Towards Reporting Guidelines for Experimental Replications: A Proposal

Jeffrey C. Carver
University of Alabama
Box 870290
Tuscaloosa, AL
+1-205-348-9829

carver@cs.ua.edu

ABSTRACT

The value of experimental replications has been well established. In order for the replicating researcher and the community to receive the greatest benefit from a replication, the right information about it must be published. This paper proposes publishing guidelines to increase the value of experimental replications. First, a review of some published replications highlights the variation in current publishing practice. Then, a set of guidelines are proposed. The goal of this paper is to provide a starting point for a discussion that will formalize and publish a set of guidelines.

Categories and Subject Descriptors

D.2.m [**Software Engineering**]: Miscellaneous

General Terms

Experimentation

Keywords

Reporting Guidelines

1. INTRODUCTION

The value of experimental replications is evident to the participants of this workshop. 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. To facilitate the usefulness of these publications, we need guidelines to ensure that a consistent set of information is published about each replication.

There are existing guidelines for reporting controlled experiments [6] and case studies [15], but none specifically for reporting experimental replications. The type of report required for an experimental replication is similar to, but is not the same as that for a controlled experiment. In a replication it is important to publish information about the original study, the context of the replication, any changes made, and the results. It is not always clear how to balance these various types of information within a replication paper. In this paper, I put forth an initial proposal of reporting guidelines for experimental replications with the goal of standardizing how replications are reported in the literature. This proposal is meant to begin a discussion that will result in formalized reporting guidelines.

While there is general agreement on the need for conducting replications, there are a variety of definitions of replications. While the goal of this paper is not to provide a definition of a replication, it is important to mention a few words about what a replication is. Recently two opposing viewpoints concerning what constitutes a valid replication appeared in the Empirical Software Engineering journal [9, 17]. A major difference in the viewpoints taken by these two papers regards the level of interaction between replicating researchers and the original researchers. Without going through the whole debate here, there are legitimate issues on both sides. To be comprehensive, the proposed guidelines provide a place to discuss this attribute.

The remainder of the paper is organized as follows. Section 2 discusses on how existing replications are reported in the literature. Section 3 proposes the new reporting guidelines. Section 4 provides some conclusions.

2. PUBLISHED REPLICATIONS

As a starting point for the proposed guidelines, I performed a small literature review. This review focused on replications that were published in the International Symposium on Empirical Software Engineering and in *Empirical Software Engineering: An International Journal*, the main conference and journal of the empirical software engineering community. While there are replications published in other venues, I focused my review on these venues under the assumption that the replication papers published there would be the most complete and consistent because the empirical software engineering community is the most experienced at performing and publishing experiments and replications. Section 2.1 discusses the process of identifying the papers included in the review. Section 2.2 illustrates the different approaches these papers took in discussing the original study. Section 2.3 focuses on how the papers compare the results of the replication with the results of the original study. Finally, Section

RESER'10, May 4, 2010, Cape Town, South Africa.

2.4 presents some of the issues that arise when a replication is part of a series or 'family' of studies.

2.1 Replication papers reviewed

To identify these replication papers, I performed a search using the terms "replication" and "replicated." After reviewing the results to ensure that the papers identified were actually experimental replications, I identified 15 papers that are the focus of the remainder of this section [1-5, 7, 8, 10-14, 16, 18, 19]. In general, the replications were not reported in a consistent manner. Each author used his or her own organization for their replication paper. In addition, the papers were not consistent in either the type of information reported or the level of detail reported.

The goal of this section is not to speak negatively of the published replications. After all, there are no existing guidelines and each author published the information that he or she deemed to be most important for their purposes. Conversely, the replications are discussed merely to illustrate the inconsistency with which replications are currently being published and to identify the types of information published by various authors, as an input to the proposed guidelines in Section 3.

2.2 Description of original study

To provide context for the replication, the replication paper must discuss the original study upon which the replication was based. The reviewed papers were not consistent in their reporting of information about the original study. In reviewing these papers, three approaches for discussing the original experiment emerged. First, some authors fully describe the original study in its own section near the beginning of the paper [2, 8, 10, 14, 18, 19]. Second, some authors provide a brief summary of the original study and its results early in the paper (but not in a separate section) [1, 5, 13]. Third, some authors do not provide a separate discussion of the original study. Rather, they merge the description of the original study with the description of the replication (only referring to the original study when a change was made for the replication) [3, 4, 7, 11, 12, 16].

A second issue is which information about the original study is reported. Even when different authors use the same approach to discuss the original study, they often do not include the same details. The superset of information reported by the replication papers reviewed includes: the goal of the original study, the experimental context, the study design, the subjects, the tasks, the hypotheses, the variables, a summary of results, and a detailed description of the results. The lack of consistency in how authors report this information suggests the need for clear guidance about which information should be published in a replication paper.

For both of the above issues, one confounding factor is whether the report is published in a conference or in a journal. Given that conference papers are shorter than journal papers, the guidelines should provide recommendations for each type of venue. This factor is addressed in Section 3.3.

2.3 Comparison of replication results with original study results

One of the main benefits of an experimental replication is that it provides researchers with the ability to confirm, refute, or deepen the conclusions drawn from an earlier study. In order to draw such conclusions, the results of the replication must be compared with the results of the original study. In reviewing the published replications, I found that there was a lack of consistency in how these comparisons were presented. There appear to be three methods used to compare the results of the replication with the results of the original study. These methods are not mutually exclusive, as some authors used multiple methods in the same paper. In addition, even when using the same method, some authors provided a large amount of detail, while others provided only a brief discussion.

The first method is to integrate the comparison of results throughout the paper as each replication result is analyzed and discussed [1, 7, 10, 11, 19]. The second method is to create a separate section solely focused on the comparison of results [2, 10, 12, 16, 18]. The third approach is to make a comparison of the results in the conclusion of the paper [1, 3-5, 7, 8, 13, 14]. In addition, one interesting method use by two studies was to present the results of the replication and the results of the original study in a summary table for easy reference [11, 19]. Finally, in one study, the authors conducted a formal meta-analysis to compare the results of the replication with those of the original study [12].

2.4 Families of replications

One interesting observation while searching for replications was that 40% (6/15) [5, 10, 11, 13, 14, 16] of the papers identified were related to a similar topic, i.e. comparing the effectiveness of checklists and various scenario-based reading techniques for supporting a requirements inspection. These replications were not all based on the same original study, but they did investigate related questions. This family of replications is interesting because even though they are all related, the papers are organized differently and include different information at different levels of detail. Specifically, some of the papers did not review the whole family of previous studies in detail. It appears that there may be some additional guidelines necessary for reporting a replication that within a family of replications than are necessary for stand-alone replications.

3. PROPOSED REPORTING GUIDELINES

In this section, I discuss the information that should be provided in the report of an experimental replication. In Section 3.1, I detail each type of information and why it should be included. In Section 3.2, I discuss what information should be included when reporting a replication that is part of a family of replications. In Section 3.3, I discuss how to adapt this proposal for a shorter conference paper which may not allow space for reporting all of the information.

3.1 What information to report

Based on the review of published replications and the understanding of the goals of an experimental replication, I propose that following items should be included in any report describing an experimental replication.

3.1.1 Information about the original study

To help the reader understand the replication, a replication paper needs to discuss some information about the original study. Authors should provide enough information about the original study to allow the reader to properly interpret and understand the replication without providing so much detail that the reader is distracted from the main goal of the replication paper. A replication paper report should provide the following information about the original study (at a minimum):

- *Research question(s)* – a description of the research question(s) that was the basis for the design,
- *Participants* –the number of participants and any relevant characteristics of the participants,
- *Design* – a graphical (or textual) description of the experimental design,
- *Artifacts* – a description of and/or links to the artifacts used,
- *Context variables* – any important context variables that affected the design of the study or interpretation of the results, and
- *Summary of the results* – a brief overview of the major findings.

3.1.2 Information about the replication

As with any experiment, the basic information about the study should be reported. This section focuses on the specific information that needs to be reported about a replication. A replication report should contain the following information (at a minimum):

- *Motivation for conducting the replication* – a description of why the replication was conducted (e.g. to validate the results, to broaden the results by changing the participant pool or the artifacts).
- *Level of interaction with original experimenters* – If the replication is external (i.e. the original researchers are not involved), the level of interaction the replicators had with the original experimenter should be reported. This interaction could range from none (i.e. simply read the paper) to a lot (i.e. original experimenter acted as consultants). If a lab package is used, then its use should be described. There has also been some discussion within the community about the acceptable level of interaction between replicators and original experimenters [9, 17]. These guidelines do not address that controversy; rather they provide a mechanism for reporting the level of interaction.
- *Changes to the original experiment* – Any changes made to the design, participants, artifacts, procedures, data collected and/or analysis techniques should be discussed along with the motivation for the change.

3.1.3 Comparison of results to original

One of the main values of a replication is the comparison of its results with the results of the original study. As was noted in Section 2.3, there is not a consistent approach to presenting this information. It is reasonable to expect that authors will embed brief comparisons of results throughout the presentation of the replications results. To make the comparison explicit, it is also important to have a section specifically devoted to comparing the results of the replication with the results of the original study. This section should highlight the following information:

- *Consistent results* – replication results that supported results from the original study, and
- *Differences in results* – results from the replication that did not coincide with the results from the original study. Authors should also discuss how changes made to the experimental design (Section 3.1.2) may have caused these differences.

3.1.4 Drawing conclusions across studies

Finally, pulling together information about the original study (Section 3.1.1), changes made for the replication (Section 3.1.2) and the comparison of results (Section 3.1.3), the authors should provide a discussion of the current state of knowledge. By combining conclusions from the original study with conclusions from the replication, the authors should be able to provide insights that would not have been evident from either study individually. In this section authors should highlight any conclusions of the original study that were strengthened. This section is also the place to propose hypotheses about new context variables that may have become evident through the analysis of multiple studies.

3.2 Reporting a replication within a family

While most replications may be isolated replications (i.e. the only replication of a study), the community really begins to gain deeper knowledge when multiple researchers replicate the same study in different contexts. A series of replications can be called a ‘family’ because insight can be gained by analyzing the results from all studies. While the guidelines presented previously apply to all types of replications, some special guidelines are in order for a replication that is part of a family.

To place the replication in the proper context within the family, it is important for the report to provide a relatively brief summary of the previous studies and replications. This summary should include information about how the studies were related, conclusions drawn and current state of knowledge about the topic. Then, it is important for the author to clearly motivate why the current replication was performed along with any changes that were made. Finally, when discussing the results of the replication, it is important for the author to place the results into the context of the entire family of studies. Conclusion should be drawn based on knowledge gained by analyzing the results of all studies.

3.3 Adaptations for shorter papers

When the publishing target for the replication paper is a venue that has a shorter page limit (e.g. a conference), it may not be possible to provide all of the details described in Section 3.1. Conversely, in order for the replication to make sense within its context, this information is necessary. The recommendation is to provide all of the information, but with less detail. Comparisons between the original study and the replication can be summarized in tables rather than fully described in the text. If information must be omitted, authors should ensure that the most important details, i.e. those details that relate to changes or differences between the original study and the replication, should be included in the report.

4. SUMMARY

This paper presents an initial proposal of reporting guidelines for publishing experimental replications. Section 2 reviewed a number of published replications to compare and contrast both the format and the content of the replication papers. With this information as a background, Section 3 proposes guidelines that should be used when reporting a replication. It is important to ensure that replications that are part of a family receive special treatment. In addition, Section 3.3 makes some suggestions about how to deal with publication in venues with shorter page limits.

The goal of this paper is to begin a discussion about standardizing the publication of experimental replications. The community can gain a great deal of knowledge from these replications. It is important to ensure that they are reported in such a way that the greatest benefit can be obtained. My hope is that this paper will begin a discussion that can result in a fully specified set of guidelines that will be publishable in *Empirical Software Engineering: An International Journal*.

5. REFERENCES

[1] Abrahao, S., Insfran, E., Gravino, C. and Scanniello, G. On the effectiveness of dynamic modeling in UML: Results from an external replication. In Anonymous *ESEM '09: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement.* (). IEEE Computer Society, Washington, DC, USA, 2009, 468-472.

[2] Andersson, C. A replicated empirical study of a selection method for software reliability growth models. Empirical Software Engineering, 12, 2 (04/01/ 2007), 161-182.

[3] Cox, K. and Phalp, K. Replicating the CREWS Use Case Authoring Guidelines Experiment. Empirical Software Engineering, 5, 3 (11/01/ 2000), 245-267.

[4] Dias-Neto, A. C. and Travassos, G. H. Evaluation of model-based testing techniques selection approaches: An external replication. In Anonymous *ESEM '09: Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement.* (). IEEE Computer Society, Washington, DC, USA, 2009, 269-278.

[5] Fusaro, P., Lanubile, F. and Visaggio, G. A Replicated Experiment to Assess Requirements Inspection Techniques. Empirical Software Engineering, 2, 1 (03/01/ 1997), 39-57.

[6] Jedlitschka, A. and Pfahl, D. Reporting Guidelines for Controlled Experiments in Software Engineering. In Shull, F., Singer, J. and Sjøberg, D. I. K. eds.*Guide to Advanced Empirical Software Engineering.* Springer, , 2008, 201-228.

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

[8] Kiper, J. D., Auernheimer, B. and Ames, C. K. Visual Depiction of Decision Statements: What is Best for Programmers and Non-Programmers? Empirical Software Engineering, 2, 4 (12/01/ 1997), 361-379.

[9] Kitchenham, B. A. The role of replications in empirical software engineering—a word of warning. Empirical Software Engineering, 13, 2 (04/01/ 2008), 219-221.

[10] Maldonado, J., Carver, J., Shull, F., Fabbri, S., Doria, E., Martimiano, L., Mendonca, M. and Basili, V. Perspective-Based Reading: A Replicated Experiment Focused on Individual Reviewer Effectiveness. Empirical Software Engineering, 11, 1 (2006), 119-142.

[11] Miller, J., Wood, M. and Roper, M. Further experiences with scenarios and checklists [software inspection]. Empirical Software Engineering, 3, 1 (1998), 37-64.

[12] Pfahl, D., Laitenberger, O., Dorsch, J. and Ruhe, G. An Externally Replicated Experiment for Evaluating the Learning Effectiveness of Using Simulations in Software Project Management Education. Empirical Software Engineering, 8, 4 (12/01/ 2003), 367-395.

[13] Porter, A., Votta, L. and Basili, V. R. Comparing Detection Methods for Software Requirements Inspections: A Replication Using Professional Subjects. Empirical Software Engineering: An International Journal, 3, 4 (1998), 355-379.

[14] Regnell, B., Runeson, P. and Thelin, T. Are the perspectives really different? Further experimentation on scenario-based reading of requirements. Empirical Software Engineering, 5, 4 (2000), 331-56.

[15] Runeson, P. and Höst, M. Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering, 14, 2 (04/01/ 2009), 131-164.

[16] Sandahl, K., Blomkvist, O., Karlsson, J., Krysander, C., Lindvall, M. and Ohlsson, N. An extended replication of an experiment for assessing methods for software requirements inspections. Empirical Software Engineering, 3, 4 (1998), 327-54.

[17] Shull, F., Carver, J., Vegas, S. and Juristo, N. The Role of Replications in Empirical Software Engineering. Empirical Software Engineering, 13, 2 (2008), 211-218.

[18] Vokáč, M., Tichy, W., Sjøberg, D. I. K., Arisholm, E. and Aldrin, M. 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 (09/01/ 2004), 149-195.

[19] Wesslén, A. A Replicated Empirical Study of the Impact of the Methods in the PSP on Individual Engineers. Empirical Software Engineering, 5, 2 (06/01/ 2000), 93-123.