Replications of software engineering experiments

Jeffrey C. Carver · Natalia Juristo · Maria Teresa Baldassarre · Sira Vegas

1 Background on Replications

Replication is an essential part of the experimental paradigm and is considered the cornerstone of scientific knowledge (Moonesinghe et al. 2007). Experiments need to be replicated in different contexts, at different times and under different conditions before they can produce generalizable knowledge (Campbell and Stanley 1963).

There are many open issues that must be addressed before the replication process can be successfully formalized in empirical software engineering research. We define replication as the deliberate repetition of the same empirical study for the purpose of determining whether the results of the first experiment can be reproduced. This definition would appear at first glance to be good. However, it needs several clarifications that have not yet been forthcoming in software engineering:

– What is the exact meaning of the same empirical study? Namely how similar should an experiment be to the baseline study for it to be considered a replication?

J. C. Carver (✉)
Department of Computer Science, University of Alabama, Tuscaloosa, AL USA
e-mail: carver@cs.ua.edu

N. Juristo · S. Vegas
Facultad de Informatica, Universidad Politecnica de Madrid, Madrid, Spain

N. Juristo
e-mail: natalia@fi.upm.es

S. Vegas
e-mail: svegas@fi.upm.es

N. Juristo
University of Oulu, Oulu, Finland

M. T. Baldassarre
Department of Informatics, University of Bari, Bari, Italy
e-mail: mariateresa.baldassarre@uniba.it

– What is the exact meaning of a result being reproduced? Namely how similar does a result have to be to the result of the baseline study for it to be considered reproduced?

These and other methodological questions need to be researched and tailored for empirical software engineering. Ours is not the only field that has need of a deeper understanding of replication. Schmidt recently noted for the social sciences that "The word replication has been used as a collective term to describe various meanings in different contexts" (Schmidt 2009).

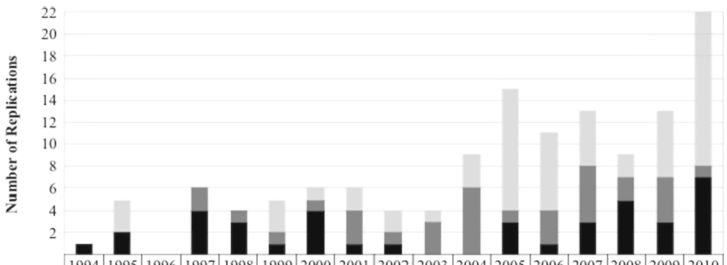
Replication of empirical software engineering studies has not yet attracted enough attention from researchers. The few early works are relatively recent (Basili et al. 1999; Brooks et al. 1995). Although over the last 10 years, there has been some more research aiming at adopting this essential part of the experimental paradigm in SE: (Gómez et al. 2010a, b; Shull et al. 2002; Vegas et al. 2006; Brooks et al. 2008; Juristo and Vegas 2009; Krein and Knutson 2010). However, there is no agreement yet on terminology, typology, purposes, operation and other replication issues. The 2008 point/counterpoint column in the EMSE journal on replication (Kitchenham 2008; Shull et al. 2008) provides a good example of divergent viewpoints.

Although much more methodological research on the adoption of replication in empirical software engineering is still necessary, we have preferred to focus this special issue on the practice rather than the theoretical/methodological issues. This focus on practice is due to the fact that most software engineering experimental results have not yet been reproduced. Three reviews provide empirical data that support this point. Let us look at their results.

Sjoberg et al. reviewed 5,453 articles published from 1993 to 2002 in leading software engineering journals and conference proceedings. They found a total of 113 controlled experiments, of which 20 (17.7 %) were described as replications (Sjøberg et al. 2005). Zannier et al. conducted a review to ascertain the number of articles containing empirical studies published over the first 29 years of the ICSE conference. Of a population of 1,227 articles, they retrieved 63 articles (5 %) containing empirical research. There was not one replication in this sample (Zannier et al. 2006). A recently published paper (Silva et al. 2012) found, from 1994 to 2010, 96 papers reporting 133 replications of 72 empirical studies (including not only experiments, but also case studies, surveys and others). The baseline studies were replicated on average 1.8 (133/72) times. Although we do not have data on how many software engineering empirical studies have been published in that same period, we do know that number is much greater than 72. So, it is reasonable to accept that the majority of the studies have not been replicated. Therefore, we can state that replication is not yet a regular practice in empirical software engineering.

As Fig. 1 shows, 70 % (94/133) of the published replications have been conducted by the same researchers who conducted the baseline study (called an Internal Replication Brooks et al. 1995). Of these internal replications, 64 % (60/94) were presented in papers that also reported the baseline study. In other words, the reported replications serve the purpose of checking that the results observed in the baseline study were not due to chance. This trend might be indicative that SE experimentation is maturing since results need to be observed more than once before they are taken as preliminary evidence.

There is still a clear need for researchers to publish replications independent of the baseline empirical study. The software engineering community learns a great deal from performing replications, reading reports of replications performed by others and aggregating the results of replications to draw deeper conclusions that would otherwise be possible. For experimental replications to have scientific value comparable to that of other types of empirical studies, they must be published in the peer-reviewed literature.



	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. 1 Temporal distribution of replications (Silva et al. 2012)

There have traditionally been limited opportunities to publish peer-reviewed replication papers in journals. It has been argued (Kitchenham 2008) that publishing isolated replications is hard. The aim of this special issue is to draw attention to the importance of publishing replications to advance the experimental research paradigm within software engineering and to provide four examples of such replications. Experimental disciplines replicate their empirical studies. Nature magazine tracked the fate of 19 papers in issue 6893 to check whether the results had been reproduced two years later (Giles 2006). In a large majority of cases they had. If we aim to fully apply the experimental paradigm to software engineering research, we need to give replication and its publication the relevance it deserves.

2 Replication Reporting Guidelines

Co-editor Carver performed a literature review focused on the information content and structure of the reports about replications. He found that replications were not reported in a consistent manner and the publications did not report the same type of information at the same detail level. Carver concluded by proposing a series of guidelines for reporting replications (Carver 2010). In this special issue we encouraged the use of Carver's guidelines with two aims: improve replication reporting and standardize the structure for the replications published in the issue. The goal of these guidelines is to provide replication authors with guidance regarding which information is important to include in the reporting of replication results so that the empirical software engineering community can obtain the greatest benefit from the results of the replication. To that end, the guidelines suggest four types of information to include in a replication report:

- Information about the original study to provide enough context for understanding the replication, including:
  - Research questions,
  - Participants,
  - Design,
  - Artifacts,
  - Context variables, and
  - Summary of results;

- Information about the replication to help readers understand specific important details about the replication itself, including:
  - Motivation for replication,
  - Level of interaction with original experimenters, and
  - Changes to original experiment;
- Comparison of replication results with original study results to illustrate commonalities and differences in the results obtained, including
  - Consistent results, and
  - Inconsistent results; and
- Conclusions across studies to provide readers with important insights that can be drawn from the series of studies that may not be obvious from a single study.

For the special issue, we solicited papers that made use of these guidelines. Although we did not require authors to explicitly follow the guidelines, the reviewers were asked to use the guidelines as an aid during the review process. After the review process, we surveyed both the authors and the reviewers to obtain a better understanding of the usefulness of the guidelines. The next section reports the results from the author and reviewer surveys.

3 Survey Results

We developed two survey instruments, one for paper authors and one for reviewers. While overall, the surveys were quite similar, each did contain a small number of questions targeted towards each specific audience. The remainder of this section is organized around the main topics covered in the surveys. For simplicity, we discuss the responses of the authors and the reviewers together within each topic. The results summarize the response from 18 authors and 28 reviewers. Note that not all authors or reviewers responded to the survey.

3.1 Topic 1: Usefulness of the Guidelines

Almost all authors (17/18) attempted to follow the guidelines. The primary reasons they gave: 1) the call for papers requested use of the guidelines and 2) the guidelines provided structure to the paper. The one author who did not follow the guidelines explained that s/he was including both a baseline study and a replication in the same paper, therefore the guidelines did not completely fit the situation. Of the 17 that followed the guidelines, 16 found them helpful in writing their paper and 14 found them clear to follow.

Of the 24 reviewers whose paper followed the guidelines, 16 reported that the guidelines made the paper easier to review, 2 said it they made it harder and 6 were not sure. In addition, 20 reviewers used the guidelines while reviewing their assigned paper (including 2 reviewers whose assigned paper did not follow the guidelines). Only one reviewer did not find the guidelines helpful during his/her review. In general, the reviewers liked the fact that the guidelines provided them with a checklist they could use to ensure the paper contained all relevant information about a replication. Conversely, the reviewers were less excited about papers that tried to follow the guidelines as if they were prescribing a paper outline (see Section 3.3).

3.2 Topic 2: Should Publication Venues Adopt or Enforce Guidelines?

In general, the authors and reviewers agreed that publication venues like ESEM or EMSE should adopt or enforce some type of guidelines (23/28 reviewers and 13/18 authors). This response supports a common belief about guidelines, namely that most authors and reviewers appreciate them. However, it is still an open question why authors do not follow guidelines as often as might be expected given their positive view of them. This conclusion is supported by the fact that there are so few papers that strictly follow the guidelines for publishing experiments, e.g., Jedlitschka and Pfahl (2005). It might be that the community has not yet arrived at an appropriate level of detail which make guidelines useful as a reference but do not constrict authors. Replications guidelines suffer the same awkwardness.

3.3 Topic 3: Paper Structure

Even though the guidelines did not prescribe a paper outline, there is an implicit structure that authors can extract from the guidelines. From the authors point of view, most (15/18) found the paper structure suggested by the guidelines to be clear, most (14/18) agreed with the paper structure, and most (15/18) followed the paper structure. Interestingly, only one of the authors who disagreed with the paper structure also found the structure unclear and only one author who agreed with the paper structure did not follow it. The concerns here regarded the level of detail to include about the baseline study, how to compare the replication results to the baseline results, and how to report non-standard situations (i.e., including the replication and baseline studies in the same paper). Notice that these results seem to suggest that authors are asking for more than a guideline should provide. Guidelines should not be expected to eliminate the need for authors to use their own abilities and knowledge in reporting their work.

From the reviewers point of view, 21/28 agreed with the paper structure suggested by the guidelines. Those that did not agree either stated that they did not believe the guidelines were suggesting a structure or that authors should be free to structure their paper in whatever manner they like given the content of the paper. These opinions indicate the belief, which is present in any type of reporting guidelines, that guidelines should describe content rather than prescribe structure. This belief still must be agreed upon within the empirical community regarding any type of guideline.

3.4 Topic 4: Content of Guidelines

The majority of authors (14/18) and reviewers (24/28) agreed that the guidelines were clear about what information should be included in each section of the paper. Those who disagreed primarily believed that the guidelines were intended to describe content not structure. Regarding the completeness of the guidelines, about half (15/28 reviewers and 8/18 authors) found information missing from the guidelines. From the authors point of view, most of the missing information was regarding the general reporting of experiments (e.g., introduction, context, threats to validity, ethics, etc). This comment is interesting because there are already existing guidelines for reporting experiments (as mentioned in Section 3.2). It is unclear whether the replication guidelines should repeat this information or simply reference those guidelines. The most common complaint from the reviewers was that the guidelines did not support reporting multiple replications. While this was not the intent of the guidelines, it is true that often replication reports might include more than one study. On one hand, it could be interesting to include guidance on this situation. On the other hand, the number of diverse

replication contexts makes it unreasonable to provide detailed guidance for all situations. The best way for replication guidelines to deal with this issue is still an open question.

In addition, 12/18 authors added information not in the guidelines. For example, some added additional information from the original study and other included a Threats to Validity Section. Two authors specifically excluded information because it was not in the guidelines. Furthermore only one reviewer and one author indicated that there was unnecessary information in the guidelines. The reviewer did not think there should be any guidelines at all. The author thought there was too much emphasis on the original study. Finally, four authors excluded information that was included in the guidelines. These authors all had unique situations for which the guidelines required information that was not relevant to their study (i.e., no human subjects).

We then asked whether the guidelines about specific sections (i.e., the Original Study, the Replication and the Synthesis of Results) were adequate. First, regarding the Original Study 20/28 reviewers and 13/18 authors said the guidelines were adequate. As mentioned earlier, the primary weaknesses concerned the lack of specificity in exactly which information to include from the original study (i.e., hypotheses or context variables). Second, regarding the Replication, 17/28 reviewers and 14/18 authors said the guidelines were adequate. The primary concerns here were related to how the results of the replication should be reported and what level of detail should be included. Finally, regarding the Synthesis of Results, 16/28 reviewers and 12/18 authors said the guidelines were adequate. The primary concerns here were regarding the specificity with which the synthesis process was performed, e.g., which statistical method was used.

3.5 Survey Summary

In summary, combining the results above along with the additional qualitative feedback, we can draw some conclusions about the use of guidelines to report replications. In general reviewers and authors view them positively, as long as they are not viewed as prescribing specific paper outlines or exact content. In some cases, the reviewers and authors would have liked more guidance on what to report, especially as it related to the statistics and synthesis. This observation may result from the fact that statistical aggregation methods are still not very well-known within the empirical software engineering community. Finally, due to the variety of replication types, there may be a need for some flexibility in the guidelines to accommodate different situations.

4 Summary of Included Papers

We were pleasantly surprised to discover that there were a sizeable number of replications ready to be submitted for publication. We received a total of 16 submissions and finally accepted 4 for this special issue. The large number of submissions was admittedly more than we expected. This level of response suggests that facilitating the publication of replications will encourage researchers to replicate more studies and to structure their replications according to formal guidelines. Of the four papers, three focused on studies related to verification and validation, which suggests that replications on this topic are of particular interest to the community.

We observed two commonalities across the set of four accepted papers. First, none of the replications would be considered exact replications, that is, all of them made some changes to the baseline experiment. The papers discussed the results in light of these

changes compared with the baseline study. Second, in three of the four cases, there was at least one researcher in common between the baseline experiment and the replication. In the other study the replicating researchers had face-to-face communication with the baseline researchers. That is, in all of the accepted papers, the replicating researchers had some level of interaction with the baseline researchers.

As mentioned earlier, there is not yet an agreement on terminology and type of replications. Therefore, we are unable to use a standard terminology to describe the type of the replications published in this issue. Furthermore, each of the four accepted papers use a different terminology to describe the replication they have conducted. Because our goal as editors is not to propose consistent terminology or taxonomy about replication nor is a guest editors introduction the appropriate venue for such a proposal, in the following paper summaries, we use the authors terminology and provide their explanation for the terms used.

In the first paper "On the role of tests in test-driven development: a differentiated partial replication", Fucci and Turhan focus on the claim that Test Driven Development has positive effects on external code quality and productivity. They conducted a *partial differentiated* replication. *Partial* because they only replicated part of the baseline experiment; and *differentiated* because they made some changes to the study context and measurements. The authors considered the replication to be *external* because the baseline researchers were not directly involved, although face-to-face communication did occur. In spite of the changes, consisting mainly of timing constraints and enforced development processes, the replicated study confirmed the results of the baseline study, pointing out the need to implement several tests in order to achieve higher baselines for quality and productivity.

The second paper, "Effectiveness for detecting faults within and outside the scope of testing techniques", authored by Fonseca, Espinosa, Dieste and Apa, evaluates the effectiveness of two unit testing techniques known as equivalence partitioning (EP) and branch testing (BT) to determine if the effectiveness of either technique varies according to whether the faults are visible (InScope) or not (OutScope) to the technique strategy. The authors classify their study as a *literal joint external* replication (Gómez et al. 2010a). *Literal* because the replication resembles the baseline study as closely as possible, with minor changes to eliminate one of the factors in the baseline study; *joint* because the baseline researchers participated in the replication; and *external* because the replication occurred at a different site than the baseline study. In general, the replication results were consistent with the results of the baseline study, at least in regards to the factors in common between the two studies. The elimination of a factor in the replication meant that the replicating researchers could only compare some of the results between the replication and the baseline study and could not provide confident discussions of any discrepancies observed. This situation highlights a typical problem with performing replications of software engineering studies, that is, because we know so little about the important context variables, researchers can obtain different results even if they think they have made no changes to the baseline study.

The third paper, "Are test cases needed? Replicated comparison between exploratory test-case-based software testing" authored by Itkonen and Mantyla, focuses on Exploratory Testing as a type of manual software testing. In particular the authors carried out a *dependent* replication (Shull et al. 2008) comparing effectiveness, efficiency and perceived differences between Exploratory Testing (ET) and Test-Case-Based testing (TCT) techniques. The replication is *dependent* because it was performed by the same researchers in the same context (university course) with a different sample of students (i.e., student population taking the same undergraduate course in subsequent years). The experimenters made four alterations to the baseline study, with the primary change being the elimination of a time restriction. In general, the results confirm the findings of the original study that there is no

difference between ET and TCT in terms of effectiveness; ET is more efficient; TCT generates more false positives with impact on the quality of the test cases; and it is more difficult to manage and report the testing activities in ET.

The fourth paper "A replicated quasi-experimental study on the influence of personality and team climate in software development" by Gomez and Acuna, reports an *internal* replication of a quasi-experimental study that analyzes how personality factors and team climate influence software development team effectiveness, product quality and team member satisfaction. The replication is *internal* because it was conducted by the same researchers who conducted the baseline experiment. The context variables were different (i.e., a different sample of the same population, students from a different year, degree and university) and each used a different type of development process. The results of the replication confirm some of the findings of the baseline study in terms of correlation between the extroversion personality factor and software quality, but none with team satisfaction. A significant relationship between the perception of the four team climate factors and team member satisfaction also emerged.

Acknowledgments We thank the following reviewers who provided responses to the survey: Özlem Albayrak, Danilo Caivano, Robert Feldt, Marcela Genero, Rachel Harrison, Andreas Jedlitschka, Jonathan Krein, Mika Mäntylä, Joao Paulo N. de Oliveira, Massimiliano Di Penta, Lutz Prechelt, Gregorio Robles, Barbara Russo, Alberto Sillitti, Fabio Q. B. da Silva, Martin Solari, Megan Squire, Richard Torkar, Guilherme H. Travassos, Burak Turhan, Diego Vallespir, and Murray Wood.

References

Basili V, Shull F, Lanubile F (1999) Building knowledge through families of experiments. IEEE Trans Softw Eng 25(4):456–473

Brooks A, Daly J, Roper M, Wood M (1995) Replication of experimental results in software engineering. Tech. Rep. Technical Report, RR/95/193 [EFoCS-17-95]

Brooks A, Roper M, Wood M, Daly J, Miller J (2008) Replication's role in software engineering. In: Shull F, Singer J, Sjøberg D (eds) Guide to advanced empirical software engineering. Springer London, pp 365–379

Campbell DT, Stanley JC (1963) Experimental and quasi-experimental designs for research

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

Giles J (2006) The trouble with replication. Nature 442:344–347

Gómez OS, Juristo N, Vegas S (2010a) Replication, reproduction and re-analysis: three ways for verifying experimental findings. In: RESER2010: proceedings of the 1st international workshop on replication in empirical software engineering research

Gómez OS, Juristo N, Vegas S (2010b) Replications types in experimental disciplines. In: Proceedings of the 4th ACM-IEEE international symposium on empirical software engineering and measurement. ACM, New York, pp 3:1–3:10

Jedlitschka A, Pfahl D (2005) Reporting guidelines for controlled experiments in software engineering. In: Proceedings of the 4th international symposium on empirical software engineering, pp 95–104

Juristo N, Vegas S (2009) Using differences among replications of software engineering experiments to gain knowledge. In: Proceedings of the 3rd international symposium on empirical software engineering and measurement. IEEE Computer Society, Washington, pp 356–366

Kitchenham B (2008) The role of replications in empirical software engineering–a word of warning. Empir Softw Eng 13(2):219–221

Krein JL, Knutson CD (2010) A case for replication: synthesizing research methodologies in software engineering. In: RESER2010: proceedings of the 1st international workshop on replication in empirical software engineering research

Moonesinghe R, Khoury MJ, Janssens ACJW (2007) Most published research findings are false but a little replication goes a long way. PLoS Med 4(2):e28

Schmidt S (2009) Shall we really do it again? the powerful concept of replication is neglected in the social sciences. Rev Gen Psychol 13(2):90–100

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: Proceedings of the 1st international symposium on empirical software engineering. IEEE Computer Society, Washington, pp 7–16

Shull FJ, Carver JC, Vegas S, Juristo N (2008) The role of replications in empirical software engineering. Empir Softw Eng 13(2):211–218

Silva FQ, Suassuna M, Frana ACC, Grubb AM, Gouveia TB, Monteiro CV, Santos IE (2012) Replication of empirical studies in software engineering research: a systematic mapping study. Empir Softw Eng 1–57

Sjøberg D, Hannay J, Hansen O, Kampenes V, Karahasanovic A, Liborg NK, Rekdal A (2005) A survey of controlled experiments in software engineering. IEEE Trans Softw Eng 31(9):733–753

Vegas S, Juristo N, Moreno A, Solari M, Letelier P (2006) Analysis of the influence of communication between researchers on experiment replication. In: Proceedings of the 5th ACM/IEEE international symposium on empirical software engineering. ACM, New York, pp 28–37

Zannier C, Melnik G, Maurer F (2006) On the success of empirical studies in the international conference on software engineering. In: Proceedings of the 28th international conference on software engineering. ACM, New York, pp 341–350