cloned code: Stable code," *J. Softw., Evol. Process*, vol. 25, no. 10, pp. 1063–1088, Oct. 2013.
[72] H. Huijgens and R. van Solingen, "A replicated study on correlating agile team velocity measured in function and story points," in *Proc. Int. Workshop Emerg. Trends Softw. Metrics (WETSoM)*, 2014, pp. 30–36.
[73] A. Idri and A. Zahi, "Software cost estimation by classical and fuzzy analogy for Web hypermedia applications: A replicated study," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, Apr. 2013, pp. 207–213.
[74] J. Itkonen and M. Mäntylä, "Are test cases needed? Replicated comparison between exploratory and test-case-based software testing," *Empirical Softw. Eng.*, vol. 19, no. 2, pp. 303–342, 2014.
[75] M. A. Javed and U. Zdun, "The supportive effect of traceability links in architecture-level software understanding: Two controlled experiments," in *Proc. Work. IEEE/IFIP Conf. Softw. Archit. (WICSA)*, Apr. 2014, pp. 215–224.
[76] J. Jung, K. Hoefig, D. Domis, A. Jedlitschka, and M. Hiller, "Experimental comparison of two safety analysis methods and its replication," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Oct. 2013, pp. 223–232.
[77] N. Juristo, S. Vegas, M. Solari, S. Abrahão, and I. Ramos, "A process for managing interaction between experimenters to get useful similar replications," *Inf. Softw. Technol.*, vol. 55, no. 2, pp. 215–225, Feb. 2013.
[78] M. Kalinowski, M. Felderer, T. Conte, R. Spínola, R. Prikladnicki, D. Winkler, D. Fernández, and S. Wagner, "Preventing incomplete/hidden requirements: Reflections on survey data from Austria and Brazil," in *Proc. Int. Conf. Softw. Quality Future Syst. Softw. Develop. (SWQD)*, 2016, pp. 63–78.
[79] M. Kalinowski, R. Spínola, T. Conte, R. Prikladnicki, D. Fernández, and S. Wagner, "Towards building knowledge on causes of critical requirements engineering problems," in *Proc. Int. Conf. Softw. Eng. Knowl. Eng. (SEKE)*, 2015, pp. 1–6.
[80] J. L. Krein, L. Prechelt, N. Juristo, A. Nanthaamornphong, J. C. Carver, S. Vegas, C. D. Knutson, K. D. Seppi, and D. L. Eggett, "A multi-site joint replication of a design patterns experiment using moderator variables to generalize across contexts," *IEEE Trans. Softw. Eng.*, vol. 42, no. 4, pp. 302–321, Apr. 2016.
[81] V. Lenarduzzi, I. Lunesu, M. Matta, and D. Taibi, "Functional size measures and effort estimation in agile development: A replicated study," in *Proc. Int. Conf. Agile Softw. Syst. Develop. (XP)*, 2015, pp. 105–116.
[82] A. Marcolino, E. Oliveira, Jr., I. Gimenes, and T. U. Conte, "Towards validating complexity-based metrics for software product line architectures," in *Proc. Brazilian Symp. Softw. Compon., Archit. Reuse (SBCARs)*, Sep./Oct. 2013, pp. 69–79.
[83] A. K. Massey, P. N. Otto, and A. I. Antón, "Evaluating legal implementation readiness decision-making," *IEEE Trans. Softw. Eng.*, vol. 41, no. 6, pp. 545–564, Jun. 2015.
[84] M. Goran and G. G. Tihana, "Co-evolutionary multi-population genetic programming for classification in software defect prediction: An empirical case study," *Appl. Soft Comput. J.*, vol. 55, pp. 331–351, Jun. 2017.
[85] N. Niu, A. Koshoffer, L. Newman, C. Khatwani, C. Samarasinghe, and J. Savolainen, "Advancing repeated research in requirements engineering: A theoretical replication of viewpoint merging," in *Proc. Int. Requirements Eng. Conf. (RE)*, Sep. 2016, pp. 186–195.
[86] B. Penzenstadler, J. Eckhardt, and D. Fernández, "Two replication studies for evaluating artefact models in RE: Results and lessons learnt," in *Proc. Int. Workshop Replication Empirical Softw. Eng. (RESER)*, Oct. 2013, pp. 66–75.
[87] J. Pow-Sang, "Evaluating and comparing perceptions between undergraduate students and practitioners in controlled experiments for requirements prioritization," in *Trends and Applications in Software Engineering*. 2016, pp. 189–199.
[88] C. Quesada-López, D. Madrigal, and M. Jenkins, "An empirical evaluation of automated function points," in *Proc. Ibero-Amer. Conf. Softw. Eng. (CIBSE)*, 2016, pp. 214–228.
[89] C. Quesada-Lápez and M. Jenkins, "An empirical validation of function point structure and applicability: A replication study," in *Proc. Ibero-Amer. Conf. Softw. Eng. (CIBSE)*, 2015, pp. 418–431.
[90] D. Reimanis, C. Izurieta, R. Luhr, L. Xiao, Y. Cai, and G. Rudy, "A replication case study to measure the architectural quality of a commercial system," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Sep. 2014, Art. no. 31.
[91] M. Ribeiro, P. Borba, and C. Kästner, "Feature maintenance with emergent interfaces," in *Proc. Int. Conf. Softw. Eng. (ICSE)*, 2014, pp. 989–1000.
[92] F. Ricca, G. Scanniello, M. Torchiano, G. Reggio, and E. Astesiano, "Assessing the effect of screen mockups on the comprehension of functional requirements," *ACM Trans. Softw. Eng. Methodol.*, vol. 24, no. 1, pp. 1:1–1:38, 2014.
[93] P. Runeson, A. Stefik, and A. Andrews, "Variation factors in the design and analysis of replicated controlled experiments," *Empirical Softw. Eng.*, vol. 19, no. 6, pp. 1781–1808, 2014.
[94] N. Salleh, E. Mendes, and J. Grundy, "Investigating the effects of personality traits on pair programming in a higher education setting through a family of experiments," *Empirical Softw. Eng.*, vol. 19, no. 3, pp. 714–752, 2014.
[95] J. A. M. Santos and M. G. de Mendonça, "Exploring decision drivers on god class detection in three controlled experiments," in *Proc. ACM Symp. Appl. Comput. (SAC)*, 2015, pp. 1472–1479.
[96] G. Scanniello, A. Marcus, and D. Pascale, "Link analysis algorithms for static concept location: An empirical assessment," *Empirical Softw. Eng.*, vol. 20, no. 6, pp. 1666–1720, Dec. 2015.
[97] G. Scanniello, C. Gravino, M. Risi, G. Tortora, and G. Dodero, "Documenting design-pattern instances: A family of experiments on source-code comprehensibility," *ACM Trans. Softw. Eng. Methodol.*, vol. 24, no. 3, pp. 14:1–14:35, May 2015.
[98] G. Scanniello, C. Gravino, G. Tortora, M. Genero, M. Risi, J. A. Cruz-Lemus, and G. Dodero, "Studying the effect of UML-based models on source-code comprehensibility: Results from a long-term investigation," in *Proc. Int. Conf. Product-Focused Softw. Process Improvement*, vol. 9459, 2015, pp. 311–327.
[99] G. Scanniello, M. Staron, H. Burden, and R. Heldal, "On the effect of using SysML requirement diagrams to comprehend requirements: Results from two controlled experiments," in *Proc. Eval. Assessment Softw. Eng. (EASE)*, 2014, p. 49.
[100] G. Scanniello and U. Erra, "Distributed modeling of use case diagrams with a method based on think-pair-square: Results from two controlled experiments," *J. Vis. Lang. Comput.*, vol. 25, no. 4, pp. 494–517, Aug. 2014.
[101] G. Scanniello, C. Gravino, M. Genero, J. A. Cruz-Lemus, and G. Tortora, "On the impact of UML analysis models on source-code comprehensibility and modifiability," *ACM Trans. Softw. Eng. Methodol.*, vol. 23, no. 2, pp. 13:1–13:26, 2014.
[102] C. A. Siebra and M. A. B. Mello, "The importance of replications in software engineering: A case study in defect prediction," in *Proc. ACM Res. Adapt. Convergent Syst. (RACS)*, 2015, pp. 376–381.
[103] M. Solari and S. Matalonga, "A controlled experiment to explore potentially undetectable defects for testing techniques," in *Proc. Int. Conf. Softw. Eng. Knowl. Eng. (SEKE)*, 2014, pp. 106–109.
[104] P. Sun and K. T. Stolee, "Exploring crowd consistency in a mechanical turk survey," in *Proc. Int. Workshop CrowdSourcing Softw. Eng. (CSI-SE)*, 2016, pp. 8–14.

[105] W. Sun, M. Aguirre-Urreta, and G. Marakas, "Effectiveness of pair and solo programming methods: A survey and an analytical approach," in *Proc. Amer. Conf. Inf. Syst. (AMCIS)*, 2015. [Online]. Available: https://dblp.org/rec/html/conf/amcis/Aguirre-UrretaM15

[106] P. Tramontana, M. Risi, and G. Scanniello, "Studying abbreviated vs. full-word identifier names when dealing with faults: An external replication," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Sep. 2014, Art. no. 64.

[107] A. Vetrò, W. Böhm, and M. Torchiano, "On the benefits and barriers when adopting software modelling and model driven techniques-An external, differentiated replication," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Oct. 2015, pp. 168–171.

[108] F. Wu, "Empirical tests of scale-free characteristic in open source software: A replicated case study," *Adv. Mater. Res.*, vol. 622, pp. 1933–1936, Feb. 2013.

[109] H. Yang, "Improved software cost estimation method based on COCOMO model and linear regression," *Adv. Mater. Res.*, vols. 989–994, pp. 1497–1500, Jul. 2014.

[110] S. Vegas, Ó. Dieste, and N. Juristo, "Difficulties in running experiments in the software industry: Experiences from the trenches," in *Proc. Int. Workshop Conducting Empirical Stud. Ind. (CESI)*, 2015, pp. 3–9.

[111] X. Chen, W. Zhang, P. Liang, and K. He, "A replicated experiment on architecture pattern recommendation based on quality requirements," in *Proc. IEEE Int. Conf. Softw. Eng. Service Sci. (ICSESS)*, Jun. 2014, pp. 32–36.

[112] F. Q. B. Da Silva, A. C. C. França, M. Suassuna, L. M. R. de Sousa Mariz, I. Rossiley, R. C. G. de Miranda, T. B. Gouveia, C. V. F. Monteiro, E. Lucena, E. S. F. Cardozo, and E. Espindola, "Team building criteria in software projects: A mix-method replicated study," *Inf. Softw. Technol.*, vol. 55, no. 7, pp. 1316–1340, Jul. 2013.

[113] A. C. Dias-Neto, S. Matalonga, M. Solari, G. Robiolo, and G. Travassos, "Toward the characterization of software testing practices in South America: Looking at Brazil and Uruguay," *Softw. Qual. J.*, vol. 25, no. 4, pp. 1145–1183, Dec. 2017.

[114] S. U. Farooq, S. M. K. Quadri, and N. Ahmad, "A replicated empirical study to evaluate software testing methods," *J. Softw., Evol. Process*, vol. 29, no. 9, p. e1883, 2017.

[115] D. Ferneández *et al.*, "Naming the pain in requirements engineering: Contemporary problems, causes, and effects in practice," *Empirical Softw. Eng.*, vol. 22, no. 5, pp. 2298–2338, 2017.

[116] O. Gómez, K. Cortés-Verdín, and C. J. Pardo, "Efficiency of software testing techniques: A controlled experiment replication and network meta-analysis," *E-Informatica Softw. Eng. J.*, vol. 11, no. 1, pp. 77–102, 2017.

[117] S. Herbold, A. Trautsch, and J. Grabowski, "Global vs. local models for cross-project defect prediction: A replication study," *Empirical Softw. Eng.*, vol. 22, no. 4, pp. 1866–1902, 2017.

[118] C. Khatwani, X. Jin, N. Niu, A. Koshoffer, L. Newman, and J. Savolainen, "Advancing viewpoint merging in requirements engineering: A theoretical replication and explanatory study," *Requirements Eng.*, vol. 22, no. 3, pp. 317–338, Sep. 2017.

[119] P. Peachock, N. Iovino, and B. Sharif, "Investigating eye movements in natural language and C++ source code—A replication experiment," in *Proc. Int. Conf. Augmented Cognition Neurocognition Mach. Learn. (AC)*, 2017, pp. 206–218.

[120] M. Riaz, J. King, J. Slankas, L. Williams, F. Massacci, C. Quesada-López, and M. Jenkins, "Identifying the implied: Findings from three differentiated replications on the use of security requirements templates," *Empirical Softw. Eng.*, vol. 22, no. 4, pp. 2127–2178, Aug. 2017.

[121] G. Scanniello, M. Risi, P. Tramontana, and S. Romano, "Fixing faults in C and Java source code: Abbreviated vs. full-word identifier names," *ACM Trans. Softw. Eng. Methodol.*, vol. 26, no. 2, pp. 6:1–6:43, 2017.

[122] D. Taibi, A. Janes, and V. Lenarduzzi, "How developers perceive smells in source code: A replicated study," *Inf. Softw. Technol.*, vol. 92, pp. 223–235, Dec. 2017.

[123] N. Yusop, M. Kamalrudin, M. Yusof, and S. Sidek, "Eliciting security requirments for mobile apps: A replication study," *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 15, pp. 3613–3622, 2017.

[124] S. Abrahão and E. Insfran, "Evaluating software architecture evaluation methods: An internal replication," in *Proc. Eval. Assessment Softw. Eng. (EASE)*, 2017, pp. 144–153.

[125] X. Franch, D. M. Fernández, M. Oriol, A. Vogelsang, R. Heldal, E. Knauss, G. H. Travassos, J. C. Carver, O. Dieste, and T. Zimmermann, "How do practitioners perceive the relevance of requirements engineering research? An ongoing study," in *Proc. Int. Requirements Eng. Conf. (RE)*, Sep. 2017, pp. 382–387.

[126] K. Lauenroth, E. Kamsties, and O. Hehlert, "Do words make a difference? An empirical study on the impact of taxonomies on the classification of requirements," in *Proc. Int. Requirements Eng. Conf. (RE)*, Sep. 2017, pp. 273–282.

[127] E. Marrese-Taylor and Y. Matsuo, "Replication issues in syntax-based aspect extraction for opinion mining," in *Proc. Student Res. Workshop Conf. Eur. Chapter Assoc. Comput. Linguistics (EACLSRW)*, 2017, pp. 23–32.

[128] R. M. de Mello, R. F. Oliveira, and A. F. Garcia, "On the influence of human factors for identifying code smells: A multi-trial empirical study," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, Nov. 2017, pp. 68–77.

[129] M. Nassif and M. P. Robillard, "Revisiting turnover-induced knowledge loss in software projects," in *Proc. IEEE Int. Conf. Softw. Maintenance Evol. (ICSME)*, Sep. 2017, pp. 261–272.

[130] S. Pearson, J. Campos, R. Just, G. Fraser, R. Abreu, M. D. Ernst, D. Pang, and B. Keller, "Evaluating and improving fault localization," in *Proc. Int. Conf. Softw. Eng. (ICSE)*, May 2017, pp. 609–620.

[131] N. Siegmund, S. Sobernig, and S. Apel, "Attributed variability models: Outside the comfort zone," in *Proc. ACM Joint Eur. Softw. Eng. Conf. Symp. Found. Softw. Eng. (ESEC/FSE)*, 2017, pp. 268–278.

[132] S. Assar, M. Borg, and D. Pfahl, "Using text clustering to predict defect resolution time: A conceptual replication and an evaluation of prediction accuracy," *Empirical Softw. Eng.*, vol. 21, no. 4, pp. 1437–1475, 2016.

[133] S. Akbarinasaji, B. Caglayan, and A. Bener, "Predicting bug-fixing time: A replication study using an open source software project," *J. Syst. Softw.*, vol. 136, pp. 173–186, Feb. 2018.

[134] M. Beck and J. Walden, "Using software birthmarks and clustering to identify similar classes and major functionalities," in *Proc. ACM Southeast Conf.*, 2018, Art. no. 11.

[135] B. Bernárdez, A. Durán, J. Parejo, and A. Ruiz-Cortés, "An experimental replication on the effect of the practice of mindfulness in conceptual modeling performance," *J. Syst. Softw.*, vol. 136, pp. 153–172, Feb. 2018.

[136] V. Caires, N. Rios, J. Holvitie, V. Leppänen, M. de Mendonça Neto, and R. Spínola, "Investigating the effects of agile practices and processes on technical debt—The viewpoint of the Brazilian software industry," in *Proc. Int. Conf. Softw. Eng. Knowl. Eng. (SEKE)*, Jul. 2018, pp. 506–511.

[137] D. Girardi, F. Lanubile, N. Novielli, and D. Fucci, "Sensing developers' emotions: The design of a replicated experiment," in *Proc. Int. Conf. Softw. Eng. (ICSE)*, 2018, pp. 51–54.

[138] S. Herbold, A. Trautsch, and J. Grabowski, "A comparative study to benchmark cross-project defect prediction approaches," *IEEE Trans. Softw. Eng.*, vol. 44, no. 9, pp. 811–833, Sep. 2018.

[139] T. Kosar, S. Gaberc, J. C. Carver, and M. Mernik, "Program comprehension of domain-specific and general-purpose languages: Replication of a family of experiments using integrated development environments," *Empirical Softw. Eng.*, vol. 23, no. 5, pp. 2734–2763, 2018.

[140] M. Mondal, S. Rahman, C. K. Roy, and K. A. Schneider, "Is cloned code really stable?" *Empirical Softw. Eng.*, vol. 23, no. 2, pp. 693–770, 2018.

[141] S. Nielebock, D. Krolikowski, J. Krüger, T. Leich, and F. Ortmeier, "Commenting source code: Is it worth it for small programming tasks?" *Empirical Softw. Eng.*, vol. 24, no. 3, pp. 1418–1457, 2018.

[142] J. F. S. Ouriques, E. G. Cartaxo, and P. D. L. Machado, "Test case prioritization techniques for model-based testing: A replicated study," *Softw. Qual. J.*, vol. 26, no. 4, pp. 1451–1482, Dec. 2018.

[143] J. I. P. Navarrete, O. Dieste, B. Marín, S. España, S. Vegas, O. Pastor, and N. Juristo, "Evaluating model-driven development claims with respect to quality: A family of experiments," *IEEE Trans. Softw. Eng.*, to be published.

[144] A. Panichella, F. M. Kifetew, and P. Tonella, "A large scale empirical comparison of state-of-the-art search-based test case generators," *Inf. Softw. Technol.*, vol. 104, pp. 236–256, Dec. 2018.

[145] K. Quille and S. Bergin, "Programming: Predicting student success early in CS1. A re-validation and replication study," in *Proc. Annu. Conf. Innov. Technol. Comput. Sci. Educ. (ITiCSE)*, 2018, pp. 15–20.

[146] G. Raura, E. R. C. Fonseca, J. W. Castro, T. Gualotuña, R. C. Mejía, M. T. Santillán, C. Pons, and O. Dieste, "Gender gap in computing: A preliminary empirical study," in *Proc. Ibero-Amer. Conf. Softw. Eng. (CIBSE)*, 2018, pp. 57–70.
[147] R. Ré, R. M. Meloca, D. N. R. Junior, M. A. da Cruz Ismael, and G. C. Silva, "An empirical study for evaluating the performance of multi-cloud APIs," *Future Gener. Comput. Syst.*, vol. 79, pp. 726–738, Feb. 2018.
[148] F. Ricca, M. Torchiano, M. Leotta, A. Tiso, G. Guerrini, and G. Reggio, "On the impact of state-based model-driven development on maintainability: A family of experiments using UniMod," *Empirical Softw. Eng.*, vol. 23, no. 3, pp. 1743–1790, Jun. 2018.
[149] A. R. Santos, I. do Carmo Machado, E. S. de Almeida, J. Siegmund, and S. Apel, "Comparing the influence of using feature-oriented programming and conditional compilation on comprehending feature-oriented software," *Empirical Softw. Eng.*, vol. 24, no. 3, pp. 1226–1258, 2018.
[150] G. Rong, H. Zhang, B. Liu, Q. Shan, and D. Shao, "A replicated experiment for evaluating the effectiveness of pairing practice in PSP education," *J. Syst. Softw.*, vol. 136, pp. 139–152, Feb. 2018.
[151] S. Shamshiri, J. M. Rojas, J. P. Galeotti, N. Walkinshaw, and G. Fraser, "How do automatically generated unit tests influence software maintenance?" in *Proc. Int. Conf. Softw. Test., Verification Validation (ICST)*, Apr. 2018, pp. 250–261.
[152] M. Shepperd, C. Mair, and M. Jørgensen, "An experimental evaluation of a de-biasing intervention for professional software developers," in *Proc. ACM Symp. Appl. Comput. (SAC)*, Apr. 2018, pp. 1510–1517.
[153] F. L. Siqueira, "Comparing the comprehensibility of requirements models: An experiment replication," *Inf. Softw. Technol.*, vol. 96, pp. 1–13, Apr. 2018.
[154] E. Souza, A. Moreira, J. Araújo, S. Abrahão, E. Insfran, and D. Silveira, "Comparing business value modeling methods: A family of experiments," *Inf. Softw. Technol.*, vol. 104, pp. 179–193, Dec. 2018.
[155] L. Vidács and M. Pinzger, "Co-evolution analysis of production and test code by learning association rules of changes," in *Proc. IEEE Int. Workshop Mach. Learn. Techn. Softw. Quality Eval.*, Mar. 2018, pp. 31–36.
[156] S. Young, T. Abdou, and A. Bener, "A replication study: Just-in-time defect prediction with ensemble learning," in *Proc. Int. Conf. Softw. Eng. (ICSE)*, May/Jun. 2018, pp. 42–47.
[157] N. Rios, R. O. Spínola, M. Mendonça, and C. Seaman, "The most common causes and effects of technical debt: First results from a global family of industrial surveys," in *Proc. ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas. (ESEM)*, 2018, Art. no. 39.
[158] S. Abrahão, C. Gravino, E. Insfran, G. Scanniello, and G. Tortora, "Assessing the effectiveness of sequence diagrams in the comprehension of functional requirements: Results from a family of five experiments," *IEEE Trans. Softw. Eng.*, vol. 39, no. 3, pp. 327–342, Mar. 2013.
[159] A. Aranda, O. Dieste, and N. Juristo, "Evidence of the presence of bias in subjective metrics: Analysis within a family of experiments," in *Proc. Eval. Assessment Softw. Eng. (EASE)*, 2014, Art. no. 24.
[160] D. Binkley, M. Davis, D. Lawrie, J. I. Maletic, C. Morrell, and B. Sharif, "The impact of identifier style on effort and comprehension," *Empirical Softw. Eng.*, vol. 18, no. 2, pp. 219–276, 2013.
[161] M. Ceccato, M. Di Penta, P. Falcarin, F. Ricca, M. Torchiano, and P. Tonella, "A family of experiments to assess the effectiveness and efficiency of source code obfuscation techniques," *Empirical Softw. Eng.*, vol. 19, no. 4, pp. 1040–1074, 2014.
[162] A. Fernandez, S. Abrahão, and E. Insfran, "Empirical validation of a usability inspection method for model-driven Web development," *J. Syst. Softw.*, vol. 86, no. 1, pp. 161–186, Jan. 2013.
[163] J. Gonzalez-Huerta, E. Insfran, S. Abrahão, and G. Scanniello, "Validating a model-driven software architecture evaluation and improvement method: A family of experiments," *Inf. Softw. Technol.*, vol. 57, no. 1, pp. 405–429, Jan. 2015.
[164] L. Guerrouj, M. Di Penta, Y.-G. Guéhéneuc, and G. Antoniol, "An experimental investigation on the effects of context on source code identifiers splitting and expansion," *Empirical Softw. Eng.*, vol. 19, no. 6, pp. 1706–1753, 2014.
[165] A. I. Molina, M. A. Redondo, M. Ortega, and C. Lacave, "Evaluating a graphical notation for modeling collaborative learning activities: A family of experiments," *Sci. Comput. Program.*, vol. 88, pp. 54–81, Aug. 2014.
[166] A. Moraes, W. L. Andrade, and P. D. L. Machado, "A family of test selection criteria for timed input-output symbolic transition system models," *Sci. Comput. Program.*, vol. 126, pp. 52–72, Sep. 2016.
[167] J. M. Morales, E. Navarro, P. Sánchez, and D. Alonso, "A family of experiments to evaluate the understandability of TRiStar and i* for modeling teleo-reactive systems," *J. Syst. Softw.*, vol. 114, pp. 82–100, Apr. 2016.
[168] B. Penzenstadler, "Sustainability analysis and ease of learning in artifact-based requirements engineering: The newest member of the family of studies (It's a girl!)," *Inf. Softw. Technol.*, vol. 95, pp. 130–146, Mar. 2018.
[169] G. Reggio, F. Ricca, G. Scanniello, F. Di Cerbo, and G. Dodero, "On the comprehension of workflows modeled with a precise style: Results from a family of controlled experiments," *Softw. Syst. Model.*, vol. 14, no. 4, pp. 1481–1504, Oct. 2015.
[170] G. Santos, T. Conte, G. Travassos, R. Prikladnicki, A. Rocha, N. Franco, and K. Weber, "Towards successful software process improvement initiatives: Experiences from the battlefield," in *Proc. Amer. Conf. Inf. Syst. (AMCIS)*, 2015, pp. 1–12.
[171] M. Shahin, P. Liang, and Z. Li, "Do architectural design decisions improve the understanding of software architecture? Two controlled experiments," in *Proc. IEEE Int. Conf. Program Comprehension (ICPC)*, Jun. 2014, pp. 3–13.
[172] P. Singh and B. Suri, "Quality metrics for conceptual model of data warehouse," in *Proc. CSI Int. Conf. Softw. Eng. (CONSEG)*, 2013, pp. 98–103.
[173] C.-A. Sun, Y. Zai, and H. Liu, "Evaluating and comparing fault-based testing strategies for general Boolean specifications: A series of experiments," *Comput. J.*, vol. 58, no. 5, pp. 1199–1213, May 2013.
[174] A. Teran-Somohano, O. Dayıbas, L. Yilmaz, and A. Smith, "Toward a model-driven engineering framework for reproducible simulation experiment lifecycle management," in *Proc. Winter Simulation Conf.*, Dec. 2014, pp. 2726–2737.
[175] M. Torchiano, G. Scanniello, F. Ricca, G. Reggio, and M. Leotta, "Do UML object diagrams affect design comprehensibility? Results from a family of four controlled experiments," *J. Vis. Lang. Comput.*, vol. 41, pp. 10–21, Aug. 2017.

MARGARITA CRUZ is currently an Assistant Professor with the University of Seville, where she was teaching databases for more than 30 years. Her current research interest includes empirical software engineering, specifically in methodological aspects of experiment replications such as the specification of changes between replications and its reporting not only in software engineering but also in other research areas.

BEATRIZ BERNÁRDEZ is currently an Assistant Professor with the University of Seville. Her current research interests include empirical software engineering, requirements engineering, and applications of mindfulness in software engineering. She was in collaboration with some international conferences such as ESEM 2007 and SPLC 2017. She has served as a Reviewer for the IEEE Transactions of Software Engineering and Empirical Software Engineering Journal.

AMADOR DURÁN is currently an Associate Professor in software engineering with the University of Seville. His current research interests include requirements engineering, software variability, empirical software engineering, business process modelling, formal methods, and metamorphic testing. He is the author of the REM requirements management tool used by universities and companies in various countries. He also serves regularly as a Reviewer for international journals and conferences.

JOSÉ A. GALINDO received the Ph.D. degree (Hons.) from the University of Seville and the Ph.D. degree (Hons.) from the University of Rennes 1, in March 2015. He is a Juan de la Cierva Research Fellow of the University of Seville. He received the Best National Thesis Award by SISTEDES. He has developed his Postdoctoral research activity with INRIA, France. His current research interests include software product lines, product configuration, testing, and the evolution of highly configurable systems.

ANTONIO RUIZ-CORTÉS is currently a Full Professor in software and service engineering and the Head of the Applied Software Engineering Group, University of Seville. His current research interests include service-oriented computing, business process management, and testing and software product lines. He was a recipient of the Most Influential Paper of SPLC 2017 Award. He is an Associate Editor of *Computing* (Springer).