Empir Software Eng (2014) 19:501–557
DOI 10.1007/s10664-012-9227-7

Replication of empirical studies in software engineering research: a systematic mapping study

Fabio Q. B. da Silva · Marcos Suassuna · A. César C. França · Alicia M. Grubb · Tatiana B. Gouveia · Cleviton V. F. Monteiro · Igor Ebrahim dos Santos

Published online: 1 September 2012

Editor: Natalia Juristo

Abstract In this article, we present a systematic mapping study of replications in software engineering. The goal is to plot the landscape of current published replications of empirical studies in software engineering research. We applied the systematic review method to search and select published articles, and to extract and synthesize data from the selected articles that reported replications. Our search retrieved more than 16,000 articles, from which we selected 96 articles, reporting 133 replications performed between 1994 and 2010, of 72 original studies. Nearly 70 % of the replications were published after 2004 and 70 % of these studies were internal replications. The topics of software requirements, software construction, and software quality concentrated over 55 % of the replications, while software design, configuration management, and software tools and methods were the topics with the smallest number of replications. We conclude that the number of replications has grown in the last few years, but the absolute number of replications is still small, in particular considering the breadth of topics in software engineering. We still need incentives to perform external replications, better standards to report empirical studies and their replications, and collaborative research agendas that could speed up development and publication of replications.

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

Preliminary and partial results of this mapping study were published and presented at the 2nd International Workshop on Replication in Empirical Software Engineering Research (RESER'2011).

F. Q. B. da Silva (✉) · M. Suassuna · A. C. C. França · T. B. Gouveia · C. V. F. Monteiro · I. E. dos Santos
Centre for Informatics, Federal University of Pernambuco, Recife, PE, Brazil
e-mail: fabio@cin.ufpe.br

A. M. Grubb
Department of Computer Science, University of Toronto, Toronto, ON, Canada

1 Introduction

Replications of empirical studies are regarded as an essential activity in the construction of knowledge in any empirical science. As pointed by Karl Popper, "We do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them" (Popper 1959). Lindsay and Ehrenberg argue that replication "… is needed not merely to validate one's findings, but more importantly, to establish the increasing range of radically different conditions under which the findings hold, and the predictable exceptions" (Lindsay and Ehrenberg 1993). For Schmidt, "[a] replication experiment to demonstrate that the same findings can be obtained in any other place by any other researcher … is the proof that the experiment reflects knowledge that can be separated from the specific circumstances (such as time, place, or persons) under which it was gained" (Schmidt 2009).

In software engineering, the first article, that explicitly reported a replication of an empirical study, was published in 1994 (Daly et al. 1994). Around the same period, Brooks et al. (1995) put together a set of principles for replicating software engineering studies, by refining and synthesizing principles from other disciplines. In the late 1990s, Basili et al. (1999) discussed a framework to organize sets of related experiments (families) and the generation of knowledge from such sets. These seminal works inspired and guided researchers developing replications, as well as researchers studying the issues associated with conducting and reporting replications.

More recently, the empirical software engineering community began to address several important issues related to replication, such as the role of lab packages to support close replications (Kitchenham 2008; Shull et al. 2008), the lack of incentives for undertaking replications (Kitchenham 2008), the importance of tacit knowledge (Shull et al. 2002), the issue of communication between researchers (Vegas et al. 2006), the need for having reporting guidelines specific for replications (Carver 2010), the difficulties of replications of studies involving human subjects (Lung et al. 2008; França et al. 2010), the role of different types of replications (Gómez et al. 2010a, b; Krein and Knutson 2010), and the generation of knowledge from differences in replications (Juristo and Vegas 2009), among others.

The growing interest of the research community in performing and studying replications resulted in the organization of two editions of the International Workshop on Replication in Empirical Software Engineering Research (RESER), in 2010[1] and 2011.[2]

Considering the importance of replications for empirical sciences in general and for empirical software engineering in particular, one would expect to find a body of knowledge that provides clear and unambiguous definitions for central questions like 'what exactly is a replication experiment?' or 'what exactly is a successful replication?' (Schmidt 2009), and 'what are all types of replication and their respective roles?' Furthermore, one would expect to find guidelines on how to perform and report replications complementing existing guidelines to perform experiments and other types of empirical studies.

As discussed above, some of these issues have been studied in software engineering over the last two decades. However, these studies and the scientific debate about their results are in an initial stage. The published work does use clear-cut definitions of terms and concepts, and there is no generally accepted taxonomy to distinguish between types of replications and their roles in generating scientific knowledge. Schmidt noted, "the word replication [was] used as a collective term to describe various meanings in different contexts" (Schmidt 2009). We expect these issues to be addressed as the field of empirical software engineering matures

[1] http://dl.acm.org/citation.cfm?doid=1838687.1838698 (last visited April, 2012)
[2] http://dl.acm.org/citation.cfm?doid=2088883.2088889 (last visited April, 2012)

and shared theoretical and practical understanding of replications increase. A comprehensive review of the replication literature in software engineering is an important step towards this understanding.

However, until the publication of a preliminary set of results by da Silva et al. (2011b), no extensive body of research had been published to provide a comprehensive picture of the available material on replications in Software Engineering. Our goal in this article is to plot the landscape of the current scientific work reporting replications of empirical software engineering research and the original replicated studies, extending and complementing the preliminary results published by da Silva et al. (2011b). We believe this extended and improved version of our results provides more in-depth information about replication work. We hope to contribute to the debate about the conceptual and practical issues related to replications and their role in software engineering research.

For the research method, we performed a systematic mapping study (Arksey and O´ Malley 2005) of the scientific literature, guided by the following general research question:

- RQ: What is the current state of the replication work of empirical studies performed in software engineering research?

We followed the guidelines proposed by Kitchenham and Charters (2007) to build the protocol for our mapping study. The automatic and manual search procedures retrieved more than 16,000 articles, from which we selected 96 articles reporting 133 replications of 72 original studies. We searched for articles published until and including 2010. The selected articles were published between 1994 and 2010. We systematically structured and analyzed data extracted from these articles to answer the following seven specific research questions:

- RQ1: What is the evolution in the number and type (internal and external) of replications over the years?
- RQ2: Which individuals and organizations are most active in replications?
- RQ3: What software engineering topics have been addressed by replications?
- RQ4: What research methods are being replicated?
- RQ5: What sets of replications were found?
- RQ6: Did the replications confirm the results of the original studies?
- RQ7: What was the elapsed time between the replications and corresponding original studies?

We provide the rationale and motivation for each research questions before presenting their results in Section 4.2.

The rest of this article is structured into sections. In Section 2, we provide a brief discussion of concepts and related work. In Section 3, we describe the systematic mapping study protocol. In Section 4, we present a comprehensive set of results of the mapping study. In Section 5, we present a discussion of the results, and limitations and threats to validity of our mapping study. Finally, in Section 6, we present some conclusions and directions for future work.

2 Background and Related Work

As briefly discussed in the Introduction, there is little agreement about nomenclature and definition of concepts about replication in many empirical sciences and in empirical software engineering in particular. Although, a deeper debate about nomenclature and conceptual definitions is out of the scope of this paper, we need some definitions in order to guide the

mapping study process. In this section, we provide such definitions, briefly describe three related works, and discuss how this article improved the preliminary results published by da Silva et al. (2011b).

2.1 Concepts and Definitions

According to La Sorte, "replication refers to a conscious and systematic repeat of an original study" (La Sorte 1972). This definition implies that a replication must be explicitly related (conscious repetition) to a previous study, the *original study*. Juristo and Vegas define replication as "the repetition of an experiment to double-check its results" (Juristo and Vegas 2009), which also implies an explicit relationship with a previous study. Similarly, A Dictionary of Social Sciences (Gould and Kolb 1964) defines replication as "a repetition of a research procedure to check the accuracy or truth of the findings reported".

According to these three definitions, empirical studies that address similar questions or hypothesis, but without explicit reference to one or more previous studies that can be considered the original study, should not be considered replications. Therefore, in our mapping study we do not consider as replications the studies that Krein and Knutson (2010) have classified as independent replications. The reason is that their definition of independent replication admits studies to be called replications without a reference (direct or indirect) to an original study.

In the literature, researchers have used different terms to define types or classifications of replications. Gómez et al. (2010b) performed a literature review looking for studies in various scientific disciplines that provided classifications of replications. They found 18 studies that proposed different classification schemes and descriptions of the roles for each type of replication in the construction of scientific knowledge.

As mentioned above, a deeper discussion about types of replications is outside the scope of this article. We use a classification of replications as internal and external as defined by Brooks et al. (2007), which is important to show the evolution of replication work through the years. A replication is defined as internal if the original researchers performed the replication. Likewise, a replication is defined as external if independent researchers performed the replication. To make this definition more precise, we considered a replication to be internal if any individual researcher involved in the original experiment was also involved in the replication. We discuss how we operationalized this definition and its limitations in Section 4.1.

As a final conceptual issue, we group replications related to the same original study into *sets* and analyze these *sets* in our mapping study. Basili et al. (1999) introduced the concept of family of experiments as a "framework for organizing sets of related studies" in such a way that "experiments can be viewed as part of common families of studies, rather than being isolated events". We do not use the concept of family to group the replications in our mapping study because this concept seemed to be more general than the concept of replication we use, in the sense that a family could have various sets of replications related to different original studies. Therefore, we prefer to use the generic term *set* to refer to all replications that refer, directly or indirectly (via other replications) to the same original study.

2.2 Related Work

Sjøberg et al. (2005) performed a systematic review of controlled experiments in software engineering published between 1993 and 2002. This review performed manual search on some leading journals and conference proceedings, but no automatic search was performed on online databases. From the 103 selected experiments, the authors found that 18 % (20/

103) were replications grouped in 14 series of experiments. Only nine replications were external, according to their definition, which was consistent with the one used in our current study. The focus of this systematic review was not replications and the review article did not provide the coverage and the depth of analysis included in our review.

Almqvist (2006) extended the scope of the studies found by Sjøberg et al. (2005) so that all studies reporting replications of a given original study were included and analyzed. The author added 11 studies to the 20 replications found by Sjøberg et al. (2005), and grouped all 31 studies into 20 series of experiments. Although the guidelines to perform a systematic review were followed, Almqvist (2006) only performed manual search over a limited set of nine journals and three conference proceedings looking for articles published between 1993 and 2002. This review identified the journals and conferences where replications were published and the researchers and institutions most active in performing replications. Almqvist (2006) classified replications as internal and external, using the same concept we used in our review, which was discussed in the previous section. The author also analyzed whether replications confirmed the results of the original studies, and found that internal replications had a tendency to confirm the original results, whereas external replications tended to disconfirm the originals, which is consistent with our findings (RQ6, Section 4.2.2). Although Almqvist's Master Dissertation has been cited in several articles, we have not found a peer-reviewed article presenting his results in conference proceedings or scientific journals.

In the first edition of the RESER workshop (2010), Carver (2010) presented the results of a literature review in which a simple automated search using "replication" and "replicated" as the search string found 15 replications. Carver's review added five new studies to the set reviewed by Almqvist (2006). The central focus of Carver's work was on the information content and structure of the reports about replications. Carver found that the replications were not reported in a consistent manner and the publications did not report the same type of information or the same level of detail. Carver concluded by proposing a series of guidelines for reporting replications, and we incorporated some of his suggestions in the quality assessment criteria described in Section 3.5 and presented in detail in Appendix B.

The 36 distinct studies reporting replications found and analyzed in the three reviews discussed above were analyzed for selection in our current review. We compare our findings with those of Sjøberg et al. (2005) and Almqvist (2006) in Section 4.2, when answering our research questions. We extend the coverage of these three related studies with 57 new articles reporting 89 more replications. Furthermore, we provide a deeper analysis of several aspects not addressed by the previous studies, such as, software engineering topics, research methods, and the quality of the studies. Our study also shows various temporal analysis that are relevant to understanding the state of replication in our field, which have not been performed in the other reviews.

2.3 Improvements from the Preliminary Results of this Mapping Study

We presented some preliminary results of this mapping study at the 2nd International Workshop on Replication in Empirical Software Engineering Research (RESER'2011) and the article was published in the workshop proceedings (da Silva et al. 2011b). In the workshop paper, we presented the results from the review of 93 articles reporting 125 replications performed between 1994 and 2010, of 76 original studies.

For this article, we improved and extended those results in six ways:

- We reanalyzed the selected papers resulting in an adjusted set of studies, which is further explained in Section 3.3.

- We added a complete account of the quality assessment process and included the results, as well as a more complete discussion about these results (see Appendix B).
- We added two new research questions to guide important discussions about the relationships between replications and the original studies (see RQ6 and RQ7).
- We largely extended and improved the presentation of the results and the discussion about the answers of each research question.
- We constructed new mappings to connect the answers to individual research questions, which allows for the visualization of their interactions.
- We provided the complete list of references to the replications and original studies (Appendix A).

Although the results presented in this article are essentially based on the same data set as da Silva et al. (2011b), the new added information and the derived discussions represent significant improvements to the preliminary results. We believe this article contributes to a comprehensive understanding of the landscape of replications of empirical studies in software engineering.

3 Review Method

The scientific literature differentiated several types of systematic reviews (Petticrew and Roberts 2006), including the following:

- Conventional systematic reviews (Petticrew and Roberts 2006), which aggregated results about the effectiveness of a treatment, intervention, or technology, and were related to specific research questions such as *Is intervention I on population P more effective for obtaining outcome O in context C than comparison treatment C*?
- Mapping (or scoping) studies (Arksey and O'Malley 2005) aimed to identify all research related to a specific topic, i.e., to answer broader questions related to trends in research. Typical questions were exploratory, e.g., *What do we know about topic T*?

In this article, we performed a *mapping study* (Arksey and O'Malley 2005) of replications of empirical studies in software engineering. This work is classified as a secondary study since it is a review of primary studies. The conceptual work on conventional systematic literature reviews (SLR) (Petticrew and Roberts 2006) and the guidelines for performing SLR in software engineering presented by Kitchenham and Charters (2007) were followed to plan and execute this mapping study. Our goal was to collect evidence that could be used to guide research and practice, so we consider this mapping study to be part of the evidence-based software engineering effort (Kitchenham et al. 2004).

In the rest of the article, we use the term *paper* to refer to the published work (article or other form of publication) analyzed in this review. We use *original study* (or simply original) or *replication* to refer to an experiment or other type of empirical study reported in the paper. In our review, a paper may have reported one or more empirical studies. In particular, one paper may have reported the original study and one or more replications together. Finally, the term *study* is used to refer to a replication or original study when the distinction is not important.

3.1 Inclusion and Exclusion Criteria

We searched the literature looking for papers that reported two types of studies (inclusion criteria):

1. Replications of empirical studies in software engineering.

2. Conceptual and theoretical works about replications, including theories, definitions, taxonomies, lessons learned, etc.

We excluded papers that met at least one of the following seven exclusion criteria:

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

3.2 Data Sources and Search Strategy

Our search process combined automatic and manual search to achieve high coverage. We performed manual search on relevant journals, conference proceedings, and on the list of primary studies analyzed in the three reports of related research (see Table 1 for a complete list of our manual sources). We looked for titles and abstracts of all papers in each source used in the manual search, using the same procedure applied to the list of papers returned in the automatic search. Therefore, both searches were compatible and created an audit trail that could be used to repeat the process. The use of manual search is supported in the literature about systematic reviews to complement and extend the coverage of automatic searches (Petticrew and Roberts 2006; Kitchenham and Charters 2007). In particular, manual searches are important to cover the cases where published articles are available in the manual sources but have not yet been indexed by the search engines used in the automatic search.

The automatic search was performed in five search engines and indexing systems (see Table 2 for a complete list of our automatic sources). Automatic searches were performed on the entire paper on all engines but Scopus, which does not perform full-text search. For this engine, the search was performed on Title and Abstract.

The search string used in the automatic search was constructed based on three search terms extracted from the general research question (see RQ in the Section 1): replication, empirical study, and software engineering. Synonyms for replication were found in the literature (in particular in the 18 studies reviewed by Gómez et al. (2010b)) and in consultation with empirical software engineering specialists. For the reasons explained in Section 2.1, we did not use "family of experiments" as synonym for replication. We used the

Table 1 Manual sources

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

Table 2 Automatic sources

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

types of empirical studies discussed in Easterbrook et al. (2007). We explicitly included plural forms of our definitions because several search engines did not treat wildcards as expected. Finally, we did not use any synonym for software engineering to build our search string; synonyms were joined with OR and the set of synonyms for each term were joined with AND (see Fig. 1).

We constructed the search string through several iterations and pilot tests to ensure that we used a comprehensive set of synonyms to allow for high coverage while keeping the number of retrieved articles under control. We used a refinement process similar to Zhang et al. (2010) to include new search terms from previously selected articles. Considering the high number of results from our automatic search process (over 16,000 articles) we believe we achieved a reasonable coverage level with our search string in the automatic search.

Five researchers performed the manual and automatic searches working individually on a given engine or set of manual sources. The researchers evaluated the results from the automatic search (n=16,055) by looking at the title and abstract, and excluding the papers that were clearly not relevant. The resulting papers (n=382) were merged with 74 potentially relevant papers found in the manual search. At this point, 58 duplicated studies were found and removed. When a study had been published in more than one journal or conference, all versions were reviewed for the purpose of data extraction. In this case, we used the first publication in all time-based analyses to track replication activities over time.

We finished the search process with a set of 398 potentially relevant papers. JabRef[3] and Mendeley[4] were used to support the search and selection process.

3.3 Study Selection

The list of 398 potentially relevant studies was analyzed for final selection (see Fig. 2). Four researchers worked in the selection process. Initially, the 398 papers were divided in two subsets of 199 papers, and each subset was assigned to a group of two researchers. In each group, each researcher worked independently to analyzed all the papers in their assigned set. The researchers applied the inclusion and exclusion criteria (see Section 3.2) on the potentially relevant papers after reading the abstract, introduction, and conclusion of each paper. Another researcher that did not participate in the original group, worked to solve the discrepancies and when an agreement was not possible, the differences were solved in a consensus meeting. This process selected 134 papers considered relevant for data extraction and analysis.

[3] http://jabref.sourceforce.net. JabRef is an open source bibliography reference manager. We used JabRef to record the data extracted from the articles, including the reference data and extracts of the text that we used to answer the research questions.

[4] http://www.mendeley.com. We used Mendeley to share the consolidated references of the selected papers on the Web, so multiple researchers could access them.

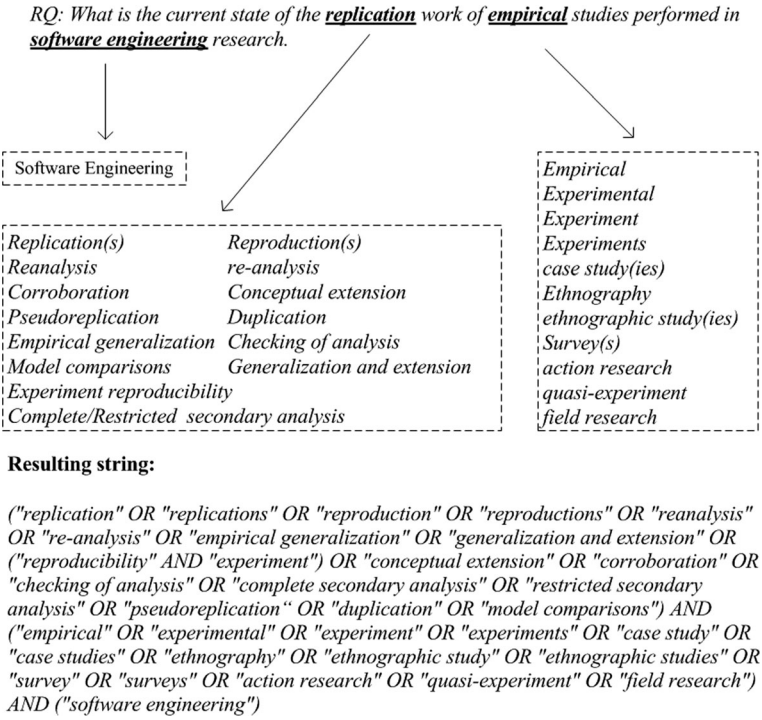


Fig. 1 Search string construction

The papers resulting from the application of inclusion and exclusion criteria were classified as reports of replications (inclusion criterion 1) or conceptual work about

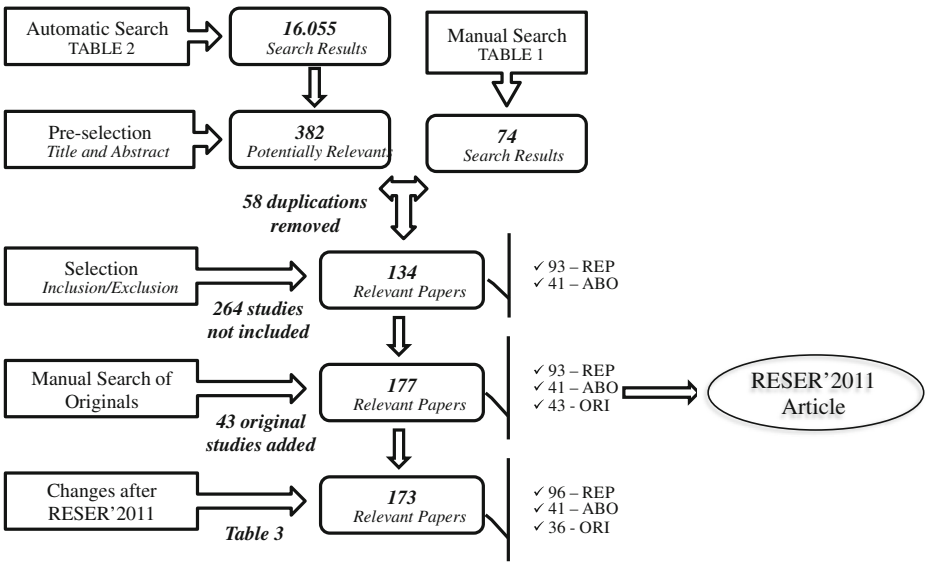


Fig. 2 Stages of the search and selection processes

replication (inclusion criterion 2). At the end of this classification, we found 93 reports of replications (REP) and 41 conceptual works about replication (ABO). In this article, we did not analyze the ABO set, which was part of our ongoing research.

We then retrieved papers reporting the original studies for each replication in the REP set. We found 43 papers reporting original studies (ORI). At this point we analyzed all 93 papers in REP and 43 papers in ORI, and published some preliminary results at the RESER'2011 (da Silva et al. 2011b).

We used the discussions during the RESER'2011workshop and the suggestions we received from participants to improve our analysis and to update the set of selected articles reporting replications and original studies. We performed these improvements in October 2011 using the same procedures defined in the review protocol. The summary of the changes is as follows:

- We decided to remove one set of articles reporting a longitudinal study, which was composed of one original study and three subsequent follow up studies. As pointed out by da Silva et al. (2011b), the inclusion of this set was a matter of debate in the preparation of the preliminary results and the discussions during the workshop convinced us that it should not have been considered a replication (see Section 5.2, Question 1).
- We regrouped three sets of replications that were analyzed separately by da Silva et al. (2011b) after consulting with the authors of the studies, and added one article reporting the original study that was common to all sets and was not included by da Silva et al. (2011b).
- We reanalyzed five articles reporting original studies in da Silva et al. (2011b) and considered that they also reported one or more replications together with the original study. These articles were moved from the set of articles reporting original studies (ORI) to the set of articles reporting replications (REP).
- We discovered one article reporting four replications, which was not found in our search process because it had not been indexed in any search engine and was not published in the manual sources we searched. This new article did not introduce a new original study, but added four new replications in a set that we had already found.

Table 3 shows the results of these changes in the main numbers of our mapping study, together with a summary of the justification for the changes.

Therefore, the number of relevant papers analyzed in this article was 132 of which 96 reported replications (REP) and 36 reported solely original studies (ORI). In the REP set, 60 papers reported replications and the original studies in the same paper and 36 reported only replications.

3.4 Data Extraction

Four researchers worked in the data extraction process. Table 4 shows the data extracted from the papers in the REP and ORI sets. Initially, the papers were divided into two subsets, and each subset was assigned to a group of two researchers. In each group, each researcher worked independently to extract data from all papers in their assigned subset, guided by an extraction form implemented in MS Excel™ and with the support of JabRef and Mendeley whenever possible. A third researcher reviewed the disagreements in the extracted data. When an agreement was not achieved, the differences were solved in a consensus meeting.

With the support of JabRef and Mendeley, we automatically extracted objective data from the papers including publication title, year of publication, authors' names, authors' affiliation, and country. Inter-rater agreement measures were calculated for the extraction of subjective data, which could have resulted in inconsistencies, including software engineering topic, research method, unit of analysis, and confirmation of original's results. In these

Table 3 Changes between da Silva et al. (2011b) and the current article

	(Da Silva et al. 2011b)	Current	Justification
Articles Reporting Replications (REP)	93	96	• Three articles reporting the longitudinal survey were removed. • Five articles were moved from the set ORI. • One new article not previously selected was added to the set.
Articles Reporting Original Studies (ORI)	43	36	• The original study of the longitudinal survey was removed. • Five articles were moved to the set REP • One article was added due to the regrouping of three sets of replications
Number of Original Studies	76	72	• One original study from the longitudinal survey was removed. • Three original studies were removed after regrouping.
Number of Replications	125	133	• Three replications of the longitudinal study were removed. • 7 new replications were added after the reanalysis of the five articles reporting originals. • Four new replications were added with inclusion of the new article not previously found.

cases, we used Cohen's Kappa (κ) coefficient (Cohen 1960) to measure the agreement level before disagreements were resolved. Cohen's Kappa coefficient is a measure of inter-rater agreement and it is used in systematic reviews to assess the internal consistency of processes that are performed by more than one researcher, such as selection of articles and data extraction. In other words, to provide a sense of how ambiguous these processes were. A high Kappa value (above 0.70) showed good agreement and indicated that the process was not ambiguous and was used coherently by the raters.

Table 4 Data extracted from each paper

Data	Description
Publication Title	Title of the paper
Authors	Name of all authors
Year	Year of publication of the paper reporting the replication
Research Organizations	Affiliation of the authors
Country	Country where the organization is located
Replication Type	Internal/External
Report Type	Original-Included/Replication-Only
Software engineering topic	Chapter and Section of SWEBOK (Abran et al. 2004)
Research Method	According to Easterbrook et al. (2007)
Unit of Analysis	Academics, Professionals, Artifacts
Confirmation of Original	Whether the replication confirmed the results of the original study

3.5 Quality Assessment

We assessed the quality of papers reporting replications (REP) using assessment criteria that addressed study design, conduct, analysis, and conclusions, as suggested in the guidelines developed by Kitchenham and Charters (2007). Our quality assessment was used only for classification purposes, not to exclude studies. For a full review and evaluation of our quality assessment (see Appendix B).

3.6 Synthesis of Results

The results from data extraction and quality assessment were integrated in spreadsheets, which were also used to generate graphs and tables. We did not perform meta-analysis or other form of meta-synthesis because it was out of the scope of our research questions. All statistics were calculated using MS Excel™ or SPSS Statistics version 17.0.

4 Results

In this section, we present the results of the mapping study providing answers to our research questions. We also provide information mappings to link the answers of the individual research questions.

4.1 Counting Replications and Original Studies

Due to the lack of standard terminology and the variability in paper structure, counting and classifying original studies and replications was not straightforward. Thus we preface our results with an explanation of replication counts.

Most papers collected in our investigation reported a single replication or an original study, which allowed for a direct one-to-one correspondence between replication and original study. However the mapping of the remaining papers was not self-evident. In our investigation, some papers reported more than one replication in a single paper and others reported one or more replications together with an original study in a single paper. In these cases, we counted each replication and original separately.

Our study resulted in 96 papers reporting 133 replications of 72 original studies. Thirty-six original studies were reported independently and the remaining 36 original studies were reported together with one or more replications. For classification purposes, we use the term *Replication-Only* reports to refer to the reports that did not include the original study. Replication-Only reports can contain internal and external replications. *Original-Included* reports refer to the reports that included the original and one or more replications in the same paper. Original-Included reports only report internal replications.

For several research questions we felt it was important to distinguish between external and internal replications, as discussed in Section 2.1. We operationalized an external replication as having no common authors between the replication and its original study; likewise an internal replication has one or more common authors. We opted-for this simple distinction because we needed our operational definition to follow directly from the data collected in the selected articles. We regard a researcher to be involved in an empirical study (original or replication) if he or she appears as one of the authors in any article reporting the

study. This distinction made our classifications objective resulting in no disagreements[5] between researchers during data extraction, according to the discussion presented in Section 3.4. We discuss the limitations of this operational definition and directions for future work in Section 6.1.

Appendix A lists the replications and original studies, together with classification information and the grouping of the sets of replications that refer to the same original study.

4.2 Answers to the Research Questions

Our results naturally fall into two groups. The first group of research questions (RQ1-RQ4) deals with the descriptive nature of individual replications and the second group of research questions (RQ5-RQ7) deals with interactions between replications.

4.2.1 Descriptive Information about Replications

In this section, we provide the answers to research questions RQ1 to RQ4, summarizing the descriptive information about the replications and original studies.

RQ1: What is the evolution in the number and type (internal and external) of replications over the years?

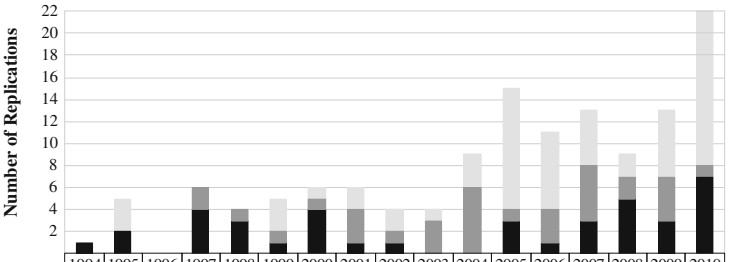
To better understand the history and development of replications in software engineering we tracked the relative growth between internal and external replications.

Figure 3 illustrates the replications in our review, grouped by publication year. Since our review's oldest replication was published in 1994, 1996 is the only year without replications. The growth in the number of replications, over the last 16 years, indicates that the software engineering research community has become aware of the importance of performing replications. However, the total number of replications (133) and replicated original studies (72) is small relative to the size of the fields as a whole and to the breadth of top software engineering (see RQ3). The original studies, in our review, were replicated on average 1.8 (133/72) times. We will show in the analysis of RQ5 that the vast majority of the original studies were only replicated once.

Figure 3 shows a clear change in the number of replications per year from 2004. In the period between 1994 and 2003, there was an average of 4 replications per year and this number grew more than threefold, to 13 replications per year in the period between 2004 and 2010. With no obvious explanation presented, we decided to investigate if there was a pattern relating to type of replication/report that would explain the threefold increase. In our review, 70 % (94/133) of the replications were internal (of which 64 % (60/94) were presented in Original-Included reports) and 30 % (39/133) were external.

By separating the number of external replications and splitting the internal replications into the Replication-Only and Original-Included internal replications (Fig. 3), we show that the increase in number of replications after 2004 was driven by the increase in internal replications, specifically by those presented in Original-Included reports. To further validate the increase in internal replications, we analyzed the average of the number of replications per year for the time period covered by this review (1994–2010) and sub-periods before and after 2004. Table 5 shows a significant difference in the means between external and internal replications. This difference was caused by the increase in the number of Original-Included

[5] We provide information about the inter-rater agreements in the specific sections below.



	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Internal– Original-Included reports	0	3	0	0	0	3	1	2	2	1	3	11	7	5	2	6	14
Internal– Replication-only reports	0	0	0	2	1	1	1	3	1	3	6	1	3	5	2	4	1
External	1	2	0	4	3	1	4	1	1	0	0	3	1	3	5	3	7
Total	1	5	0	6	4	5	6	6	4	4	9	15	11	13	9	13	22

Fig. 3 Temporal distribution of replications

reports, as discussed above. The behavior of external and Replication-Only internal replications was very similar with regards to the number of publications in the two sub-periods.

The higher rate of growth in the Original-Included internal replications over the external and Replication-Only internal replications could be indicative of the difficulty in performing and publishing isolated replications. Perhaps, authors chose instead to group one or more replications together with their original study into an Original-Included report. Another possible reason is that several empirical studies might be necessary before a reliable result is achieved and become ready for publication.

Considering the rate of growth of Original-Included reports of replications (since 2004), we believe Original-Included replications may have been a strategic advantage adopted by researchers to publish their work, amid the belief that publishing sole replications was difficult, as previously discussed by Kitchenham (2008). Given the increases in Table 5, this explicit or implicit strategy seems to be working. However, as we discussed further in RQ5, these results suggested a publication bias in the replication research when taken with other evidence.

RQ2: Which individuals and organizations are most active in replications?

Building off of the evolution of replications presented in RQ1, we wanted to gage the distribution of replications across the field of software engineering researchers and institutions. Based on our anecdotal evidence observed throughout the review process we postulated that there were hubs of institutions and researchers that were major contributors of replications.

In the 96 papers reporting replications, 194 distinct co-authors were found. Table 6 ranks the researchers more actively involved in performing replications. By separating external

Table 5 Average of replications per year

Period	Total		External		Internal					
					Total		Replication-Only		Original-Included	
	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ
1994–2003	4	2	2	1	2	2	1	1	1	1
2004–2010	13	4	3	2	10	3	3	2	7	4
1994–2010	8	6	2	2	6	5	2	2	4	4

Table 6 Most frequent researchers involved in replications

External Replications		Internal Replications	
Author Organization	External/Total	Author Organization	Internal/Total
James Miller University of Strathclyde	4/5	Mario Piattini University of Castilla-La Mancha	20/20
Emilia Mendes University of Auckland	3/5	Marcela Genero University of Castilla-La Mancha	16/16
Chris Lokan University of Auckland	3/4	Esperanza Manso University of Valladolid	9/9
Hong Mei Peking University	3/4	Conrado Aaron Visaggio University of Sannio	7/7
Per Runeson Lund University	3/4	Filippo Ricca University of Genova	7/7
Minghui Zhou Peking University	3/3	Gerardo Canfora University of Sannio	7/7
Murray Wood University of Strathclyde	3/3	Mariano Ceccato Fondazione Bruno Kessler	7/7
Marc Roper University of Strathclyde	3/3	Marco Torchiano Politecnico di Torino	7/7
Xiujuan Ma Peking University	3/3	Massimiliano Di Penta University of Sannio	7/7

and internal replications in Table 6, we observed two disjoint groups of researchers, one that performed mostly external replications and one that performed mostly internal replications. In fact, only 8 % (15/194) of the researchers performed both internal and external replications, while 66 % (128/194) performed only internal replications and the remaining 28 % (51/194) performed only external replications.

In the 96 papers reporting replications we found 97 distinct organizations (universities, research institutions, and companies) located in 19 different countries. The University of Sannio (Italy), the University of Castilla-La Mancha (Spain), Simula Research Laboratory (Norway), and the University of Valladolid (Spain) were the most active organizations in producing replications. All replications produced in these organizations were internal. Table 7 ranks the most active organizations in our review.

Consistent with our author analysis, the set of organizations that perform mostly external replications was also disjoint from the set of organizations that performed mostly internal replications.

Figure 4 shows the geographic distribution of the replication work, with the countries where two or more replications were found; Argentina, Denmark, Hong Kong, Ireland, and Turkey were excluded from the chart because they originated only one replication each.

According to Fig. 4, 91 % (122/133) of the replications originated from six countries and in this set only UK originated more external replications. Italy, Spain, and the USA concentrated nearly 60 % (79/133) of all replications. The percentage of internal replications originated from these three countries is 93 % (74/79), much larger than in the entire set of replications (70 %), which was discussed in RQ1.

Table 7 Most frequent organizations involved in replications

External Replications		Internal Replication	
Organization	External/Total	Organization	Internal/Total
Lund University	5/6	U. of Castilla-La Mancha	21/21
U. of Strathclyde	4/5	U. of Sannio	14/14
Middlesex U.	3/5	Simula Research Laboratory	9/9
U. of Auckland	3/5	U. of Valladolid	9/9
U. of Newcastle	3/4	U. of Maryland, College Park	8/8
Peking University	3/3	Fraunhofer IESE	7/8
U. of Bari	2/5	Fondazione Bruno Kessler	7/7
U. of Kaiserslautern	2/4	Politecnico di Torino	7/7
Brunel University	2/2	U. of Genova	7/7
California State U.	2/2	Fraunhofer CESE	6/6
CalTech	2/2	Ghent University	5/5
Miami University	2/2	U. degli Studi dell'Insubria	5/5

Therefore, there was a clear distinction between the groups of researchers, organizations, and countries with respect to the type of replication they perform. One could ask the questions of whether internal and external replications were distinct "styles of empirical research" or stem from different research cultures. We could not provide an answer for this question, from the data in this paper, but we believed that other research designs could be used in future research.

RQ3: What software engineering topics have been addressed by replications?

After establishing when, where, and by whom replications were published in our review, we wanted to identify the software engineering topics addressed by replications of empirical studies and to analyze coverage and concentration. We operationalized a software engineering topic as a chapter and a section of the Software Engineering Book of Knowledge (SWEBOK) (Abran et al. 2004).

We believe that replications of empirical studies are supposed to be relevant to the practice of software engineering and the SWEBOK is intended to structure the knowledge about software engineering topics needed by practitioners (Kitchenham et al. 2010).

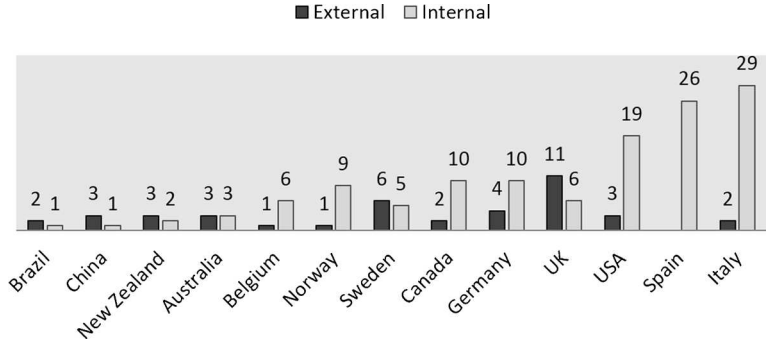


Fig. 4 Geographic distribution of the work on replication

Each replication in our review was assigned to the SWEBOK chapter that described the technology or treatment under investigation in the replication. For instance, a study investigating the "effectiveness of software testing strategies" was classified under Chapter 5: Software Testing. To ensure objectivity in our classifications, we performed a pilot classification of ten papers in a research meeting. The entire classification (excluding the pilot) was deemed to be 'good' with an inter-rater agreement of κ=0.82.

Table 8 relates the SWEBOK chapters (and 26 distinct sections) to the assigned replications. Software Requirements (Chapter 2) was assigned the largest number of replications (32) covering five of its seven sections. Thirty-one replications in this chapter were internal. Software Quality (Chapter 11) and Software Construction (Chapter 4) follow in second and third places respectively with over twenty replications each. Together, these three topics concentrate over 55 % (75/133) of the review. The least represented topics, with only one replication, were Software Configuration Management (Chapter 7), and Software Engineering Tools and Methods (Chapter 10). Software Design (Chapter 3), Software Configuration Management (Chapter 7) and Software Engineering Management (Chapter 8) showed the largest gaps in section coverage.

The concentration in certain chapters could be related to a similar concentration of the empirical research in software engineering, with the chapters with more empirical studies being the chapters with more replications and vice-versa. That is, as pointed out by one of the reviewers of an earlier version of this article, "We may see less replications for some chapters simply because the community performs fewer experiments about those chapters, so there are simply fewer candidate studies to replicate". However, we cannot test this hypothesis since there is no comprehensive review about empirical studies at large that uses the SWEBOK classification.

Figure 5 shows the distribution of the published replications plotted by SWEBOK chapter over time. Each circle represents the number of replications per year for a given chapter. External replications are shown in black and internal replications are shown in grey. This illustration allows us to understand that some chapters are covered by mostly internal

Table 8 Coverage of SWEBOK chapters and sections

SWEBOK Chapters	Number of Sections	Internal Replications		External Replications		Total		Gap in Section Coverage
		# Replic.	# Sect.	# Replic.	# Sect.	# Replic.	# Sect.	# Sect.
Ch. 2: Software Requirements	7	31	5	1	1	32	5	−2
Ch. 3: Software Design	6	5	2	0	0	5	2	−4
Ch. 4: Software Construction	3	10	1	11	2	21	2	−1
Ch. 5: Software Testing	5	2	1	5	3	7	4	−1
Ch. 6: Software Maintenance	4	8	2	7	3	15	3	−1
Ch. 7: Software Configuration Management	6	2	1	0	0	2	1	−5
Ch. 8: Software Engineering Management	6	12	2	5	1	17	2	−4
Ch. 9: Software Engineering Process	4	5	2	2	2	7	3	−1
Ch. 10: Software Engineering Tools and Methods	2	4	1	1	1	5	1	−1
Ch. 11: Software Quality	3	15	3	7	2	22	3	0
Total	46	94	20	39	15	133	26	

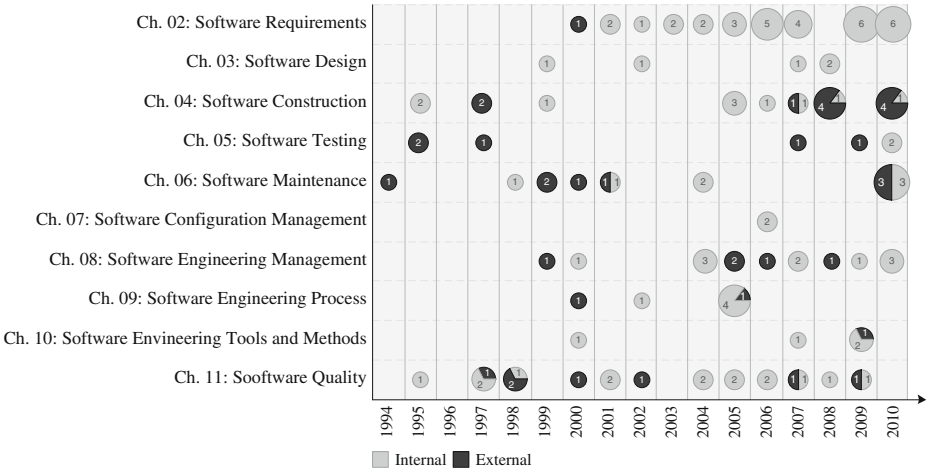


Fig. 5 Mapping SWEBOK chapters, year of publication, and type of replication

replications and other chapters are covered by mostly external replications. Chapter 02, 03, 07, 08, 09, and 10 are covered almost entirely by internal replications, Chapter 05 is mainly covered by external replications, and there is a balance in Chapters 04, 06, and 11. Figure 5 also illustrates the temporal nature of chapter coverage. Chapters 02, 04, 08, and 09 are primarily replicated after 2003, as opposed to the uniform distribution exhibited by other chapters.

RQ4: What research methods are being replicated?

To further understand what was replicated, we analyzed which research methods were present in replications. We classified the studies in our review using the classification of research methods presented by Easterbrook et al. (2007), which included the distinction between controlled experiments and quasi-experiments. Two team members assigned each replication with a research method. Upon merging our results, we achieved a good inter-rater agreement (κ=0.76), prior to resolving disagreements.

In our review, we found examples of controlled experiments, quasi-experiments, surveys, and case studies (Table 9). But the vast majority of replications were quasi-experiments (experiments in which "the subjects are not assigned randomly to the treatments" (Easterbrook et al. 2007)). Together, experiments and quasi-experiments comprise 88 % of the replications. We did not find any ethnography or action research studies in our review. This seems natural since these methods are less common in software engineering research than experiments. However, another explanation could be that these methods are harder to replicate since they involve more tacit knowledge that is difficult to code in the reports of the original studies.

Another interesting finding is that external replications have a much higher percentage of case studies than of experiments, when compared to the internal replications. From the answer to RQ2, we know that the researchers performing external replications are, in general, different from those performing internal replications. These two findings could suggest that there are two distinct research cultures in the software engineering community. One that prefers to perform internal replications of experiments and another that prefers to perform external replications of case studies. However, we cannot test this argument with our collected data. We would need to interview the researchers of these studies to gain relevant insights into their culture.

Table 9 Research methods

Method	Internal (*N*=94)	External (*N*=39)	Total (*N*=133)
Case Study	7 % (7/94)	21 % (8/39)	11 % (15/133)
Experiment	34 % (32/94)	10 % (4/39)	27 % (36/133)
Quasi-Experiment	57 % (54/94)	69 % (27/39)	61 % (81/133)
Survey	1 % (1/94)	0 % (0/39)	1 % (1/133)

We were surprised by how few Surveys were replicated. Each year new survey instruments are published and these instruments require repeated reliability assessments. We would expect that the research community would be more active in performing such assessments.

In addition to research method we also extracted the unit of analysis in each replication. The results show that the vast majority (59 %) of the replications used students and academic researchers (collectively called Academics) as units of analysis. Fifteen percent (20/133) of the studies used a mix of Academics and practitioners from industry (Professionals), and 12 % (16/133) used solely Professionals as participants. Artifacts, such as program code or specifications, were used as unit of analysis in 14 % (19/133) of the replications.

The concentration of replications using Academics is consistent with the results of Sjøberg et al. (2005), where 87 % percent of the reviewed experiments used students as subjects and 9 % used solely Professionals. Almqvist (2006) also showed a similar trend, with 70 % of the replications performed with students and 22 % solely with professionals. Although there are good reasons for conducting empirical studies with students (Carver et al. 2003; Ciolkowski et al. 2004), we agree with Sjøberg et al. (2005) in that "the low proportion of professionals used in software engineering experiments reduces experimental realism, which in turn may inhibit the understanding of industrial software processes and, consequently, technology transfer from the research community to industry". This argument is also valid for the replication work; thus strategies to increase replications with professional subjects are needed.

Table 10 presents the number of internal and external replications. Although there are proportionally more external replications using artifacts, we cannot observe significant distinctions between the two types of replications regarding their orientation towards using Academics or Professionals in the empirical studies.

Combining the results of Tables 9 and 10, Fig. 6 maps research methods, unit of analysis, and type of replication. As indicated in the previous tables, case studies with professionals and artifacts have been mostly performed by external replications.

4.2.2 Interactions between Replications

After understanding the descriptive nature of individual replications (RQ1-RQ4), we investigated the interactions between replications, starting with the grouping of replications that refer to the same original study.

RQ5: What sets of replications were found?

We identified the number of replications that replicated each original study. We then labeled each set of replications with a size (number of replications in the set), a type of replication (internal, external, or both) and a type of report (Replication-Only, Original-Included, or Both). The size of each replication set varied between one and ten (represented

Table 10 Types of units of analysis in replications

Unit of Analysis	Internal (N=94)	External (N=39)	Total (N=133)
Academics and Professionals	20 % (19/94)	3 % (1/39)	15 % (20/133)
Academics	61 % (57/94)	54 % (21/39)	59 % (78/133)
Professionals	11 % (10/94)	15 % (6/39)	12 % (16/133)
Artifacts	9 % (8/94)	28 % (11/39)	14 % (19/133)

as S1 to S10 respectively). See Appendix A for a description of the sets of replications and their original studies.

Table 11 shows the 72 sets of replications (based on the 72 original studies). Each row in the table represents the counts of replication sets with a specific size. For example, row four includes the counts for the two sets of replications with four replications of an original study in each set. Table 11 is missing rows six through eight because no sets were found of these sizes. The majority of the sets (61 %) have only one replication or a size of one. Only 21 % of the sets are composed of three or more replications.

Broken down by type of replication, 70 % of the sets are composed solely of internal replications, 25 % of the sets are composed only of external replications, and only 5 % of the sets are mixed with both internal and external replications. These numbers are consistent with the distribution of internal and external replications (discussed in RQ1). Replication work is mainly partitioned into studies that have been replicated only internally or only externally, with a very small cross-section that have been replicated both internally and

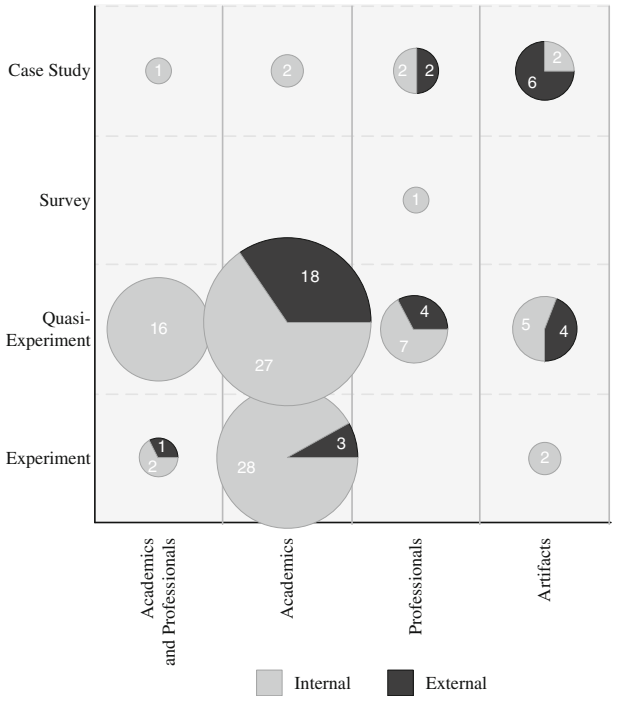


Fig. 6 Map of research method, unit of analysis, and type of replication

Table 11 Summary of sets of replications

Set Id.	Rep./Set	Internal				External (N=18)	Both Internal and External (N=4)	Total (N=72)
		Internal Original-Included (N=31)	Internal Replication-Only (N=17)	Internal both types of Reports (N=2)	Internal Total (N=50)			
S1	1	48 % (15/31)	94 % (16/17)		62 % (31/50)	72 % (13/18)		61 % (44/72)
S2	2	29 % (9/31)		50 % (1/2)	20 % (10/50)	6 % (1/18)	25 % (1/4)	17 % (12/72)
S3	3	19 % (6/31)			12 % (6/50)	22 % (4/18)	25 % (1/4)	15 % (11/72)
S4	4	3 % (1/31)		50 % (1/2)	4 % (2/50)			3 % (2/72)
S5	5						25 % (1/4)	1 % (1/72)
S9	9		6 % (1/17)		2 % (1/50)			1 % (1/72)
S10	10						25 % (1/4)	1 % (1/72)

externally. These results are consistent with the findings of Almqvist (2006), where 55 % (11/20) of the sets were composed solely of internal replications, 30 % (6/20) were composed of external replications, and 15 % (3/20) were mixed. Possible causes and implications of this partitioning are discussed below in our analysis of RQ6 and RQ7, and in Section 5.

Although the concentration on small sets (S1 and S2) happens for all types of replications, there are proportionally more S1 sets with only external replications (72 %) than with only internal replications (62 %). This proportional concentration changes when the set of internal replications is divided into Replication-Only and Original-Included reports. In this case, 94 % of the sets with only Replication-Only replications are composed of only one replication, whereas only 48 % of sets are composed only of Original-Included replications.

One possible explanation for this is the publication bias discussed in RQ1: it is easier to publish a paper in which several replications are bundled with the original study, than to publish a single replication, internal or external, in a paper. If this bias is true, the tendency would be to bundle as many replications as possible in a single paper. Our review shows evidence of this because 58 % of the sets with two or more replications are reported in Original-Included papers. Complementary, it is possible that several repetitions of a study are necessary before sufficient results are produced for publication, also explained the grouping of several replications in one report.

RQ6: Do the replications confirm the results of the original studies?

After describing consolidated information about the replications and understanding the groups of replications with respect to their original studies, we investigated whether the replications confirmed the results of the original studies. At this point, we also wanted to explore patterns in replication results related to internal or external replications.

We defined three categories of Confirmation to categorize the replications: Yes, the results are confirmatory; No, the results are non-confirmatory; and Partial, some sub-results are confirmatory and others are non-confirmatory. In order to fairly assign a confirmation category, we used the conclusions about confirmation as reported by the authors whenever possible. When the authors were not explicit about the relationships between the results of their replications and the original study, we looked at the study results of both and performed our own analysis. This assessment resulted in a borderline inter-rater agreement (κ=0.65) before disagreements were resolved, which we discussed in Section 5.4.

Four cases of Original-Included internal replications were excluded from this analysis because their replications did not test their original results. In these cases, the authors made clear distinctions between the replications and the original studies, but combined the results of both in an effort to synthesize new hypotheses. Therefore, the total number of replications in RQ6 is 129. Table 12 shows a clear difference between the results of internal and external replications. While 82 % of internal replications confirmed and 9 % partially confirmed the original results, 46 % of the external replications did not confirm the original results. This divide increases when we consider the set of Original-Included internal replications, in which 89 % of the replications fully confirmed the original results.

Figure 7 visualizes this difference between internal and external replications, and also compares our results with the findings of Almqvist (2006) and Sjøberg et al. (2005). This data shows a clear tendency for internal replications to report confirmatory results. While the opposite tendency (reporting non-confirmatory results), is not as strong in the set of external replications. But, it is undeniable that the internal and external replications differ with respect to this factor. These trends are consistent with the work of Almqvist (2006) and Sjøberg et al. (2005).

Table 12 Confirmation of the original results

Replication Type	Confirmation			Total
	No	Partial	Yes	
External	46 % (18/39)	28 % (11/39)	26 % (10/39)	39
Internal	9 % (8/90)	9 % (8/90)	82 % (74/90)	90
Original-Included	5 % (3/56)	5 % (3/56)	89 % (50/56)	56
Replication-Only	15 % (5/34)	15 % (5/34)	71 % (24/34)	34
TOTAL	20 % (26/129)	15 % (19/129)	65 % (84/129)	129

Three (inter-related) types of publication bias could explain the confirmatory tendency of internal replications: first, researchers would not even try to publish non-confirmatory results of their own research; second, even in the case that there was an attempt to publish non-confirmatory results, they were not accepted, because "negative" results are perceived as not worthy of being published; and third, although it is difficult to publish non-confirmatory results, it would be easier to publish non-confirmation of the work of others (external replications) than of one's own work.

Another explanation could be that researchers replicating their own work were unintentionally biasing the result towards confirmation during the study development or unintentionally being more lenient with the results during analysis. Finally, internal replications would naturally be closer to the original study methodology (less variations) than external replications, increasing the likelihood of confirmatory results, as noted by Amiqvist (2006) and Sjøberg et al. (2005).

The non-confirmatory tendency of external replications could be explained by at least three reasons: first, intentional or unintentional variations in experimental design are more likely to happen in external replications, and these variations may lead to non-confirmatory results; second, differences in experimental context (cultural, social, and organizational factors) could also have affected the results even if the design was followed as close as possible (as reported in REP118, REP119, and REP120); and third, researchers replicating the work of others would not have access to essential tacit information about the original study (Shull et al. 2002) and could (inadvertently) induce the non-confirmatory results either during study development or results analysis.

Independent of which reason prevails in a particular case, it seems likely that internal replications will have a tendency towards confirmation and external replications will have a

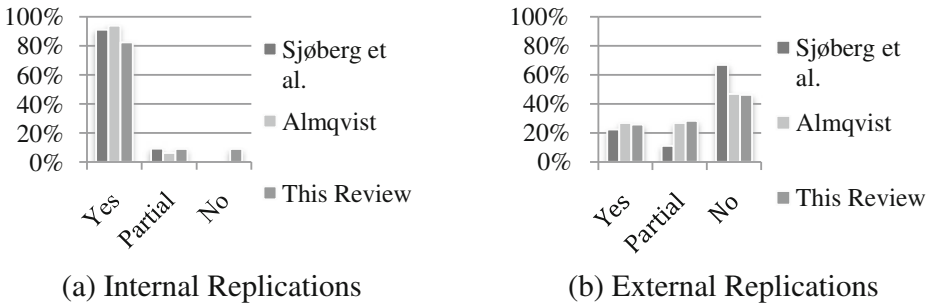


Fig. 7 Confirmation of original results in internal and external replications

tendency towards non-confirmation. Therefore, confirmatory internal replications and non-confirmatory external replications are not that helpful in isolation, unless they are part of a set composed of a mix of internal and external replications with in which comparison of variations and opposite results can be performed. However, the bulk of the replications analyzed in our review are isolated confirmatory internal or non-confirmatory external, and do not contribute to substantial knowledge building in our field.

Sets composed of a mix of replication type will lead to more scientific advances than sets consisting of solely internal or external replications. We assert that a mixed set of replications (having both internal and external replications) would have allowed researchers to analyze results affecting variations in experimentation. This analysis would have contributed to a better understanding of the conditions under which results hold or break (Juristo and Vegas 2009).

However, considering that only four sets of replications, in Table 11, were composed of internal and external replications, this type of knowledge building was very limited. Therefore, having only internal or external replications of a given study, as demonstrated in the answer of RQ5, is a limitation that must be addressed in the empirical research in software engineering.

RQ7: What is the elapsed time between the replications and corresponding original studies?

Finally, we analyzed the temporal distance between the development of the original study and its replications. We conducted this analysis with the goal of understanding how long it takes for the information about a replication and, consequently, its results compared to the original study, to be publicly available. Considering the fast pace of technological advances in software engineering and the recent findings that "the aging of the computing literature is not atypical compared with other scientific research disciplines" (Sjøberg 2010), this information would inform discussion about the impacts of empirical studies both in research and in practice.

One possible way to operationalize the elapsed time between a replication and its original study would be to get the elapsed time in months or years between the end date of the original study and the end date of its replication. However, information about when the studies were performed is not provided in any reliable form in most papers, making this approach infeasible. An alternative approach is to consider the elapsed time in years between the publication dates of the papers reporting the original study and its replications. Although this operationalization would not reflect the exact temporal distance between the studies, it does provide useful information regarding how long it takes for the external public (researchers and practitioners not involved in the development of the studies) to access information about the replications. Therefore, we chose this operationalization because of its viability and usefulness. More precisely, we used the first publication date of the paper as it is printed in the paper. Therefore, first available date in journals and the conference date in conferences.

Table 13 shows the mean, standard deviation, and the median distance between the publication of the original study and replications. The mean distance between replication and original is just above 2 years. Internal replications are published less than 1 year after the publication of the original result. The decreased delay in publications of internal replications could be explained by two reasons: first, since the same team of researchers performed the original and replication, it is possible that the two studies were conducted (and published) in parallel or very close together; and second, the full set of papers reporting internal replications include the Original-Included internal replications, in which original and replication

Table 13 Mean and median of time elapsed between replication and original in years

Papers Reporting Replication	Mean	σ	Median
All Papers (N=96)	2.1	2.8	1.0
Internal Replications (N=65)	0.8	1.7	0.0
External Replications (N=31)	4.8	2.8	4.0

were reported in the same paper. When the 36 Original-Included reports were excluded from the full set, the mean time distance increases to 3.4 years (σ=2.9).

Considering only the external replications, the mean temporal distance increases to almost 5 years. Since internal and external replications were almost disjoint sets with respect to original studies (see RQ5), for almost 50 % of the original studies (31/72), it took an average 5 years for a replication to be published.

We illustrate the elapsed time between publishing original studies and replication in Figs. 8 and 9. Figure 9 displays only the temporal distance between an original study and its first replication, whereas Fig. 8 displays all replications. In both figures, sub-figure (a) shows the number of replications for each time distance to its corresponding original study, and (b) shows the percentage of internal and external replications for each time interval (based on the totals in (a)).

It is clear from both charts, in Fig. 8, that internal replications are published with less time elapsed than external replications and that as time passes the number of internal replications drop substantially. From Fig. 9, we observed that the first internal publications are published in no more than 4 years after the publication of the original study, and the number of internal replications published over 2 years after the original decreases drastically. We observed the first external replications appeared 2 or more years after the original.

When combined with our result that internal and external replications form disjoint sets (RQ5), these results (see Table 13, Figs. 8 and 9) indicate that, in general, internal replications are published close to the originals, and after a certain period of time, no more internal replications are published (there are only two internal replications published after 4 years of the original). Thus, after a research group published a set of internal replications in a short period of time, the experiment was not internally replicated again. We believe that research groups moving to new research agendas after completion of their internal replications could explain this phenomenon.

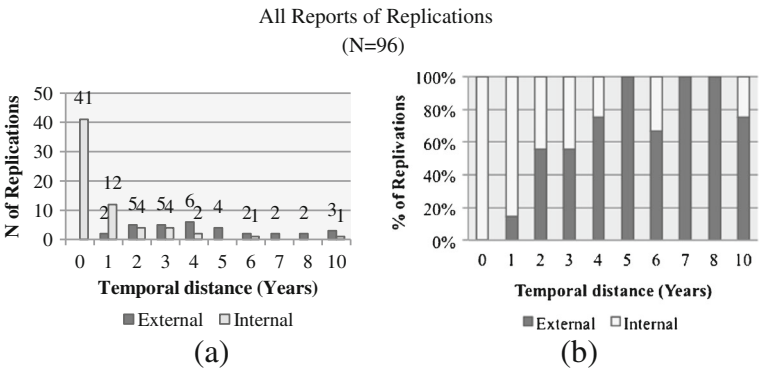


Fig. 8 Elapsed time between the publication of original and replications

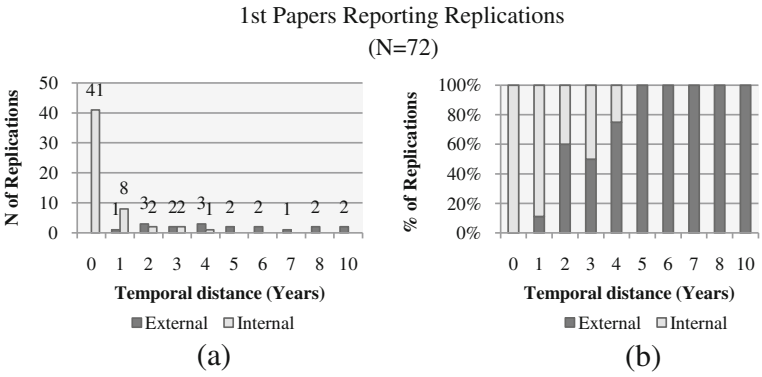


Fig. 9 Elapsed time between original and the 1st replication

External replications are more evenly spread, but more than 50 % of them took more than 5 years to be published. This seems natural because the research groups involved were distinct. In most cases, the original study's published report was the only source researchers used to design their external replication, so design could have only begun after the publication and discovery of the original report.

Because external replications tended to have non-confirmatory results (RQ6) and over 50 % or them were produced after the original researchers had stopped working on the problem, it is unlikely that the non-confirmatory results of this set of replications served as feedback to the original researchers, either to challenge their results or to improve their investigations. Therefore, over 50 % of external replications, which are likely to present non-confirmatory results, were not readdressed by an internal replication. Empirical software engineering research was not using the potential benefits of mixed types of replications, with both external and internal replications happening early in the research cycle.

5 Discussion

Our goal, in this review of empirical software engineering research, is to plot the general landscape of the body of work on replications. In this section, we discuss our results and their implications for software engineering research and practice.

5.1 Strengths and Limitations of the Replication Work

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

- The number of replications is increasing. In 2004, the rate intensified to 13 replications per year.
- Empirical replications cover a diverse set of topics in software engineering, although certain topics are found more concentrated.
- A diverse set of researchers at research organizations located in several countries is executing replications. This identification of researchers/organizations, interested in performing replications, could foster collaborations and support the growth of replications.

However, the overall results raise some concerns due to important limitations:

- Replication reports do not provide complete information about their original studies, making it difficult to relate the replication results to the original studies. Consequently, our community cannot explore under which conditions empirical results hold and break.
- The absolute number of replications (133) and replicated original studies (72) is small and the rate of growth is not enough to keep the pace of the growth of the empirical studies in general.
- The absolute and relative number of external replications is very small (39), and since 1994 the growth rate of three new replications per year is more or less constant.
- The majority of replication sets are small (1 replication per set). An average of 1.8 replications have been performed per original study.
- Only four sets of replications are composed of both external and internal replications. Internal and external replications differ, which limits our ability to generate knowledge from the sets of replications.

These strengths and limitations have implications for researcher and practitioners, which we discuss in the next section.

5.2 Implications for Research

We believe that our results have important implications for empirical research in software engineering. We summarize the six questions our implications raise, which require attention from the research community.

Question 1: What should be considered replication?

We found at least three situations where it is not clear whether a given study should be considered a replication.

Multiple case studies: Yin (2009) argued that in multiple-case designs, each case study should be seen as one (internal) replication, which was consistent with our definition of replication. We analyzed reports of case studies and counted each case as unique replications. However, some reports did not clearly describe the study design, so it was not possible to determine whether the study was a multiple-case design or an embedded single-case design (with multiple cases as the unit of analysis). We decided to count all case studies as replications. This decision may have artificially increased the number of replications in our study. Researchers reporting case studies should precisely describe their study design to allow for a clear-cut identification of actual replications.

Longitudinal surveys: In the manual search of journal articles, we found a series of four surveys performed with the same population over a period of 10 years (Davidsen and Krogstie 2010; Holgeid et al. 2000; Krogstie and Sølvberg 1994; Krogstie et al. 2006). It was not clear whether the three subsequent surveys acted as replications of the first one or part of one ongoing study (e.g. a cohort or longitudinal survey (Kitchenham and Pfleeger 2007)). We consulted survey and replication specialists and received conflicting opinions. In the RESER'2011 paper (da Silva et al. 2011b) we considered these studies as replications because the initial study was not positioned as part of a longitudinal study (Krogstie and Sølvberg 1994). However, other participants at RESER'2011 convinced us that such studies were not a set of replications. Therefore, we removed them from our analysis for this article.

Change in treatment: Several reports of experiments that compared two or more treatments considered each treatment to be replication. It some cases, it was evident that the researchers were describing a single experiment with multiple treatment assignments for the subjects. We did not consider these cases to be replications. However, in some cases this was not clear, which may have resulted in inappropriate selection. We may have considered two distinct experiments as a replication and an original, instead of treating the second experiment as a variation of the first (original) study.

Question 2: What should be considered a good report of a replication?

At RESER'2010, Carver (2010) discussed guidelines for reporting replications. We incorporated some of his suggestions in our quality assessment criteria. As a result, the reports in our analysis scored considerably lower in the replication-specific criteria than on the generic criteria. One possible reason for these lower score is page limits. The replication-specific criteria require a description of various aspects of the original study, which consume space in the report. Faced with space restrictions, researchers may have opted not to include this information in the reports. Alternatively, the replication authors were missing complete information about the original studies, and reported only what they could find.

It become clear during data extraction and analysis that the poor quality of the information about the original study impacted the quality of the replication analysis and, consequently, the potential use of its results both in research and practice. Furthermore, the lack of explicit information about variations between replications and their original studies made it more difficult or impossible to compare replications, which resulted in more disagreements between researchers when evaluating RQ6. Ultimately, the lack of reporting standards, for replications, impacted our ability to use and generate knowledge from these sets of replications.

Question 3: Why do external and internal replications differ with respect to the various factors analyzed in this review and do these differences matter (if so how)?

The answers for most research questions presented clear evidence that internal and external replications differ with respect to four factors: the researchers who performed the replications, the confirmation of original results, the grouping into sets of replications, and the time elapsed since the publication of the original study. While some of these differences seem obvious, they are also indicative of several types of limitations and publication biases. We believe these limitations threaten the validity of replications in general and have not been explicitly addressed in any study. In order to improve the quality of future replications, our community should investigate why external and internal replications are so different.

Question 4: How to increase the number of replications?

Researchers used corroborating internal replications to improve the reliability of their work. Studies are more likely to be published if their results are deemed reliable through reproduction.

As discussed in RQ1, the number of Original-Included reports of replications published since 2004 has increased over other types of reports. We expect this trend to continue, as the production of internal replications enters more empirical software engineering research agendas.

This motivation does not apply to external replications. In fact, we do not know what motivates researchers to perform external replications. External replications increase the range of results related to a given original study, so we believe it is important to understand the motivations behind external replications in order to stimulate their publication.

Question 5: How do we build more sets with a mix of internal and external replications?

As discussed in RQ6, sets containing a mix of internal and external replications provide increased knowledge over homogeneous sets of studies. However, we found researchers prefer to perform either internal or external replications (RQ2), but not both. Furthermore, researchers who performed external replications did not replicate studies with preexisting internal replications. Therefore, it is unlikely that a large number of mixed sets will be produced.

We found that internal replications tend to confirm results and were published shortly after the original study (or even in the same paper). This partially explained the low number of mixed sets because after internal replications were published (confirming the original study), there was no motivation to perform external replications. We also found that over half of the external replications were preformed a long time after the publication of the original studies and tended to report non-confirmatory results. This also explained the low number of mixed sets because, for over 50 % of the external replications, by the time the non-confirmatory results were published, the original researchers would have shifted their focus to a new area and would lack the motivation to refute the non-confirmatory results with an internal replication.

Our inspection of the four sets with mixed types of replications (REP134, REP132, ORI092, and REP053) provided insight into mixed sets. By contrasting them with the homogeneous sets, we hoped to find ways to stimulate mixed sets. In all four cases, the external replications were performed immediately after the internal ones, and in three cases the external replications were followed by more internal replications. The time elapsed between the first and the last replication was 4 years for REP134 (which was the largest set of replications), and 3 years for REP132 and REP053. These were below the average elapsed time for external replications. The short time of the entire set seems to be an important factor for the existence of mixed sets.

The time elapsed between REP098 and ORI092 was 10 years (one of the largest in our review and the largest of an internal replication). This seems to contradict the above argument. However, upon further investigation of REP098, we found out that it was the result of a collaborative project between the authors of ORI092 and an external group of researchers from a different country, “with the goal of effectively replicating software engineering experiments”. The original experiment was used as one of the cases in this project performed by this external group of researchers in collaboration with the original researchers. This explained the long elapsed time and also the existence of an internal replication 10 years after the original (an outlier in our set of internal replications). In this case, a collaborative project focused on replication stimulated the development of a late internal replication and the existence of a set with mixed types of replications. All this discussion leads to our final question.

Question 6: How to decrease the temporal distance between external replications and originals?

RQ7 showed that, in general, internal replications were published recently after the originals, and after a period of time, no more internal replications were published. We found that this trend happened 4 years after of the original study was published. External replications were more evenly spread, with more than half published after 4 years of the original study. Thus, it is unlikely that this set of replications, which is likely to be non-confirmatory (RQ6), would have served as feedback to the original researchers and would not be readdressed by an internal replication. Therefore, it would be beneficial, for software engineering research, to shorten the temporal distance between original studies and their external replications.

One reason for the long elapsed time, already discussed above, was that the design of the external replication was likely to be based on the paper reporting the original study and, therefore, the development of the replication can have only started after the original was published. Furthermore, the original study reports were often missing complete design information and researchers performing the replications had to complement this information either by consulting the original researchers or by trial and error.

In any case, we believe one way of reducing this temporal distance is to make public the design of the empirical studies as early as possible. One potential venue for this could be to include tracks in conferences for the presentation of empirical studies in the design phase. In this way, researchers could choose to work on an external replication while the original experiment or one of its internal replications is being developed. An increased level of communication between researchers could help researchers correct limitations in their study designs before publication. How much collaboration is useful is a matter of debate (Vegas et al. 2006), since too much communication could mischaracterize the study as an external replication.

5.3 Implication for Practice

Empirical studies are important to increase our confidence on the effectiveness of solutions proposed to deal with practical problems in software engineering. Therefore, replications of these studies would be important in corroborating study results and would ultimately be important for the practical adoption of proposed solutions. The results of our review have at least two implications for practitioners. First, both internal and external replications tend to be biased in their confirmation of results. Therefore, it is important to know which type of replication is being reported and how to interpret its results with respect to confirmation, before deciding whether to use the proposed solution. Furthermore, practitioners would gain more knowledge by looking at sets with a mix of internal and external replications, but such cases are rare, representing a limitation for practitioners as well as for researchers, as discussed above.

Second, “context is golden” when interpreting the results of empirical studies and in particular of the replications. It is important to understand the context of a study or replication before concluding on the transferability of its results to another context. To guide this contextual interpretation, it is important to look at the variations between replications and their original study. However, as discussed in the quality assessment, reports of replications, in particular the Replication-Only reports, provide poor descriptions of the

replication-specific criteria, including the context of the original study. This context specific interpretation is left to the interested reader, who will have to explicitly compare all studies in a set, to form his or her opinion on the variations in context.

5.4 Limitations of this Review

The most common limitations in a systematic review are limited coverage, possible biases introduced in the selection process, and inaccuracies during data extraction and quality assessment. These were also the main limitations in this review. We tried to minimize these limitations by using well-established guidelines for our research protocol. Three co-authors of this mapping study previously participated in a tertiary study (da Silva et al. 2011a), which minimized these limitations further because they had previous experience performing and analyzing systematic reviews, and understood the potential pit falls of the review process.

The combination of automatic search (in multiple search engines) and manual search (on relevant publications) improved the coverage of the review. We searched over 16,000 articles and did not impose any time constraint in the automatic search. As discussed above, there was no standard terminology for replication work in software engineering. It was possible that we missed replication reports that did not use "replication" or any of our synonyms. It was also possible that we missed papers because they were not indexed in the search engines at the time we performed our automatic search. The use of manual search reduced this possibility and we are confident we achieved a high coverage. Nevertheless, coverage is an inherent threat to validity in any review (systematic or not). In general, it is not possible to achieve 100 % coverage and it is very difficult to confidently estimate the level of coverage a review achieves.

Based on the recommendation of Kitchenham and Charters (2007), we used a multistage selection process where each pair of researchers recorded (at each stage) their inclusion and exclusion criteria for each study. When agreement was not reached, within each pair, a third independent researcher assessed the disagreement. When conflict persisted, the final decisions, for inclusion or exclusion of a study, were made through a consensus meeting. As a rule, we decided to postpone exclusion of a potentially relevant study as much as possible to avoid excluding relevant papers.

Data extraction and quality assessments were also performed by at least two researchers, and conflicts were solved either by a third party or in consensus meetings. The reviewed studies varied with respect to the use of terminology and reporting standards. This made some data extraction more difficult and error prone, in particular in the quality assessment and in RQ6. In our group meetings, we conducted pilot extractions and evaluations, solved possible sources of ambiguity in the extraction and assessment, and tested our inter-rater agreements using Kappa coefficients. The quality assessment and the extraction of RQ6 provided a Kappa coefficient just below 0.7 (0.69 and 0.65 respectively) before disagreements were solved. We believe this is not a strong limitation since we thoroughly reviewed all disagreements before data analysis was performed.

6 Conclusions

In this article, we presented a systematic mapping study of the empirical software engineering literature with regards to replications of empirical studies. From over 16,000 papers found during our extensive manual and automatic searches, we selected

173 relevant papers and divided them in three sets: papers describing replications (96), papers presenting original studies (36), and papers describing conceptual or theoretical work about replication (41).

In this article, we used a subset of 132 papers, which presented replications and original studies, to answer our research questions. These papers reported 133 replications of 72 distinct original studies.

We extended the previous literature by including 60 papers that have not been reviewed by Sjøberg et al. (2005); Almqvist (2006), or Carver (2010). This article also extended the preliminary results of our RESER'2011 workshop paper (da Silva et al. 2011b) by adding two new research questions, providing more information about each question, building a more comprehensive mapping of the information, and presenting a deeper analysis and discussion about the results.

The total number of replications is increasing. Specifically, the average number of replications produced each year is increasing from four (1994–2003) to 13 (2004 and 2010). The number of internal replications is growing faster than the external replications. This indicates that although the research community is more aware of the importance of performing replications, performing replications of studies by others is not an established practice yet.

6.1 Future Work

In generating this mapping study, we made a substantial effort to select papers from the large volume generated from our automatic search results. We believe that future extensions and updates will benefit from our efforts, because these extensions will only have to apply the search procedure on the publications after 2011. We expect to perform continuous updates to this mapping study, using the same research protocol described in this article.

From our findings, we believe that several aspects of replication work require further investigation. Using data extracted from this mapping study, we are currently working on three topics:

- Analysis of goals and motivation: We plan to study the goals and motivations for performing replications with the hope of fostering more replications. In particular, we would like to understand how the goals and motivations of those performing internal and external replications differ. This could explain the existence of two distinct groups of researchers in our review. In this case, the data analysis extracted from this paper must be complemented by data collected directly from the researchers. We believe that interviews, followed by qualitative analysis, are a suitable approach for this investigation.
- Analysis of variations: We partially collected data with regards to the variations between replications and their original study. As a future study, we will comprehensively extract and systematically analyze sets of replications, to study the variations between replications and their impact. We hope to identify patterns that would indicate types or categories of variation, such as those proposed by Basili et al. (1999). In particular, we are interested in investigating how intentional and non-intentional variations, (unavoidable) human subject variations (Lung et al. 2008), and cultural and contextual variations cause discrepancies in replications. Finally, we hope to contribute to Juristo and Vegas' (2009) discussion of the problem of building knowledge from sets of replications.
- Lessons learned: Some authors reported their *lessons learned* from developing replications in their papers. We plan to build a structured catalog of these experiences, to guide and assist future replications.

In addition to these three topics, we believe that guidelines, for performing and reporting replications, as well as conducting empirical studies in general, are still needed. Carver (2010) offers a good starting point. We believe that the results of this mapping study with an added catalog of lessons learned can contribute to the construction of guidelines.

Another topic for future work is to improve our operational definition of internal and external replications. In this study, we used paper authorship as a proxy for participation in the study. Although our definition was non-ambiguous and consistent with previous related work it has limitations and other operational definitions could provide complementary insights. However, it is not currently possible to ascertain the level of participation of an author from the contents of their papers. Therefore, other research designs are needed to use other operational definitions. For example, a survey of the authors of the papers could investigate their levels of participation.

As part of the interpretation of our findings, we stated several hypotheses and propositions, which seem to be supported by our data. However, we have not tested these hypotheses because they require other study designs. We believe that testing these hypotheses through case studies or survey research would add to our field.

Finally, in our search and selection processes, we found 41 papers addressing conceptual and practical issues related to software engineering replications. We intend to analyze and synthesize the results of these papers to complement our understanding of replications in empirical software engineering.

6.2 Final Considerations

In this article, we discussed questions that we believe should be addressed by the empirical software engineering research community. We need clearer definitions of what constitutes a replication and better guidelines for performing and reporting replications. We must increase the number of replications, the number of sets of replications, and sets with mixed types of replications. We believe that it is important to understand, what motivates researchers to perform internal and external replications. Finally, we should stimulate shorter temporal distances between external replications and their original studies, because this could contribute to the development of more mixed sets.

These improvements require further collaboration between replication investigators, as well as systematically structuring information about experiments and replications, in persistent repositories. We believe this mapping study is one step to achieve these improvements.

Acknowledgments Fabio Q. B. da Silva holds a research grant from the Brazilian National Research Council (CNPq), process #314523/2009-0. This article was written while Prof. Fabio Silva was in a sabbatical leave at the University of Toronto, receiving a CAPES research grant process # 6441/10-6. A. César C. França is a doctoral student at the Center of Informatics of the Federal University of Pernambuco where he receives a scholarship from the Brazilian National Research Council (CNPq), process #141156/2010-4. We would like to thank Prof. Steve Easterbrook, Jonathan Lung, and Elizabeth Patitsas for many discussions, comments, and criticisms that lead to important improvements in the content and structure of this article. We also thank Prof. André Santos, Rodrigo Lopes, João Paulo Oliveira, and Leonardo Oliveira, for their participation in the earlier version of this study published at RESER'2011. Finally, we are grateful for the partial support of the Samsung Institute for Development of Informatics (Samsung/SIDI) for this research.

Appendix A—Selected Primary Studies

In this Appendix, we describe the selected papers that report replications and primary studies, which form the dataset of our mapping study. In Section A.1, we describe the two types of

papers reporting replications: Original-Included reports and Replication-Only reports. In Section A.2, we present the complete list of references of these papers. In Section A.3, we present the complete reference list of the papers reporting solely original studies.

A.1 Selected Papers Reporting Replications

In this section we presente a summary of the papers reporting replications in Section A.1.1. We then provide details about the papers that compose the sets of replications in Section A.1.2.

A.1.1 Descriptive Information of Papers Reporting Replications

In Table 14, we presente an overview of the 96 papers reporting the 133 replications. This table is ordered by the year of publication of the paper reporting the replication. The quality score is discussed in detail in Appendix B. The column Original Ref. presents the reference to the paper reporting the original study. When this column is empty, it indicates that the paper reporting the replication also reported the original study, i.e., it is an Original-Included report of internal replications.

Table 14 Overview of the replication papers

Replication Ref.	Year	Quality Score	Replication Type	Report Type	SWEBOK Chapter	Research Method	Original Ref.
REP073	1994	72 %	External	Replication-Only	6	Quasi-Experiment	ORI073
REP053	1995	94 %	Internal	Original-Included	11	Quasi-Experiment	
REP085	1995	78 %	External	Replication-Only	5	Experiment	ORI085
REP089	1995	72 %	Internal	Original-Included	4	Experiment	
REP047	1997	78 %	External	Replication-Only	11	Experiment	REP053
REP048	1997	69 %	External	Replication-Only	5	Quasi-Experiment	ORI085
REP086	1997	75 %	External	Replication-Only	4	Quasi-Experiment	ORI086
REP087	1997	72 %	Internal	Replication-Only	11	Quasi-Experiment	ORI087
REP113	1997	78 %	Internal	Replication-Only	11	Experiment	ORI092
REP050	1998	91 %	External	Replication-Only	11	Experiment	REP053
REP088	1998	88 %	External	Replication-Only	11	Quasi-Experiment	REP053
REP112	1998	81 %	Internal	Replication-Only	11	Quasi-Experiment	REP053
REP058	1998	44 %	External	Replication-Only	6	Quasi-Experiment	ORI034
REP045	1999	50 %	External	Replication-Only	8	Quasi-Experiment	ORI045
REP052	1999	100 %	Internal	Original-Included	6	Quasi-Experiment	
REP055	1999	66 %	Internal	Replication-Only	4	Case Study	ORI055
REP083	1999	100 %	Internal	Original-Included	3	Quasi-Experiment	
REP046	2000	63 %	Internal	Replication-Only	8	Quasi-Experiment	ORI046
REP049	2000	100 %	Internal	Original-Included	10	Experiment	
REP060	2000	78 %	External	Replication-Only	2	Quasi-Experiment	ORI060
REP090	2000	59 %	External	Replication-Only	6	Quasi-Experiment	ORI034
REP092	2000	44 %	External	Replication-Only	11	Quasi-Experiment	ORI092
REP106	2000	81 %	External	Replication-Only	9	Quasi-Experiment	ORI106
REP033	2001	88 %	Internal	Replication-Only	6	Quasi-Experiment	ORI033
REP034	2001	81 %	External	Replication-Only	6	Quasi-Experiment	ORI034

Table 14 (continued)

Replication Ref.	Year	Quality Score	Replication Type	Report Type	SWEBOK Chapter	Research Method	Original Ref.
REP051	2001	94 %	Internal	Original-Included	11	Quasi-Experiment	
REP122	2001	47 %	Internal	Replication-Only	2	Experiment	ORI122
REP132	2002	72 %	Internal	Original-Included	9	Case Study	
REP091	2002	89 %	Internal	Original-Included	3	Quasi-Experiment	
REP123	2002	63 %	Internal	Replication-Only	2	Experiment	ORI122
REP043	2002	72 %	External	Replication-Only	11	Quasi-Experiment	ORI043
REP036	2003	78 %	Internal	Replication-Only	8	Quasi-Experiment	ORI036
REP040	2003	89 %	Internal	Original-Included	8	Survey	
REP124	2003	100 %	Internal	Replication-Only	2	Experiment	ORI122
REP032	2004	91 %	Internal	Replication-Only	6	Quasi-Experiment	ORI032
REP038	2004	94 %	Internal	Replication-Only	11	Experiment	ORI038
REP041	2004	75 %	Internal	Replication-Only	11	Quasi-Experiment	ORI041
REP082	2004	67 %	Internal	Original-Included	8	Quasi-Experiment	
REP093	2004	56 %	Internal	Replication-Only	6	Experiment	ORI093
REP125	2004	82 %	Internal	Replication-Only	2	Experiment	ORI122
REP024	2005	56 %	Internal	Replication-Only	11	Experiment	ORI024
REP035	2005	100 %	Internal	Original-Included	9	Quasi-Experiment	
REP037	2005	78 %	External	Replication-Only	8	Quasi-Experiment	ORI025
REP039	2005	78 %	External	Replication-Only	8	Quasi-Experiment	ORI039
REP068	2005	100 %	Internal	Original-Included	2	Experiment	
REP071	2005	78 %	Internal	Original-Included	4	Quasi-Experiment	
REP076	2005	69 %	External	Replication-Only	9	Case Study	REP132
REP101	2005	89 %	Internal	Original-Included	2	Experiment	
REP103	2005	94 %	Internal	Original-Included	11	Experiment	
REP129	2005	78 %	Internal	Original-Included	2	Experiment	
REP133	2006	83 %	Internal	Original-Included	11	Quasi-Experiment	
REP134	2006	44 %	Internal	Original-Included	4	Quasi-Experiment	
REP019	2006	89 %	Internal	Original-Included	7	Experiment	
REP025	2006	63 %	External	Replication-Only	8	Quasi-Experiment	ORI025
REP026	2006	94 %	Internal	Original-Included	2	Quasi-Experiment	
REP098	2006	81 %	Internal	Replication-Only	11	Quasi-Experiment	ORI092
REP130	2006	59 %	Internal	Replication-Only	2	Experiment	REP129
REP020	2007	84 %	External	Replication-Only	5	Case Study	ORI020
REP021	2007	78 %	External	Replication-Only	11	Case Study	ORI021
REP027	2007	88 %	Internal	Replication-Only	3	Quasi-Experiment	ORI027
REP029	2007	100 %	Internal	Original-Included	2	Quasi-Experiment	
REP030	2007	100 %	Internal	Original-Included	8	Case Study	
REP061	2007	28 %	Internal	Original-Included	10	Quasi-Experiment	
REP107	2007	69 %	Internal	Replication-Only	11	Quasi-Experiment	REP133
REP119	2007	56 %	External	Replication-Only	4	Quasi-Experiment	REP134
REP126	2007	94 %	Internal	Replication-Only	2	Experiment	ORI122
REP128	2007	47 %	Internal	Replication-Only	4	Quasi-Experiment	REP134
REP028	2008	81 %	External	Replication-Only	8	Quasi-Experiment	ORI025

Table 14 (continued)

Replication Ref.	Year	Quality Score	Replication Type	Report Type	SWEBOK Chapter	Research Method	Original Ref.
REP023	2008	50 %	Internal	Replication-Only	3	Quasi-Experiment	ORI023
REP066	2008	89 %	Internal	Original-Included	3	Quasi-Experiment	
REP102	2008	78 %	Internal	Original-Included	11	Quasi-Experiment	
REP118	2008	94 %	External	Replication-Only	4	Quasi-Experiment	REP134
REP127	2008	69 %	Internal	Replication-Only	4	Quasi-Experiment	REP134
REP003	2009	100 %	Internal	Original-Included	2	Quasi-Experiment	
REP005	2009	89 %	Internal	Original-Included	11	Case Study	
REP006	2009	63 %	External	Replication-Only	11	Case Study	ORI006
REP009	2009	81 %	External	Replication-Only	5	Quasi-Experiment	ORI009
REP010	2009	63 %	External	Replication-Only	10	Case Study	ORI010
REP011	2009	72 %	Internal	Replication-Only	2	Quasi-Experiment	ORI011
REP014	2009	100 %	Internal	Original-Included	10	Quasi-Experiment	
REP031	2009	100 %	Internal	Original-Included	2	Quasi-Experiment	
REP105	2009	69 %	Internal	Replication-Only	8	Case Study	ORI105
REP111	2009	78 %	Internal	Replication-Only	10	Experiment	ORI111
REP131	2009	94 %	Internal	Replication-Only	2	Experiment	REP129
REP094	2010	100 %	Internal	Original-Included	2	Quasi-Experiment	
REP001	2010	84 %	External	Replication-Only	4	Quasi-Experiment	ORI001
REP007	2010	100 %	Internal	Original-Included	6	Quasi-Experiment	
REP012	2010	81 %	Internal	Replication-Only	2	Quasi-Experiment	ORI012
REP015	2010	100 %	Internal	Original-Included	5	Experiment	
REP016	2010	83 %	Internal	Original-Included	2	Experiment	
REP065	2010	72 %	Internal	Original-Included	4	Case Study	
REP072	2010	100 %	Internal	Original-Included	6	Experiment	
REP095	2010	100 %	Internal	Original-Included	8	Quasi-Experiment	
REP104	2010	63 %	External	Replication-Only	4	Quasi-Experiment	ORI104
REP120	2010	88 %	External	Replication-Only	4	Quasi-Experiment	REP134
REP121	2010	63 %	External	Replication-Only	6	Case Study	ORI121

A.1.2 Composition of the Sets of Replications

In Table 15, we presente the sets with two or more replications and their original studies, grouped by SWEBOK chapter, which enables the reader to identify the members of the sets and their corresponding original studies. Table 15 also shows the dates of publication of the original study and the replications, the type of each replication in the set, and the number of replications that were reported in each paper.

Table 15 Sets with two or more replications

SWEBOK Ch.	Original Study Id.	Replications in the Set	# of Rep.
Ch. 02: Software Requirements	ORI122 (2001)	REP122 (2001)-Internal	2
		REP123 (2002)-Internal	1
		REP124 (2003)-Internal	2

Table 15 (continued)

SWEBOK Ch.	Original Study Id.	Replications in the Set	# of Rep.
		REP125 (2004)-Internal	2
		REP126 (2007)-Internal	2
	REP129 (2005)	REP129 (2005)-Internal	1
		REP130 (2006)-Internal	2
		REP131 (2009)-Internal	1
	REP026 (2006)	REP026 (2006)-Internal	3
	REP029 (2007)	REP029 (2007)-Internal	2
	REP003 (2009)	REP003 (2009)-Internal	3
	REP094 (2009)	REP094 (2010)-Internal	3
	REP016 (2010)	REP016 (2010)-Internal	2
Ch. 04: Software Construction	ORI086 (1991)	REP086 (1997)-External	2
	REP089 (1995)	REP089 (1995)-Internal	2
	REP071 (2005)	REP071 (2005)-Internal	3
	REP134 (2006)	REP134 (2006)-Internal	1
		REP119 (2007)-External	1
		REP128 (2007)-Internal	1
		REP118 (2008)-External	1
		REP127 (2008)-External	3
		REP127 (2008)-Internal	1
		REP120 (2010)-External	2
Ch. 05: Software Testing	ORI085 (1987)	REP085 (1995)-External	2
		REP048 (1997)-External	1
	REP015 (2010)	REP015 (2010)-Internal	2
Ch. 06: Software Maintenance	ORI034 (1996)	REP058 (1998)-External	1
		REP090 (2000)-External	1
		REP034 (2001)-External	1
	REP052 (1999)	REP052 (1999)-Internal	2
	ORI121 (2000)	REP121 (2010)-External	3
	REP072 (2010)	REP072 (2010)-Internal	2
Ch. 07: Software Configuration Management	REP019 (2006)	REP019 (2006)-Internal	2
Ch. 08: Software Engineering Management	ORI025 (2001)	REP037 (2005)-External	1
		REP025 (2006)-External	1
		REP028 (2008)-External	1
	REP082 (2004)	REP082 (2004)-Internal	3
	REP030 (2007)	REP030 (2007)-Internal	2
	REP095 (2010)	REP095 (2010)-Internal	3
Ch. 09: Software Engineering Process	REP132 (2002)	REP132 (2002)-Internal	1
		REP076 (2005)-External	1
	REP035 (2005)	REP035 (2005)-Internal	4
Ch. 11: Software Quality	ORI092 (1996)	REP113 (1997)-Internal	1
		REP092 (2000)-External	1
		REP098 (2006)-Internal	1
	REP053 (1995)	REP053 (1995)-Internal	1
		REP047 (1997)-External	1

Table 15 (continued)

SWEBOK Ch.	Original Study Id.	Replications in the Set	# of Rep.
		REP050 (1998)-External	1
		REP088 (1998)-External	1
		REP112 (1998)-Internal	1
	REP051 (2001)	REP051 (2001)-Internal	2
	REP133 (2006)	REP133 (2006)-Internal	1
		REP107 (2007)-Internal	1

A.2 Reference List of Replications

[REP001] English M, Buckley J, Cahill T (2010) A replicated and refined empirical study of the use of friends in C ++ software. The Journal of Systems & Software 83(11):2275–2286. doi:10.1016/j.jss.2010.07.013

[REP003] Abrahão S, Poels G (2009) A family of experiments to evaluate a functional size measurement procedure for Web applications. The Journal of Systems & Software 82 (2):253–269. doi:10.1016/j.jss.2008.06.031

[REP005] Zhang H (2009) An Investigation of the Relationships between Lines of Code and Defects. 2009 IEEE International Conference on Software Maintenance 274–283.

[REP006] Huynh T, Miller J (2009) Another viewpoint on "evaluating web software reliability based on workload and failure data extracted from server logs." Empirical Software Engineering 14(4):371–396. doi:10.1007/s10664-008-9084-6

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

[REP009] Dias-Neto AC, Travassos GH (2009) Evaluation of {model-based} Testing Techniques Selection Approaches: an External Replication. 3rd International Symposium on Empirical Software Engineering and Measurement 269–278.

[REP010] Geet JV, Demeyer S (2009) Feature Location in COBOL Mainframe Systems: an Experience Report. IEEE International Conference on Software Maintenance 361–370.

[REP011] Abrahão S, Insfran E, Gravino C, Scanniello G (2009) On the Effectiveness of Dynamic Modeling in UML: Results from an External Replication. 3rd International Symposium on Empirical Software Engineering and Measurement 468–472.

[REP012] Ricca F, Scanniello G, Torchiano M, Reggio G, Astesiano E (2010) On the Effectiveness of Screen Mockups in Requirements Engineering: Results from an Internal Replication. Empirical Software Engineering and Measurement.

[REP014] Ceccato M, Penta MD, Nagra J, Falcarin P, Ricca F, Torchiano M, Tonella P (2009) The Effectiveness of Source Code Obfuscation: an Experimental Assessment. 17th International Conference 178–187.

[REP015] Do H, Mirarab S, Tahvildari L, Rothermel G (2010) The Effects of Time Constraints on Test Case Prioritization: A Series of Controlled Experiments. IEEE Transactions on Engineering Management 36(5):593–617.

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

[REP019] Du G, McElroy J, Ruhe G (2006) A Family of Empirical Studies to Compare Informal and Optimization-based Planning of Software Releases. International Symposium on Empirical Software Engineering 212–221.

[REP020] Andersson C (2007) A replicated empirical study of a selection method for software reliability growth models. Empirical Software Engineering 12(2):161–182. doi:10.1007/s10664-006-9018-0

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

[REP023] Falessi D, Capilla R, Cantone G (2008) A Value-Based Approach for Documenting Design Decisions Rationale: A Replicated Experiment. International Conference on Software Engineering 63–70.

[REP024] Ardimento P, Baldassarre MT, Caivano D, Visaggio G (2006) Assessing multiview framework (MF) comprehensibility and efficiency: A replicated experiment. Information and Software Technology 48(5):313–322. doi:10.1016/j.infsof.2005.09.010

[REP025] Lokan C, Mendes E (2006) Cross-company and Single-company Effort Models Using the ISBSG Database: a Further Replicated Study. International Symposium on Empirical Software Engineering 75–84.

[REP026] Staron M, Kuzniarz L, Wohlin C (2006) Empirical assessment of using stereotypes to improve comprehension of UML models: A set of experiments. Journal of Systems and Software 79(5):727–742. doi:10.1016/j.jss.2005.09.014

[REP027] Canfora G, Cimitile A, Garcia F, Piattini M, Visaggio CA (2007) Evaluating performances of pair designing in industry. Journal of Systems and Software 80(8):1317–1327. doi:10.1016/j.jss.2006.11.004

[REP028] Mendes E, Lokan C (2008) Replicating studies on cross- vs single-company effort models using the ISBSG Database. Empirical Software Engineering 13(1):3–37. doi:10.1007/s10664-007-9045-5

[REP029] Ricca F, Penta MD, Torchiano M, Tonella P, Ceccato M (2007) The Role of Experience and Ability in Comprehension Tasks supported by UML Stereotypes. 29th International Conference on Software Engineering 375–384.

[REP030] Baresi L, Morasca S (2007) Three Empirical Studies on Estimating the Design Effort of Web Applications. ACM Transactions on Software Engineering and Methodology 16(4):15. doi:10.1145/1276933.1276936

[REP031] Ricca F, Torchiano M, Penta MD, Ceccato M, Tonella P (2009) Using acceptance tests as a support for clarifying requirements: A series of experiments. Information and Software Technology 51(2):270–283. doi:10.1016/j.infsof.2008.01.007

[REP032] Vokác M, Tichy W, Sjøberg DIK, Arisholm E, Aldrin M (2004) A Controlled Experiment Comparing the Maintainability of Programs Designed with and without Design Patterns—A Replication in a Real Programming Environment. Empirical Software Engineering 9(3):149–195.

[REP033] Briand LC, Bunse C, Daly JW (2001) A Controlled Experiment for Evaluating Quality Guidelines on the Maintainability of Object-Oriented Designs. IEEE Transactions on Software Engineering 27(6):513–530.

[REP034] Prechelt L, Unger B, Philippsen M, Tichy W (2001) A Controlled Experiment on Inheritance Depth as a Cost Factor for Code Maintenance. Journal of Systems and Software 65(2):115–132.

[REP035] Canfora G, García F, Piattini M, Ruiz F, Visaggio CA (2005) A family of experiments to validate metrics for software process models. Journal of Systems and Software 77(2):113–129. doi:10.1016/j.jss.2004.11.007

[REP036] Mendes E, Mosley N, Counsell S (2003) A Replicated Assessment of the Use of Adaptation Rules to Improve Web Cost Estimation. International Symposium on Empirical Software Engineering 100–109.

[REP037] Mendes E, Lokan C, Harrison R, Triggs C (2005) A Replicated Comparison of Cross-company and Within-company Effort Estimation Models using the ISBSG Database. 11th IEEE International Software Metrics Symposium 331–340.

[REP038] Thelin T, Andersson C, Runeson P, Dzamashvili-fogelström N (2004) A Replicated Experiment of Usage-Based and Checklist-Based Reading. 10th International Symposium on Software Metrics.

[REP039] Shepperd M, Cartwright M (2005) A Replication of the Use of Regression Towards the Mean (R2M) as an Adjustment to Effort Estimation Models. 11th IEEE International Software Metrics Symposium

[REP040] Herbsleb JD, Mockus A (2003) An Empirical Study of Speed and Communication in Globally Distributed Software Development. IEEE Transactions on Software Engineering 29(6):481–494.

[REP041] Lanubile F, Mallardo T, Calefato F, Denger C, Ciolkowski M (2004) Assessing the Impact of Active Guidance for Defect Detection: A Replicated Experiment. 10th International Symposium on Software Metrics 269–278.

[REP043] Thelin T, Petersson H, Runeson P (2002) Confidence intervals for capture—recapture estimations in software inspections. Information and Software Technology 44:683–702.

[REP045] Myrtveit I, Stensrud E (1999) A Controlled Experiment to Assess the Benefits of Estimating with Analogy and Regression Models. IEEE Transactions on Software Engineering 25(4):510–525.

[REP046] Briand LC, Langley T, Wieczorek I (2000) A replicated Assessment and Comparison of Common Software Cost Modeling Techniques. International Conference on Software Engineering (2):377–386.

[REP047] Fusaro P, Lanubile F, Visaggio G (1997) A Replicated Experiment to Assess Requirements Inspection Techniques. Empirical Software Engineering 2:39–57.

[REP048] Roper Marc, Wood Murray, Miller James (1997) An empirical evaluation of defect detection techniques. Information and Software Technology 39:763–775.

[REP049] Miller J, Macdonald F (2000) An empirical incremental approach to tool evaluation and improvement. Journal of Systems and Software 51(1):19–35.

[REP050] Sandahl K, Blomkvist O, Karlsson J, Krysander C, Lindvall M, Ohlsson N (1998) An Extended Replication of an Experiment for Assessing Methods for Software Requirements Inspections. Empirical Software Engineering 3:327–354.

[REP051] Laitenberger O, Emam KE, Harbich TG (2001) An Internally Replicated Quasi-Experimental Comparison of Checklist and Perspective-Based Reading of Code Documents. IEEE Transactions on Software Engineering 27(5):387–421.

[REP052] Visaggio G (1999) Assessing the maintenance process through replicated, controlled experiments. Journal of Systems and Software 44(3):187–197.

[REP053] Porter A, Votta LG, Basili V (1995) Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment. IEEE Transactions on Software Engineering 21(6):563–575.

[REP055] Briand LC, Wüst J, Ikonomovski SV, Lounis H (1999) Investigating Quality Factors in Object-Oriented Designs: an Industrial Case Study. International Conference on Software Engineering 345–354.

[REP058] Cartwright M, Shepperd M (1998) An Empirical View of Inheritance. Information and Software Technology 40(14):795–799.

[REP060] Cox K, Phalp K (2000) Replicating the CREWS Use Case Authoring Guidelines Experiment. Empirical Software Engineering 5(3):245–267.

[REP061] Lindvall M, Rus I, Donzelli P, Menon A, Zelkowitz MV, Can-Betin A, Bultan T, et al. (2007) Experimenting with software testbeds for evaluating new technologies. Empirical Software Engineering 12(4):417–444. doi:10.1007/s10664-006-9034-0

[REP065] Shah HB, Görg C, Harrold MJ (2010) Understanding Exception Handling: Viewpoints of Novices and Experts. IEEE Transactions on Software Engineering 36(2):150–161.

[REP066] Lui KM, Chan KCC, Nosek JT (2008) The Effect of Pairs in Program Design Tasks. IEEE Transactions on Software Engineering 34(2):197–211.

[REP068] Anda B, Sjøberg DIK (2005) Investigating the Role of Use Cases in the Construction of Class Diagrams. Empirical Software Engineering 10(3):285–309.

[REP071] Hochstein L, Carver J, Shull F, Asgari S, Basili V, Hollingsworth JK, Zelkowitz MV (2005) Parallel Programmer Productivity: A Case Study of Novice Parallel Programmers. ACM/IEEE Supercomputing Conference 35–43.

[REP072] Lucia AD, Gravino C, Oliveto R, Tortora G (2010) An experimental comparison of ER and UML class diagrams for data modelling. Empirical Software Engineering 15 (5):455–492. doi:10.1007/s10664-009-9127-7

[REP073] Daly J, Brooks A, Miller J, Roper M, Wood M (1994) Verification of Results in Software Maintenance Through External Replication. IEEE International Conference on Software Maintenance 50–57.

[REP076] Dinh-Trong TT, Bieman JM (2005) The FreeBSD Project: A Replication Case Study of Open Source Development. IEEE Transactions on Software Engineering 31 (6):481–494.

[REP082] Jørgensen M, Teigen KH, Moløkken K (2004) Better Sure Than Safe? Overconfidence in Judgment Based Software Development Effort Prediction Intervals. Journal of Systems and Software 70(1):79–93.

[REP083] Agarwal R, De P, Sinha AP (1999) Comprehending Object and Process Models: An Empirical Study. IEEE Transactions on Software Engineering 25(4):541–556.

[REP085] Kamsties E, Lott CM (1995) An Empirical Evaluation of Three Defect-Detection Techniques. Information and Software Technology 39(11):763–775.

[REP086] Kiper J, Auerheimer B (1997) Visual Depiction of Decision Statements: What is Best For Programmers and Non-programmers.

[REP087] Land LPW, Jeffery R, Sauer C (1997) Validating the Defect Detection Performance Advantage of Group Designs for Software Reviews: Report of a Replicated Experiment. Engineering 17–26.

[REP088] Miller J, Wood M, Roper M (1998) Further Experiences with Scenarios and Checklists. Empirical Software Engineering 3:37–64.

[REP089] Zweben SH, Edwards SH, Weide BW, Hollingsworth JE (1995) The Effects of Layering and Encapsulation on Software Development Cost and Quality. IEEE Transactions on Software Engineering 21(3):200–208.

[REP090] Harrison R, Counsell S, Nithi R (2000) Experimental assessment of the effect of inheritance on the maintainability of object-oriented systems. Journal of Systems and Software 52:173–179.

[REP091] Prechelt L, Unger-lamprecht B, Philippsen M, Tichy WF (2002) Two Controlled Experiments Assessing the Usefulness of Design Pattern Documentation in Program Maintenance. IEEE Transactions on Software Engineering 28(6):595–606.

[REP092] Regnell B, Runeson P, Thelin T (2001) Are the Perspectives Really Different?—Further Experimentation on Scenario-Based Reading of Requirements. Empirical Software Engineering 5(1):331–356.

[REP093] Arisholm E, Sjøberg DIK (2004) Evaluating the Effect of a Delegated versus Centralized Control Style on the Maintainability of Object-Oriented Software. IEEE Transactions on Software Engineering 30(8):521–534.

[REP094] Ricca F, Penta MD, Torchiano M, Tonella P, Ceccato M (2010) How Developers' Experience and Ability Influence Web Application Comprehension Tasks Supported by UML Stereotypes: A Series of Four Experiments. IEEE Transactions on Software Engineering 36(1):96–118.

[REP095] Jørgensen M (2010) Identification of more risks can lead to increased over-optimism of and over-confidence in software development effort estimates. Information and Software Technology 52(5):506–516. doi:10.1016/j.infsof.2009.12.002

[REP098] Maldonado C, Carver J, Shull F, Fabbri S, Dória E, Martiniano L, Mendonça M, et al. (2006) Perspective-Based Reading: A Replicated Experiment Focused on Individual Reviewer Effectiveness. Empirical Software Engineering 11(1):119–142. doi:10.1007/s10664-006-5967-6

[REP101] Verelst JAN (2005) The Influence of the Level of Abstraction on the Evolvability of Conceptual Models of Information Systems. Empirical Software Engineering 10(4):467–494.

[REP102] Koru AG, Emam KE, Zhang D, Liu H, Mathew D (2008) Theory of relative defect proneness - Replicated studies on the functional form of the size-defect relationship. Empirical Software Engineering 13:473–498. doi:10.1007/s10664-008-9080-x

[REP103] Müller MM (2005) Two controlled experiments concerning the comparison of pair programming to peer review. Journal of Systems and Software 78:166–179. doi:10.1016/j.jss.2004.12.019

[REP104] Calefato F, Gendarmi D, Lanubile F (2010) Investigating the use of tags in collaborative development environments: a replicated study. International Symposium on Empirical Software Engineering and Measurement 24:1–24:9.

[REP105] Mendes E, Lokan C (2009) Investigating the Use of Chronological Splitting to Compare Software Cross-company and Single-company Effort Predictions: A Replicated Study. 32nd Australian Conference on Computer Science.

[REP106] Wesslén A (2000) A Replicated Empirical Study of the Impact of the Methods in the PSP on Individual Engineers. Empirical Software Engineering 5:93–123.

[REP107] Phongpaibul M, Boehm B (2007) A Replicate Empirical Comparison between Pair Development and Software Development with Inspection. 1st International Symposium on Empirical Software Engineering and Measurement 265–274. doi:10.1109/ESEM.2007.33

[REP111] Lucia AD, Oliveto R, Tortora G (2009) Assessing IR-based traceability recovery tools through controlled experiments. Empirical Software Engineering 14:57–92. doi:10.1007/s10664-008-9090-8

[REP112] Porter A, Votta LG (1998) Comparing Detection Methods For Software Requirements Inspections: A Replication Using Professional Subjects. Empirical Software Engineering 3:355–379.

[REP113] Ciolkowski M, Differding C, Laitenberger O, Münch J (1997) Empirical Investigation of Perspective-based Reading: A Replicated Experiment.

[REP118] Lung J, Aranda J, Easterbrook S, Wilson G (2008) On the Difficulty of Replicating Human Subjects Studies in Software Engineering. International Conference on Software Engineering 191–200.

[REP119] Caspersen ME, Bennedsen J, Larsen KD (2007) Mental Models and Programming Aptitude. ACM SIGCSE Bulletin 39(3):206–210.

[REP120] França ACC, da Cunha PRM, Da Silva FQB (2010) The Effect of Reasoning Strategies on Success in Early Learning of Programming: Lessons Learned from an External

Experiment Replication. 14th International Conference on Evaluation and Assessment in Software Engineering 1–10.

[REP121] Ma X, Zhou M, Mei H (2010) How Developers Participate in Open Source Projects: a Replicate Case Study on JBossAS, JOnAS and Apache Geronimo. Workshop on Replication in Empirical Software Engineering Research.

[REP122] Genero M, Piattini M, Jiménez L (2001) Empirical validation of class diagram complexity metrics. 21st International Conference of the Chilean Computer Science Society 95–104. doi:10.1109/SCCC.2001.972637

[REP123] Genero M, Jiménez L, Piattini M (2002) A Controlled Experiment for Validating Class Diagram Structural Complexity Metrics. OOIS'02 Proceedings of the 8th International Conference on Object-Oriented 372–383.

[REP124] Genero M, Piattini M, Manso E, Cantone G (2003) Building UML Class Diagram Maintainability Prediction Models Based on Early Metrics. 9th International Software Metrics Symposium 263–275.

[REP125] Genero M, Piatini M, Manso E (2004) Finding "Early" Indicators of UML Class Diagrams Understandability and Modifiability. 2004 International Symposium on Empirical Software Engineering 207–216.

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

[REP127] Bornat R, Dehnadi S, Simon (2008) Mental models, Consistency and Programming Aptitude. ACE'08 Proceedings of the tenth conference on Australasian computing education.

[REP128] Wray S (2007) SQ Minus EQ can Predict Programming Aptitude. Psychology of Programming Interest Group 19th Annual Workshop 243–254.

[REP129] Cruz-Lemus JA, Genero M, Manso ME, Piattini M (2005) Evaluating the Effect of Composite States on the Understandability of UML Statechart Diagrams. Lecture Notes in Computer Science 3713:113–125.

[REP130] Cruz-Lemus JA, Genero M, Piattini M, Morasca S (2006) Improving the Experimentation for Evaluating the Effect of Composite States on the Understandability of UML Statechart Diagrams. 5th ACM-IEEE International Symposium on Empirical Software Engineering (ISESE 2006) 9–11.

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

[REP132] Mockus A, Fielding RT, Herbsleb JD (2002) Two Case Studies of Open Source Software Development: Apache and Mozilla. ACM Transactions on Software Engineering and Methodology 11(3):309–346.

[REP133] Phongpaibul M, Boehm B (2006) An Empirical Comparison Between Pair Development and Software Inspection in Thailand. International Symposium on Empirical Software Engineering 85–94.

[REP134] Dehnadi S, Bornat R (2006) The camel has two humps (working title). Little Psychology of Programming Interest Group 2(23):1–21.

A.3 Reference List of Original Studies

[ORI001] Counsell S (2000) Use of friends in C++ software: an empirical investigation. Journal of Systems and Software 53(1):15–21. doi:10.1016/S0164-1212(00)00004-2

[ORI006] Tian J, Rudraraju S, Li Z (2004) Evaluating Web software reliability based on workload and failure data extracted from server logs. IEEE Transactions on Software Engineering 30(11):754–769. doi:10.1109/TSE.2004.87

[ORI009] Vegas S, Basili V (2005) A Characterisation Schema for Software Testing Techniques. Empirical Software Engineering 10(4):437–466. doi:10.1007/s10664-005-3862-1

[ORI010] Eisenbarth T, Koschke R, Simon D (2003) Locating Features in Source Code. IEEE Transactions on Software Engineering 29(3)

[ORI011] Gravino C, Scanniello G, Tortora G (2008) An Empirical Investigation on Dynamic Modeling in Requirements Engineering. MoDELS 615–629.

[ORI012] Ricca F, Scanniello G, Torchiano M, Reggio G, Astesiano, E (2010) Can screen mockups improve the comprehension of functional requirements? http://www.scienzemfn.unisa.it/scanniello/ScreenMockupExp/material/main.pdf. Accessed 13 October 2011

[ORI020] Stringfellow C, Falls W, Andrews AA, Science C (2002) An Empirical Method for Selecting Software Reliability Growth Models. Empirical Software Engineering 7:319–343.

[ORI021] Fenton NE, Ohlsson N (2000) Quantitative Analysis of Faults and Failures in a Complex Software System. IEEE Transactions on Software Engineering 26(8):797–814.

[ORI023] Falessi D, Cantone G, Kruchten P (2008) Value-Based Design Decision Rationale Documentation: Principles and Empirical Feasibility Study. 7th Working IEEE/IFIP Conference on Software Architecture 189–198. doi:10.1109/WICSA.2008.8

[ORI024] Baldassarre MT, Caivano D, Visaggio G (2003) Comprehensibility and efficiency of multiview framework for measurement plan design. International Symposium on Empirical Software Engineering. 89–98. doi:10.1109/ISESE.2003.1237968

[ORI025] Jeffery R, Ruhe M, Wieczorek I (2001) Using Public Domain Metrics to Estimate Software Development Effort. 7th International Software Metrics Symposium 16–27.

[ORI027] Canfora G, Cimitile A, Garcia F, Piattini M, Visaggio CA (2006) Performances of pair designing on software evolution: a controlled experiment. Conference on Software Maintenance and Reengineering 197–205. doi:10.1109/CSMR.2006.40

[ORI032] Prechelt L, Unger B, Tichy WF, Brössler P, Votta LG (2001) A Controlled Experiment in Maintenance Comparing Design Patterns to Simpler Solutions. IEEE Transactions on Software Engineering 27(12):1134–1144.

[ORI033] Briand LC, Bunse C, Daly JW, Differding C (1996) An experimental comparison of the maintainability of object-oriented and structured design documents. International Conference on Software Maintenance 130–138. doi:10.1109/ICSM.1997.624239

[ORI034] Daly J, Brooks A, Miller J, Roper M, Wood M (1996) Evaluating Inheritance Depth on the Maintainability of Object-Oriented Software. Empirical Software Engineering 1(2):109–132.

[ORI036]Mendes E, Mosley N, Counsell Steve (2003) Do Adaptation Rules Improve Web Cost Estimation? 14th ACM Conference on Hypertext and Hypermedia 174–183.

[ORI038] Thelin T, Runeson P, Wohlin C (2003) An Experimental Comparison of Usage-Based and Checklist-Based Reading. IEEE Transactions on Software Engineering 29 (8):687–704.

[ORI039] Jørgensen M, Indahl U, Sjøberg D (2003) Effort Estimation: Software Effort Estimation by Analogy and "Regression Toward the Mean." Journal of Systems and Software 68:253–262.

[ORI041] Denger C, Ciolkowski M, Lanubile F (2003) Does Active Guidance Improve Software Inspections? A Preliminary Empirical Study. IEEE Transactions on Software Engineering 29(6):408–413.

[ORI043] Miller J (1999) Estimating the number of remaining defects after inspection. Software Testing, Verification and Reliability, 9(3): 167–189.

[ORI045] Shepperd M, Schofield C, Kitchenham B (1996) Effort Estimation Using Analogy. International Conference on Software Engineering 170–178.

[ORI046] Briand LC, Emam KE, Surmann D, Wieczorek I, Maxwell KD (1999) An Assessment and Comparison of Common Software Cost Estimation Modeling Techniques. 21st International Conference on Software Engineering 313–322.

[ORI055] Briand LC, Daly JW, Porter V, Wüst J (1998) A Comprehensive Empirical Validation of Product Measures for Object-Oriented Systems. 5th International Software Metrics Symposion 246–257.

[ORI060] Achour CB, Rolland C, Maiden NAM, Souveyet C (1998) Guiding Use Case Authoring: Results of an Empirical Study. 4th IEEE International Symposium on Requirements Engineering.

[ORI085] Basili VR, Selby RW (1987) Comparing the Effectiveness of Software Testing Strategies. IEEE Transactions on Software Engineering SE-13(12):1278–1296.

[ORI087] Land LPW, Sauer C, Jeffery R (1997) Validating the Defect Detection Performance Advantage of Group Designs for Software Reviews: Report of a Laboratory Experiment Using Program Code. European Sofrware Engineering Conference 22–25.

[ORI092] Basili VR, Green S, Laitenberger O, Lanubile F, Shull F, Sørumgård S, Zelkowitz MV (1996) The Empirical Investigation of Perspective-Based Reading Reading. Empirical Software Engineering 1(2):133–164.

[ORI093] Arisholm E, Sjøberg DIK, Jørgensen M (2001) Assessing the Changeability of two Object-Oriented Design Alternatives - a Controlled Experiment. Empirical Software Engineering 6(3):231–277.

[ORI104] Treude C, Storey M-anne (2009) How Tagging helps bridge the Gap between Social and Technical Aspects in Software Development. International Conference on Software Engineering 12–22.

[ORI105] Lokan C, Mendes E (2008) Investigating the Use of Chronological Splitting to Compare Software Cross-company and Single-company Effort Predictions. 12th International Conference on Evaluation and Assessment in Software Engineering.

[ORI106] Hayes W, Over JW (1997) The Personal Software Process (PSP): An Empirical Study of the Impact of PSP on Individual Engineers.

[ORI111] Lucia AD, Fasano F, Oliveto R (2007) Recovering Traceability Links in Software Artifact Management Systems using Information Retrieval Methods. Transactions on Software Engineering and Methodology 16(4):13. doi:10.1145/1276933.1276934

[ORI121] Mockus A, Fielding RT, Herbsleb J (2000) A Case Study of Open Source Software Development: The Apache Server. International Conference on Software Engineering 263–272.

[ORI122] Genero M, Jiménez L, Piattini M (2001) A Prediction Model for OO Information System Quality Based on Early Indicators. Advances in Databases and Information Systems.

Appendix B—Quality Assessment

In this mapping study, we are interested in evaluating the quality of the empirical studies in general and the specific quality aspects of replication reports. We did not use quality assessment to exclude reports, only to allow comparative analysis about the studies related to the quality of the information reported in the papers.

B.1 Quality Assessment Criteria

Kitchenham and Charters (2007, pp. 25), and Dybå and Dingsøyr (2008) informed our choice of quality assessment criteria to assess issues that were related to empirical studies in general. We added replication specific criteria extracted from the propositions presented by Carver (2010). Table 16 shows the complete set of criteria, which includes the nine replication-specific criteria marked (RS), and the seven generic criteria marked (GC).

Table 16 Quality assessment criteria for replication-only replications (internal and external)

Study Phase	Quality Criteria	Quality Score		
	Has this paper clearly…	Yes	Partially	No
Design	(GC) … defined its objective?			
	(RS) … described the original study's objective?			
	(GC) … described its research method?			
	(RS) … described the original study's research method?			
	(RS) … described what research parameters are different (variations) from the original?			
Conduct	(GC) … described the environment where it was conducted?			
	(RS) … described the environment where the original study was performed?			
Analysis	(GC) … described its data analysis method?			
	(RS) … described the original study's data analysis methods?			
Conclusion	(GC) … achieved its results (answered research questions or tested hypothesis)?			
	(GC) … described its results?			
	(RS) … described the original study's results?			
	(RS) … described the comparability of its results to the original study's results?			
	(CG) … discussed threats to its validity?			
	(RS) … discussed threats to validity of the original study?			
	(RS) … described limitations/difficulties regarding the replication process?			

It is important to note that we assessed the quality of the papers reporting replications, not the quality of each individual replication reported in the papers. When a paper reported more than one replication, we assessed the quality of the entire study not each replication separately. Although this might be a limitation, the lack of detailed information about each individual replication within the papers made individual quality assessment difficult or even impossible.

In the initial definition of the review protocol, we built one set of quality criteria to be used with all papers reporting replications. As explained above, these criteria included items to assess the replication specific aspects, including the quality of the description of the original study, which was advocated to be an essential part of replication reports by Carver (2010). During the quality assessment process, we noticed that several papers reporting internal replications presented the replication (or a set of replications) and the original study in the same paper. For most of these papers, we could not find a clear-cut way to separate the description of the original study from the

description of the replications and therefore could not evaluate most of the (RS) criteria.

At this point in the analysis, we found it necessary to separate the papers that reported one or more replications together with an original study (called Original-Included reports) from the papers that reported the original study separately (Replication-Only reports). Replication-Only replications (including both internal and external replications) were assessed using the criteria presented in Table 16 (the initial set of criteria). Original-Included internal replications were assessed using the criteria in Table 17 (a subset of the initial set with seven (RS) criteria removed). We updated the mapping study protocol to reflect this deviation from the initial plan and discuss the implications of having two sets of criteria in Section B.2.

Table 17 Quality assessment criteria for original-included replications (internal)

Study Phase	Quality Criteria	Quality Score		
	Has this paper clearly…	Yes	Partially	No
Design	(GC) … defined its objective?			
	(GC) … described its research method?			
	(RS) … described what research parameters are different (variations) among original and replications?			
Conduct	(GC) … described the environment where it was conducted?			
Analysis	(GC) … described its data analysis method?			
Conclusion	(GC) … achieved its results (answered research questions or tested hypothesis)?			
	(GC) … described its results?			
	(RS) … described the comparability of the results of the multiple studies?			
	(GC) … discussed threats to its validity?			

Two researchers performed the assessment of each paper by assigning a score for each quality item on a three-point scale. A third researcher solved the disagreements. When a solution was not reached by the third assessment, the final quality score was defined in a consensus meeting with the rest of the research team. We used Cohen's Kappa (κ) coefficient (Cohen 1960) to measure the agreement level between assessments before disagreements were resolved.

B.2 Quality Assessment Results

In the quality assessment process, we achieved an acceptable inter-rater agreement (κ=0.69) between two researchers, before a third researcher or a consensus meeting solved the disagreements. In Table 18, we presente references to the replication papers, grouped by quartiles of the quality score, and types of reports (Original-Included Internal, Replication-Only Internal, and External). The scores were presented as a percentage of the maximum score, making it possible to compare these three sets.

From Table 18, we can see that Original-Included internal replications papers seem to score higher than Replication-Only, internal and external replications. Furthermore, it seems that external replications scored lower than internal ones. These tendencies are better visualized in Fig. 10.

Table 18 Quartiles of quality scores

Quality Score Quartiles	Rep./Type		
	Internal		External
	Original-Included	Replication-Only	
Q4 [92 %, 100 %]	REP003 (100 %)	REP124 (100 %)	REP118 (94 %)
	REP007 (100 %)	REP038 (94 %)	
	REP014 (100 %)	REP126 (94 %)	
	REP015 (100 %)	REP131 (94 %)	
	REP029 (100 %)		
	REP030 (100 %)		
	REP031 (100 %)		
	REP035 (100 %)		
	REP049 (100 %)		
	REP052 (100 %)		
	REP068 (100 %)		
	REP072 (100 %)		
	REP083 (100 %)		
	REP094 (100 %)		
	REP095 (100 %)		
	REP051 (94 %)		
	REP053 (94 %)		
	REP026 (94 %)		
	REP103 (94 %)		
Q3 [79 %, 91 %]	REP005 (89 %)	REP032 (91 %)	REP050 (91 %)
	REP019 (89 %)	REP027 (88 %)	REP088 (88 %)
	REP040 (89 %)	REP033 (88 %)	REP120 (88 %)
	REP066 (89 %)	REP125 (82 %)	REP001 (84 %)
	REP091 (89 %)	REP012 (81 %)	REP020 (84 %)
	REP101 (89 %)	REP098 (81 %)	REP009 (81 %)
	REP016 (83 %)	REP112 (81 %)	REP028 (81 %)
	REP133 (83 %)		REP034 (81 %)
			REP106 (81 %)
Q2 [70 %, 78 %]	REP071 (78 %)	REP036 (78 %)	REP021 (78 %)
	REP102 (78 %)	REP111 (78 %)	REP037 (78 %)
	REP129 (78 %)	REP113 (78 %)	REP039 (78 %)
	REP065 (72 %)	REP041 (75 %)	REP047 (78 %)
	REP089 (72 %)	REP011 (72 %)	REP060 (78 %)
	REP132 (72 %)	REP087 (72 %)	REP085 (78 %)
			REP086 (75 %)
			REP043 (72 %)
			REP073 (72 %)
Q1 [0 %, 69 %]	REP082 (67 %)	REP105 (69 %)	REP048 (69 %)
	REP134 (44 %)	REP107 (69 %)	REP076 (69 %)
	REP061 (28 %)	REP127 (69 %)	REP006 (63 %)

Table 18 (continued)

Quality Score Quartiles	Rep./Type		
	Internal		External
	Original-Included	Replication-Only	
		REP055 (66 %)	REP010 (63 %)
		REP046 (63 %)	REP025 (63 %)
		REP123 (63 %)	REP104 (63 %)
		REP130 (59 %)	REP121 (63 %)
		REP024 (56 %)	REP090 (59 %)
		REP093 (56 %)	REP119 (56 %)
		REP023 (50 %)	REP045 (50 %)
		REP122 (47 %)	REP058 (44 %)
		REP128 (47 %)	REP092 (44 %)

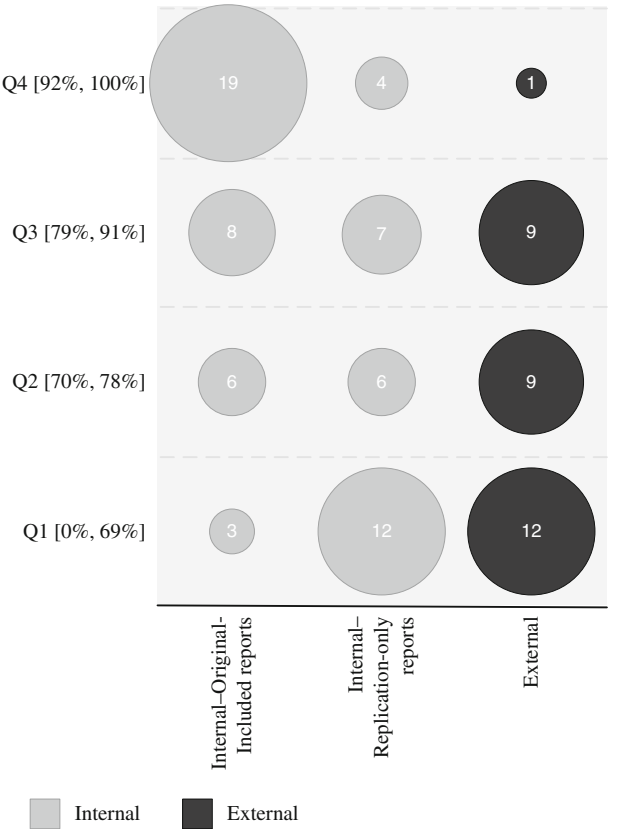


Fig. 10 Quartile of quality Sscores for internal and external replications

We calculated the average of the quality scores for each replication type (Table 19). Since the scores were obtained using two different sets of criteria, we did not calculate the average for the entire set of replications nor averaged the quality of Original-Included and Replication-Only reports together.

Table 19 Mean quality scores

Sub-sets of Report Types	N	Mean	Std. Dev.
Replication-Only (Internal and External)	60	73 %	0.14
Replication-Only External Replications	31	72 %	0.13
Replication-Only Internal Replications	29	74 %	0.15
Original-Included Internal Replications	36	88 %	0.16

As visualized in Table 18 and Fig. 10, the entire set of Replication-Only reports of replications (internal and external) scored significantly lower than the Original-Included reports of internal replications ($t=-4.769$, df=94, $p<0.001$). Similarly, the Replication-Only internal and Replication-Only external sub-sets scored significantly lower than Original-Included internal sub-set, ($t=-3.629$, df=63, $p<0.001$) and ($t=-4.246$, df=65, $p<0.001$) respectively. As predicted by looking at Fig. 10, no significant difference in the means of quality score was found between Replication-Only internal and external replications.

These differences are explained by looking at the individual quality assessment criteria. Replication-Only external and internal replications were assessed with respect to the replication specific criteria (RS criteria), which included criteria that evaluated the level of information about the original study, and they scored consistently lower in most of these criteria than on the rest of the criteria, as can be seen in the scores highlighted in boldface in Table 20. As explained above, most of the (RS) criteria were not used to assess Original-Included internal replications, due to a lack of consistent information in the papers. When we use the same sub-set of criteria to evaluate Original-Included reports of replications (removing the criteria highlighted in boldface in Table 20), the average quality score of external replications increased to 87 %, and the average for Replication-Only (internal and external together) increase to 89 %, both very close to the 88 % of the set of Original-Included internal replications.

Table 20 Scores of individual quality assessment items for external replications

Study Phase	Quality Criteria Has this paper clearly…	Score (%)
Design	(GC) … defined its objective?	97 %
	(RS) … described the original study's objective?	76 %
	(GC) … described its research method?	90 %
	(RS) … described the original study's research method?	66 %
	(RS) … described what research parameters are different (variations) from the original?	89 %
Conduct	(GC) … described the environment where it was conducted?	94 %
	(RS) … described the environment where the original study was performed?	55 %
Analysis	(GC) … described its data analysis method?	87 %
	(RS) … described the original study's data analysis methods?	47 %
Conclusion	(GC) … achieved its results (answered research questions or tested hypothesis)?	95 %

Table 20 (continued)

Study Phase	Quality Criteria Has this paper clearly…	Score (%)
	(GC) … described its results?	100 %
	(RS) … described the original study's results?	69 %
	(RS) … described the comparability of its results to the original study's results?	81 %
	(CG) … discussed threats to its validity?	55 %
	(RS) … discussed threats to validity of the original study?	16 %
	(RS) … described limitations/difficulties regarding the replication process?	42 %

These results indicated that the overall quality of the papers is good if we considere generic quality assessment criteria to evaluate the empirical studies in general (GC criteria discussed in Section B.1). However, if we adde replication specific (RS) quality criteria, as proposed by Carver (2010), the quality of the Replication-Only reports decrease.

One could argue that our choice of criteria biased the results; since we picked the criteria that the replication reports (in particular the Replication-Only reports) scored consistently low on. However, it is important to remember that the choice of criteria was made during protocol development, before we selected the papers, and was based on accepted guidelines. Although it is true that the two sets of replications scored similarly using the (CG) criteria, the differences in the scores using replication specific criteria was relevant for this study, because we were interested in the evaluation of replication reports. Furthermore, Original-Included and Replication-Only reports of replications were also distinct with respect to other factors presented in the results section. These distinctions must be carefully assessed, because they may be indicative of limitations and threats to validity of replication studies. In particular, we argue these distinctions are indicative of publication bias, as discussed in the main part of our article.

We also compared the quality scores of papers published in journals with the scores of papers published in other sources (conferences proceedings, etc.). Journal papers scored significantly higher than non-journal papers (t=3,269; df=131; sig.=0,001). Furthermore, we compared the quality of original-included reports with the other types of reports in journals and in non-journals sources. In both cases, original-included reports scored significantly higher than the other types of reports, with (t=3,703; df=70; sig.=001) for the difference in the journal papers and (t=4,732; df=59; sig.=000) for the difference in the non-journal papers. The scores of the Original-included reports in journals and non-journal are not significantly different. Replication-only reports also have non-significant difference in scores between journals and non-journal papers. Table 21 shows the mean score and standard deviation for each subset of report type.

Table 21 Quality scores for journal and conference proceedings papers

Sub-sets of Report Types	N	Mean	Std. Dev.
Journal papers	72	85 %	0.16
Original-included	43	90 %	0.15
Replication-only	29	77 %	0.14
Other sources (non-journal papers)	61	76 %	0.16
Original-included	17	89 %	0.14
Replication-only	44	71 %	0.14

In general, replication reports (specifically the Replication-Only reports) lack descriptive information detailing the original studies and other replication specific information. We can think of three reasons to explain these poor descriptions:

First, a complete description of the original study required space and may have been left out in certain publications that had constraints on page count (typical in conference proceedings, but also found in some journals or in special issues). Second, the researchers were unaware that descriptive information should have been included in their papers. Third, detailed information about the original study was not available and the researchers reported what they could find. Replication reporting guidelines should have addressed the second reason. A persistent repository to store data about experiments and replications could have helped with the first and third reasons. If the third reason applied, it would have helped to explain our findings in RQ6 related to confirmation of results.

References

Abran A, Moore J, Bourque P, Dupuis T (Eds.) (2004) Guide to software engineering body of knowledge, IEEE Computer Society. 204

Almqvist JPF (2006) Replication of controlled experiments in empirical software engineering —a survey. Master's Thesis, Department of Computer Science, Faculty of Science, Lund University, Sweden. 129

Arksey H, O'Malley L (2005) Scoping studies: towards a methodological framework. Int J Soc Res Meth 8:19–32

Basili V et al (1999) Building knowledge through families of experiments. IEEE Trans Software Eng 25:456–473. doi:10.1109/32.799939

Brooks A et al. (1995) Replication of Experimental Results in Software Engineering. Technical Report, EFoCS-17-95 [RR/95/193], Dept. of Computer Science, Univ. of Strathclyde. 38

Brooks A et al. (2007) Replication's role in software engineering. In F Shull, J Singer, and DIK Sjøberg (eds) Guide to Advanced Empirical Software Engineering. Springer, pp 365–379

Carver JC. (2010) Towards Reporting Guidelines for Experimental Replications: A Proposal. In RESER'2010: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research, Cape Town, South Africa. 4

Carver JC et al. (2003) Issues in using students in empirical studies in soft —ware engineering education. In *Proceedings of the 9th International Software Metrics Symposium (METRICS2003),* pp239–249

Ciolkowski M et al. (2004) Using academic courses for empirical validation of software development processes. In Proceedings of the 30th Euromicro Conference, pp 354–361

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46. doi:10.1177/001316446002000104

da Silva FQB et al (2011a) Six years of systematic literature reviews in software engineering: an updated tertiary study. Inform Software Tech 53(9):899–913. doi:10.1016/j.infsof.2011.04.004

da Silva FQB et al. (2011b) Replication of empirical studies in software engineering: Preliminary findings from a systematic mapping study. Proceedings of the 2nd International Workshop on Replication in Empirical Software Engineering Research RESER'2011, pp 61–70

Daly J, Brooks A, Miller J, Roper M, Wood M (1994) Verification of Results in Software Maintenance Through External Replication. IEEE International Conference on Software Maintenance, pp. 50–57

Davidsen MK, Krogstie J (2010) A longitudinal study of development and maintenance. Inform Software Tech 52(7):707–719

Dybå T, Dingsøyr T (2008) Empirical studies of agile software development: a systematic review. Inform Software Tech 50:833–859

Easterbrook SM et al. (2007) Selecting Empirical Methods for Software Engineering Research.. In: F Shull, J Singer and D Sjøberg (eds.) Guide to Advanced Empirical Software Engineering. Springer, pp 285–311

França A César C et al. (2010) The Effect of Reasoning Strategies on Success in Early Learning of Programming: Lessons Learned from an External Experiment Replication. In EASE'2010: 14th International Conference on Evaluation and Assessment in Software Engineering, Keele University, UK. 10

Gómez G, Omar S, Juristo N, Vegas N (2010a) Replication, Reproduction and Re-analysis: Three ways for verifying experimental findings. In RESER'2010: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research. Cape Town, South Africa. pp 42–44

Gómez G, Omar S, Juristo N, Vegas N (2010b) Replications Types in Experimental Disciplines. In ESEM'2010: Proceedings of the ACM/IEEE 4th International Symposium on Empirical Software Engineering and Measurement, September 16–17, Bolzano-Bozen, Italy. pp. 1–10

Gould J, Kolb WL (eds) (1964) A dictionary of the social sciences. Tavistock Publications, London, 761

Holgeid KK, Krogstie J, Sjøberg DIK (2000) A study of development and maintenance in Norway: assessing the efficiency of information systems support using functional maintenance. Inform Software Tech 42:687–700

Juristo N, Vegas S (2009) Using differences among replications of software engineering experiments to gain knowledge. In ESEM'09: Proceedings of the ACM/IEEE 3rd International Symposium on Empirical Software Engineering and Measurement. IEEE Computer Society, Washington, DC, USA, pp 356–366

Kitchenham B (2008) The role of replications in empirical software engineering—a word of warning. Empir Software Eng 13:219–221

Kitchenham B, Charters S (2007) Guidelines for performing systematic literature reviews in software engineering, Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University

Kitchenham BA, Pfleeger SL (2007) Personal Opinion Surveys. In: F Shull, J Singer, D. Sjøberg (eds), pp. 63–92, Guide to Advanced Empirical Software Engineering, Springer

Kitchenham B, Dybå T, Jørgensen M (2004) Evidence-based Software Engineering. In ICSE'2004: Proceedings of the 26th International Conference on Software Engineering, Washington DC, USA. pp 273–281

Kitchenham B et al (2010) Literature reviews in software engineering—a tertiary study. Inform Software Tech 52:792–805

Krein Jonathan L, Knutson Charles D (2010) A Case for Replication: Synthesizing Research Methodologies in Software Engineering. In RESER'2010: Proceedings of the 1st International Workshop on Replication in Empirical Software Engineering Research, Cape Town, South Africa. 10

Krogstie J, Sølvberg A (1994) Software Maintenance in Norway: a survey investigation. In ICSM'1994: Proceedings of the International Conference on Software Maintenance. pp 304–313

Krogstie J, Jahr A, Sjøberg DIK (2006) A longitudinal study of development and maintenance in Norway: report from the 2003 investigation. Inform Softw Technol 48:993–1005

La Sorte MA (1972) Replication as a verification technique in survey research: a paradigm. Socio Q 13 (2):219–227

Lindsay RM, Ehrenberg A (1993) The design of replicated studies. Am Stat 47(3):217–228

Lung J et al. (2008) On the difficulty of replicating human subjects studies in software engineering. In ICSE'2008: Proceedings of the 13th international conference on Software engineering, New York, USA: ACM Press. pp 191–201

Petticrew M, Roberts H (2006) Systematic Reviews in the Social Sciences. Blackwell Publishing. 336

Popper K (1959) The Logic of Scientific Discovery. Hutchinson & Co. 513

Schmidt S (2009) Shall we really do it again? The powerful concept of replication is neglected in the social sciences. Rev Gen Psychol 13:90–100. doi:10.1037/a0015108

Shull F, Basili V, Carver J, Maldonado JC, Travassos GH, Mendonça M, Fabbri S (2002) Replicating software engineering experiments: Addressing the tacit knowledge problem. In ISESE'2002: Proc. Int. Symp. on Empirical Softw. Eng., Washington, DC, USA, IEEE Computer Society. 10

Shull F, Carver J, Vegas S, Juristo N (2008) The Role of Replications in Empirical Software Engineering. Empir Software Eng 13:211–218

Sjøberg D (2010) Confronting the myth of rapid obsolescence in computing research. Commun ACM 53 (9):62–67

Sjøberg D et al (2005) A survey of controlled experiments in software engineering. IEEE Trans Software Eng 31:733–753

Vegas S et al. (2006) Analysis of the Influence of Communication between Researchers on Experiment Replication. In ISESE'2006: Proceedings of the 5th International Symposium on Empirical Software Engineering. September 20–21, Rio de Janeiro, Brazil. pp 28–37

Yin RK (2009) Case study research: Design and methods, 4th edn. Sage Publications, London, 240

Zhang H, Babarb MA, Tell P (2010) Identifying relevant studies in software engineering. Inform Software Tech 53(6):625–637, http://dx.doi.org/10.1016/j.infsof.2010.12.010

Fabio Q. B. da Silva has a degree of Ph.D. in Computer Science from the Laboratory for Foundations of Computer Science, University of Edinburgh, Scotland, since 1992. He is an Associate Professor at Center of Informatics of Federal University of Pernambuco, in Brazil, since 1993. His research interests include human and social aspects of software engineering, empirical methods in software engineering, management of innovation, entrepreneurship, and new enterprise creation. He is one of the founders of the Recife Center for Advances Studies and Systems (CESAR) and of the Porto Digital Science Park, of which he was the first President between 2000 and 2003.

Marcos Suassuna is a Ph.D. candidate at the Centre for Informatics, UFPE, where he is studying the role of replications in empirical software engineering research. He received his M.Sc. in 2011 and his B.Sc. in 1980. He is also a partner and Director of MEGA Consultants Associated Ltd., where he conduct consultancy in the field of Information Technology and Organizational Management for over 30 years. His research interests include empirical software engineering, replications, strategy and innovation management.

A. César C. França is a Ph.D. candidate at the Center of Informatics of the Federal University of Pernambuco where he receives a scholarship from the Brazilian National Research Council (CNPq), process #141156/2010-4. He is also a lecturer on Software Engineering related areas at the College of Philosophy, Science, and Letters of Caruaru (Brazil).

Alicia M. Grubb is currently a Ph.D. candidate, under the supervision of Steve Easterbrook, at the University of Toronto. She received her M.Sc. from the University of Toronto in 2009 and her B.S.E. from the University of Waterloo in 2008. Her research interests include systems thinking, software engineering for development, and empirical replication.

Tatiana Bittencourt Gouveia is a doctoral student in computer science at the Center for Informatics, of the Federal University of Pernambuco, Brazil. She received the MSc degree in Management from the Federal University of Rio Grande do Sul, Brazil, in 2006. She now works as Human Resources Manager at a software development project, and her main research interest is team cohesion in software engineering.

Cleviton Monteiro is a Ph.D. candidate at the Center of Informatics of the Federal University of Pernambuco, where he receives a scholarship from CAPES, since 2010. His research interests include management of innovation and creativity, empirical software engineering, and human aspects in software engineering. He is one of the founders of FAST Soluções Tecnológicas Ltda.

Igor Ebrahim dos Santos is a M.Sc. candidate in Computer Science at the Federal University of Pernambuco (Brazil). He received his B.Sc. in Computer Science from the same university. He is also a Requirement Analyst and Innovation Management Consultant at FAST Soluções Tecnológicas Ltda.