Information and Software Technology 56 (2014) 1033–1048

Contents lists available at ScienceDirect

Information and Software Technology

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

Understanding replication of experiments in software engineering: A classification

Omar S. Gómez [a,*], Natalia Juristo [b,c], Sira Vegas [b]

[a] Facultad de Matemáticas, Universidad Autónoma de Yucatán, 97119 Mérida, Yucatán, Mexico
[b] Facultad de Informática, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain
[c] Department of Information Processing Science, University of Oulu, Oulu, Finland

ARTICLE INFO

Article history:
Received 22 August 2013
Received in revised form 3 April 2014
Accepted 3 April 2014
Available online 13 April 2014

Keywords:
Software engineering
Experimental software engineering
Experimentation
Replication

ABSTRACT

Context: Replication plays an important role in experimental disciplines. There are still many uncertainties about how to proceed with replications of SE experiments. Should replicators reuse the baseline experiment materials? How much liaison should there be among the original and replicating experimenters, if any? What elements of the experimental configuration can be changed for the experiment to be considered a replication rather than a new experiment?
Objective: To improve our understanding of SE experiment replication, in this work we propose a classification which is intend to provide experimenters with guidance about what types of replication they can perform.
Method: The research approach followed is structured according to the following activities: (1) a literature review of experiment replication in SE and in other disciplines, (2) identification of typical elements that compose an experimental configuration, (3) identification of different replications purposes and (4) development of a classification of experiment replications for SE.
Results: We propose a classification of replications which provides experimenters in SE with guidance about what changes can they make in a replication and, based on these, what verification purposes such a replication can serve. The proposed classification helped to accommodate opposing views within a broader framework, it is capable of accounting for less similar replications to more similar ones regarding the baseline experiment.
Conclusion: The aim of replication is to verify results, but different types of replication serve special verification purposes and afford different degrees of change. Each replication type helps to discover particular experimental conditions that might influence the results. The proposed classification can be used to identify changes in a replication and, based on these, understand the level of verification.

1. Introduction

Experimentation is an essential part of SE research. "[In SE] Experimentation can help build a reliable base of knowledge and thus reduce uncertainty about which theories, methods, and tools are adequate" [68]. Replication is at the heart of the experimental paradigm [61] and is considered to be the cornerstone of scientific knowledge [53].

To consolidate a body of knowledge built upon evidence, experimental results have to be extensively verified. Experiments need replication at other times and under other conditions before they can produce an established piece of knowledge [13]. Several replications need to be run to strengthen the evidence.

Most SE experiments have not been replicated. Sjøberg et al. [66] reviewed 5453 articles published in different SE-related journals and conference proceedings. They found a total of 113 controlled experiments, of which 20 (17.7%) are described as replications. Silva et al. [65] have conducted a systematic review of SE replications. They found 96 papers reporting 133 replications of 72 original studies run from 1994 to 2010.

If an experiment is not replicated, there is no way to distinguish whether results were produced by chance (the observed event occurred accidentally), results are artifactual (the event occurred

* Corresponding author. Address: Anillo Periférico Norte, Tablaje Cat. 13615, (Room EA-7) Colonia Chuburna Inn, Mérida, Yucatán, Mexico. Tel.: +52 (999) 942 31 40x1117.
E-mail addresses: omar.gomez@uady.mx (O.S. Gómez), natalia@fi.upm.es (N. Juristo), svegas@fi.upm.es (S. Vegas).

http://dx.doi.org/10.1016/j.infsof.2014.04.004
0950-5849/

because of the experimental configuration but does not exist in reality) or results conform to a pattern existing in reality. Different replication types help to clarify which of these three types of results an experiment yields.

Most aspects are unknown when we start to study a phenomenon experimentally. Even the tiniest change in a replication can lead to inexplicable differences in the results. For immature experimental knowledge, the first step is replications closely following the baseline experiment to find out which experimental conditions should be controlled [10]. As Collins [16] explained for experiments in physics, "the less that is known about an area the more power a very similar positive experiment has to confirm the initial result. This is because, in the absence of a well worked out set of crucial variables, any change in the experiment configuration, however trivial in appearance, may well entail invisible but significant changes in conditions". For mature knowledge, the experimental conditions that influence results are better understood and artifactual results might be identified by running less similar replications. By using different experimental protocols, it is possible to check whether the results correspond to experiment-independent events. "As more becomes known about an area however, the confirmatory power of similar-looking experiments becomes less." [16]

The immaturity of experimental SE knowledge has been an obstacle to replication. Context differences usually oblige SE experimenters to adapt experiments for replication. As key experimental conditions are yet unknown, slight changes in replications have led to differences in the results which prevent verification. Attempts at combining replication results (Hayes [26], Miller [49–51], Hannay et al. [25], Jørgensen [35], Pickard et al. [55], Shull et al. [62], Juristo et al. [32]) have reported that it was not possible to verify results because of differences in experimental conditions.

There is no agreement in SE about what a replication is in terms of how many changes can be made to the baseline experiment and the purpose of such changes (as we will see in Section 2).

A classification of replications for SE may help form a better understanding of the particular verification purpose of each type of replication and what changes are valid for each type.

This paper is organized as follows. Section 2 discusses replication classifications proposed in SE. Section 3 describes different types of replication proposed in other disciplines. Section 4 outlines the research method that we have followed. The remainder of the paper reports each step of the research method. Section 5 describes the elements of an experimental configuration in SE. Section 6 introduces what specific verification purposes a replication can have. Section 7 describes a classification of replication types for SE experiments. Section 8 discusses the advantages of systematic changes in replications. Section 9 compares our proposal with other SE classifications proposed in the literature. Section 10 discusses researcher positions on SE replications. Finally, Section 11 presents the conclusions.

2. Related work

We have not found any research that specifically aims to classify replications in experimental SE. We have identified three works that have classified replications as part of the research.

Basili et al. [5] present a framework for organizing sets of related studies. They describe different aspects of the framework. One framework aspect defines a three-category classification of replications: (1) replications that do not vary any research hypothesis, (2) replications that vary the research hypotheses and (3) replications that extend the theory.

Basili et al. [5] identify two replication types that do not vary any research hypothesis:

- Strict replications, which duplicate as accurately as possible the original experiment.
- Replications that vary the manner in which the experiment is run. These studies seek to increase confidence in experimental results. To do this, they test the same hypotheses as previous experiments, but alter the details of the experiments in order to address certain internal threats to validity.

They identify three replication types that vary the research hypotheses:

- Replications that vary variables which are intrinsic to the object of study. These replications investigate what aspects of the process are important by systematically changing intrinsic properties of the process and examining the results.
- Replications that vary variables which are intrinsic to the focus of the evaluation. They may change the ways in which effectiveness is measured in order to understand the dimensions of a task for which a process results in most gain. For example, a replication might use a different effectiveness measure.
- Replications that vary context variables in the environment in which the solution is evaluated. They can identify potentially important environmental factors that affect the results of the process under investigation and thus help understand its external validity.

Replications that extend the theory are not further sub-divided. These replications help determine the limits to the effectiveness of a process by making big changes to the process, product, and context models to see if basic principles still hold.

In his master thesis, Almqvist [2] studies the use of controlled experiment replication in SE. He surveys 44 articles describing 51 controlled experiments and 31 replications. Categories are defined to organize the identified experiments. One of the categories develops a classification for pigeonholing the identified replications. As a reference, Almqvist [2] takes the concept of close and differentiated replication described in the accounting area by Lindsay and Ehrenberg [41] (depending on whether the replication attempts to keep almost all the known conditions of the study much the same or very similar at least, or have deliberate variations in fairly major aspects of the conditions of the study), to which he adds the internal and external replications used by Brooks et al. [11] (depending on whether the replication is run by the same experimenters or independently by other experimenters). Based on these classifications, Almqvist [2] defines the following four replication types:

1. Similar-Internal Replications.
2. Improved-Internal Replications.
3. Similar-External Replications.
4. Differentiated-External Replications.

Krein and Knutson [39] propose a unifying framework for organizing research methods in SE. They build a taxonomy of replications as part of such framework. The taxonomy defines four types of replication:

- Strict replication. An experiment that is meant to replicate a prior study as precisely as possible.
- Differentiated replication. An experiment that intentionally alters aspects of the prior study in order to test the limits of that study's conclusions.
- Dependent replication. A study that is specifically designed with reference to one or more previous studies, and is, therefore, intended to be a replication study.