Replication Types: Towards a Shared Taxonomy

Maria Teresa Baldassarre
University of Bari - DIB
via E.Orabona 4, Bari Italy
mariateresa.baldassarre@uniba.it

Jeffrey Carver
Department of Computer Science,
University of Alabama,
Tuscaloosa, AL USA
carver@cs.ua.edu

Oscar Dieste, Natalia Juristo
Facultad de Informatica, Universidad
Politecnica de Madrid
Madrid, Spain
[natalia, odieste]@fi.upm.es

ABSTRACT

Context: The software engineering community is becoming more aware of the need for experimental replications. In spite of the importance of this topic, there is still much inconsistency in the terminology used to describe replications.

Objective: Understand the perspectives of empirical researchers about various terms used to characterize replications and propose a consistent taxonomy of terms.

Method: A survey followed by plenary discussion during the 2013 International Software Engineering Research Network meeting.

Results: We propose a taxonomy which consolidates the disparate terminology. This taxonomy had a high level of agreement among workshop attendees.

Conclusion: Consistent terminology is important for any field to progress. This work is the first step in that direction. Additional study and discussion is still necessary.

1. INTRODUCTION

Experiments need to be replicated in different contexts, at different times, and under different conditions before they can produce generalizable knowledge. In general, one can define a replication as: "the repetition of an experiment, either as closely following the original experiment as possible, or with a deliberate change to one or several of the original experiment's parameters, in order to achieve, or ensure, greater validity in software engineering research" [1]. Recently the software engineering community has been embracing replications more readily, as shown by recent relevant literature [5, 6, 11, 12]. Da Silva et al.'s review indicated that reporting of replications began in 1995 and has been increasing since 2004 [12].

There are two primary motivations for conducting replications that may be drawn from scientific and industrial needs. The first motivation is that different types of replications are necessary to solve relevant problems and collect evidence. Empirical investigations gain credibility when they are replicated and reproduce generalizable results. External replications are also important for confirming and providing evidence to software engineering principles. From an industrial perspective, replications are valuable because they provide reliable information about the cost-benefit ratio of a software engineering practice which derives from replications on various types of projects. Industrial stakeholders can use this information to support adoption decisions.

Even with the increased interest in replications, there is no agreement yet on terminology, typology, purposes, operation and other replication issues. A 2008 point/counterpoint column series on replications published in the *Empirical Software Engineering* journal [8, 13] provides a good example of divergent viewpoints. The lack of consistent terminology allows different authors to use different definitions for the same term and different terms to refer to the same concept when characterizing their studies. The inconsistent terminology makes it more difficult to consistently understand how these replications fit into the overall software engineering landscape.

Various communities have identified and discussed the need for a consolidated terminology. For example, the Workshop on Replication in Empirical Software Engineering (RESER) has been conducted in 2010, 2011, and 2013. In addition, during the 2011 and 2013 International Software Engineering Research Network (ISERN) meetings, we have organized sessions about replications. This paper reports on the results of a survey and subsequent discussion from the 2013 ISERN workshop. Finally, there is an upcoming special issue of *Empirical Software Engineering* focused on replications in which the terminology issues were again evident.

The contribution and goal of this work is to: illustrate and comment the results of the ISERN session and formulate a proposal for a consistent taxonomy of replication types that researchers can agree upon for reporting replications.

Section 2 describes some open issues about replications and current classifications from the literature. Section 3 reports the results of the survey along with the proposed taxonomy. Finally, Section 4 concludes the paper and outlines future work.

2. RELATED WORK

A review of the software engineering literature shows that researchers have used several definitions and terms to classify and characterize replications. This confusion of terminology has made it difficult to classify a replication consistently. In this study we have used the most relevant terms and definitions from the software engineering literature, as shown in Table 1. The remainder of this section discusses these definitions in a bit more details to provide the context from which they were defined.

Table 1: Replication terminology in literature

Term	Definition
Close [10]	Conditions are kept similar to the original
Differentiated [10]	Deliberate variations in conditions of the study
Internal [3]	Same researchers
External [3]	Different researchers
Same hypothesis [2]	No changes in the experiment
Different hypothesis [2]	Dependent, independent or context variables change
Extend theory [2]	Aim of the study is to extend the theory it relies on
Similar – external [1]	No changes to the study carried out by others
Improved-internal [1]	Same experimenters change the study
Similar-internal [1]	Same experimenters with no changes to the study
Differentiated-external [1]	Changes to the study carried out by others
Exact [13]	Same research question, reuse the original procedures
Conceptual [13]	Same research question, different experimental procedure, different researchers
Dependent-strict [9]	Study replicated as the original
Dependent-differentiated [9]	Study intentionally changes some aspects
Independent [9]	Same research question. Experimenters are not aware of the previous study

There are two important factors that underlie these definitions. Each definition uses one or both of these factors to characterize the type of replication. The factors are: 1) procedure, i.e. the steps followed in the study and 2) people, i.e. the experimenters conducting the replication.

2.1 Procedure

With regards to the procedure, most definitions focus on how closely the replicating experimenters follow the steps of the original study. First, the studies that stay as close to the original procedures as possible have been labeled as **close** (i.e. keep all known conditions the same or similar) [10], **exact-dependent** (i.e. procedures are followed as closely as possible) [13], **strict-dependent** (i.e. study builds on a previous study and follows it as closely as possible) [9]. This type of replication is also described by Basili et al., but is not given a specific name [2]. Second, there are studies in which the replicating researchers consciously change the procedures of the original study. These types of replications are called **differentiated** (i.e. deliberate or known variations in the materials, methods or subjects of the study) [10], **differentiated-dependent** (i.e. study builds on a previous study but intentionally alters aspects to validated conclusions) [9] or **exact-independent** (i.e. replication varies one or more major aspect of the experimental conditions) [13]. This type of replication is also described by Basili et al as intentionally varying the research hypotheses or variables [2]. Finally, there are studies that seek to replicate the results by designing a replication without relying upon the original study design. These replications are called **independent** (i.e. addresses the same questions or hypotheses, but conducted without knowledge of the previous study) [9] or **conceptual** (i.e. same research question with different experimental procedures carried out by different researchers). Basili et al describe this type of replication as trying to extend the theory [2].

2.2 People

With regards to the people, Brooks et al distinguish between **internal** (i.e. same researchers) and **external** (i.e. different researchers) replications [3]. With this definition it is unclear how to characterize a study when there is only a partial overlap among the original and replicating researchers.

2.3 People and Procedure

Finally, some researchers include both factors in their characterizations of replications. Almqvist [1] characterizes replications as: **similar-external** (i.e. close replication performed by other experimenters), **improved-internal** (i.e. same experimenters carry out the experiment under different conditions, in different settings or with modified tasks), **similar-internal** (i.e. close replication performed by the same researchers), and **differentiated-external** (other experimenters carry out an experiment under different conditions). Shull et al label the last type as **conceptual** (i.e. same research question with different procedure and different experimenters [13].

3. TOWARDS A TAXONOMY

ISERN is an international community of approximately forty academic and industrial members who are experts in the theory and practice of various types of empirical studies. During the annual meeting, the members exchange ideas and form working groups to advance the practice of empirical software engineering. During the 2013 edition of ISERN, the authors of this paper organized a session on Replications. We organized the session in two parts.

The first part of the session provided background information from each participant regarding his or her experiences about:

Q1. The topics of experiments they executed;

Q2. The type and number of replications they have conducted;

Q3. For the replications, what changes they made to the original experiment; and

Q4. The types of interactions between original experimenters and the replicators;

The second part of the session focused on the definitions of replication terminology. We gave each attendee a worksheet with the following 11 terms: Internal, External, Exact, Dependent, Independent, Conceptual, Differentiated, Strict, Improved, Similar, and Close. We asked each attendee to individually provide his or her own definition of each term and to add any terms that were omitted. We also asked them to identify any terms that they thought referred to the same concept. After completing the form individually, the participants met with two other workshop attendees to discuss their answers and develop an agreed-upon set of definitions.

After completing the exercise, we conducted a large group discussion. This discussion was helpful in understanding the perspectives of the workshop attendees and helped us to draw appropriate conclusions from the gathered data. The remainder of this section describes the results.

3.1 Demographic Results

Table 2 provides the results for Q1 and Q2 regarding the topics on which the attendees studies have focused. Most studies focused on software inspections and testing followed by requirements and modeling.

Table 2: Topics participants have replicated on

Topic	Percentage distribution
Software inspections and testing	47%
Software requirements and modeling	22%
Conway's Law	9.5%
Agile development	8.6%
Elicitation	8.6%
Software maintenance	4.3%

Regarding their practice when conducting replications, the answers to Q3 and Q4 indicate that few participants (15%) made slight changes to the original experiment. In particular they changed: instrumentation such as programs used or software platform adopted; time schedule provided to experimental subjects; subjects from students to experts or vice versa. However most participants (85%) indicated that they made no changes to the original experiment when conducting their replication. Finally in most cases (78%) the researchers conducting the replication were the same as those who conducted the original experiment (or least had a large overlap). This trend was also reported in a recent literature study where about 70% of the replications identified were classified as having been carried out by the same researchers as the initial study [12].

3.2 Definition Results

There was general agreement among workshop attendees on all of the terms provided on the data collection form. No one added any new terms to the list. These results suggest that, even though there are multiple terms in the literature, the participants were familiar with their definitions as an answer was provided in almost all cases. The most frequently used terms to describe replications were: Internal, External, Close and Differentiated.

Based on these results, Table 3 shows our proposal for a consistent set of terms to describe replications that consolidate terms with similar definitions. The taxonomy in Table 3 summarizes the results of the analysis from the individual and group survey forms. In particular, we have specified the term that participants suggested should be used, a definition for the term, and whether there are similarities with other terms used in literature. The last column reports the percentage of the participants that agreed on the term and its definition.

Table 3: Proposal of Taxonomy

Term to use	Similarity with other definitions in literature	Definition	Level of agreement on the term and definition
Internal	Dependent	Same experimenters (or most of the original ones) carry out the replication.	100%
External	Independent	Different experimenters (or most of them are different from the original group of experimenters) carry out the replication	100%
Close	Exact, Strict, similar	Established that a software engineering experiment cannot be replicated exactly as the original (due to its nature) this type of replication is carried out as close as possible to the original study: design, hypothesis, context, measurements remain the same.	90%
Differentiated	Improved	Some changes are intentionally made to the original experiment: design, hypothesis, context, measurements,	100%
Conceptual	-	Everything in the experimental design setting is different. Only the research question or hypothesis is the same.	80% (note that only 65% of the participants provided an answer. Others left it blank)

For some terms there was a high level of agreement among the workshop participants. However, as it appears from a more detailed analysis though, there were cases with controversial answers:

- In the case of *close* replications, a small number of participants thought they should be called *differentiated* or *improved* (i.e. a study that involved some changes to the original design).
- The participants preferred the term *differentiated* to *improved* because, in their opinion a change in the design does not necessarily mean an improvement.
- In spite of the agreement on the meaning, the definition of *conceptual* was only reported on 65% of the forms. Several comments questioned on the need for this type of replication and whether it applies to software engineering. This result could arise from the fact that

this type of replication study is not common among the community. Indeed, the results of the background survey support this conclusion as most replications were close or internal. For completeness we have included this term in the proposal because we consider it important for the community to discuss this type of study further before deciding whether to exclude it.

We would like to suggest that researchers conform to this proposed taxonomy when reporting a replication [4]. The plenary discussion carried out after having collected single and group survey forms pointed out that when a researcher plans a replication it is important to consider aspects like: who is replicating the study, what differences are being introduced, why the study is being replicated. Therefore, the replication type ends up necessarily being a combination of the terms listed in the table, depending on the characteristics of the study.

4. CONCLUSIONS AND FUTURE WORK

This paper discusses the importance of having a consistent taxonomy for reporting replications to reduce the confusion and wide range of terminology currently present in literature. To this end, we presented the results of a survey conducted during an ISERN session on replications. The results pointed out that the participants, all knowledgeable in the area of empirical software engineering, find the need for a consistent taxonomy to follow.

The plenary discussion among the participants pointed out that any taxonomy that is proposed must take into account elements such as: *who* is carrying out the study (researchers/institutions, same as previous, different than the original study); *why*, i.e. the motivation for replicating the study; what, i.e. *what* conditions are being changed (variables, context, materials, etc.). Therefore, these three axes must be included in any set of terms used to characterize replications.

It was interesting for the authors of this paper to note that the issue of taxonomies faced in the ISERN session was also raised during the RESER workshop by another group of researchers who were not part of ISERN [14, 15]. This consistent observation supports the importance of the topic and the need for guidance to researchers in the field to assure that they all consistently move in the same direction. We look forward to collaborations with other researchers to extend this preliminary study and survey results to identify a consolidated taxonomy of replication types that can be shared with the entire empirical software engineering community.

5. ACKNOWLEDGMENTS

Our thanks go to all the members of the ISERN community that have contributed to the Replications Session during the ISERN 2013 meeting held in Baltimore, USA.

6. REFERENCES

[1] Almqvist J.P.F, Replication of controlled experiments in empirical software engineering – a survey. Master thesis, Department of computer science, faculty of science, Lund University, Sweden, 2006.

[2] Victor R. Basili, Forrest J. Shull, and Filippo Lanubile. Building knowledge through families of experiments. IEEE Transactions on Software Engineering , 25(4):456–473, July/ August 1999. doi:10.1109/ 32.799939.

[3] Brooks A., Daly J.W., Miller J., Roper M., Wood M. Replication's role in experimental computer science. Technical Report EFoCS-5-941 [RR/ 94/ 172], Department of Computer Science, University of Strathclyde, Glasgow, Scotland, UK, 1995.

[4] Carver J, Towards reporting guidelines for experimental replications: a proposal. In: RESER'2010: proceedings of the 1st international workshop on replication in empirical software engineering research. 2010

[5] Carver J, Juristo N., Baldassarre M.T., Vegas S., Replications of Software Engineering Experiments - Editorial, Empirical Software Engineering Journal, DOI 10.1007/s10664-013-9290-8, online 05.12.2013

[6] Gomez OS, Juristo N, Vegas S, 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, 2010

[7] International Software Engineering Research Network - ISERN, http://isern.iese.de/Portal/

[8] Kitchenham B, The role of replications in empirical software engineering–a word of warning. EmpirSoftw Eng 13(2):219–221, 2008

[9] Krein JL, Knutson CD, 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. 2010

[10] Lindsay R.M., Ehrenberg A.S.C. The design of replicated studies. American Statistician , 47(3):217–228, August 1993

[11] RESER Workshop series, http://sequoia.cs.byu.edu/lab/?page=reser2013

[12] 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. Empirical Software Engineering Journal. DOI 10.1007/s10664-012-9227-7, pp 1–57.

[13] Shull F, Carver J, Vegas S, Juristo N. The role of replications in empirical software engineering. Empirical software engineering 13(2) pp.211-218, 2008.

[14] de Magalhaes C.V, da Silva FQ, Towards a taxonomy of replications in empirical software engineering research: a research proposal, In: proceedings of the 3rd international workshop on replication in empirical software engineering research. 2013, pp.50-55, DOI: 10.1109/RESER.2013.10

[15] de Magalhães CV, da Silva FQ, Santos R. A preliminary analysis of conceptual studies about replication of empirical studies in software engineering, to appear in Proceedings of the 18th international conference on Evaluation and Assessment in Software Engineering, London UK, 2014.