Packaging Researcher Experience to Assist Replication of Experiments

V. Basili, S. Green, O. Laitenberger, F. Lanubile, F. Shull,
S. Sorumgard, M. Zelkowitz

Abstract

The external validity of an empirical investigation can be incrementally strengthened by replications performed by independent researchers. External replications can be promoted by the availability of experimental kits which package the experience of the original researchers. We present the lab manual of a controlled experiment for evaluating the effectiveness of a software reading technique known as Perspective-Based Reading (PBR). The lab manual includes the experimental design, the instruments for its execution, and the procedures for statistical tests. We have made the lab manual available over the World Wide Web. We hope to use the experiences we have gained through this process to evolve a lab package that meets the needs of software engineering researchers, to develop a consistent method for packaging experiments, and to facilitate the replication of experiments.

1. Introduction

Research in computer science often involves the development of new systems, methods and techniques in order to solve problems. However, once developed there exists little experimental data to demonstrate the usefulness of the new approach. In many cases no comparison is made between a new development and the state of the practice.

The only way to discover how applicable a new method, technique, or tool is in a given environment is to experiment with its use in that environment. Generalizations about the object of study cannot be made until the experiment or case study has been replicated in a number of different environments. Yet there are a number of difficulties that have hindered such replicated experiments in our field. In order to address these difficulties and encourage replication of a particular experiment we have developed a lab package. The package is designed to facilitate replication of an experiment for evaluating the effectiveness of Perspective-Based Reading (PBR) [BGLLSSZ96]. We refer to it throughout this paper as a concrete example to illustrate our discussion of essential elements of a useful lab package.

The lab package has three major objectives:

1. Description of the particular technique under study (PBR).
2. Documentation of the way a particular controlled experiment (comparing PBR against another reading technique) was prepared, conducted and analyzed in order to discuss various issues in experimentation.
3. Explanation of the experiment to make it understandable and repeatable for other researchers.

To make the replication more attractive to other researchers, we have developed an experimental design that contains both generic and domain-specific sections. That is, we ask researchers to customize parts of the experiment. While the generic portion is meant to achieve a basis of comparison with other replications of this experiment, the customizability is intended to allow results that are directly applicable in the researcher's experimental environment. How to best strike a balance between these two aims is a matter that deserves some debate.

In order to provide greater availability, the complete lab package is accessible over the World Wide Web.
(http://www.cs.umd.edu/projects/SoftEng/ESEG/manual/manual.html)

2. Contents of an Example Lab Package

The remainder of this paper uses each section of the lab manual to motivate discussion of larger issues in packaging researcher experience for the replication of experiments.

2.1. Introduction to the Experiment

This section of the package provides the overview necessary to understand the background and context of the experiment. First, the research framework is described in terms of an instantiation of the Quality Improvement Paradigm [Br88]. The second subsection provides a discussion of the differences between reading techniques and other verification methods such as reviews, walk-throughs, inspections, and audits.

In the next subsection we present a classification system for reading techniques. For example, the PBR technique which is being evaluated is systematic (i.e., the reader knows how to review the document), specific (i.e., the reader is focused to look for a specific defect class), and distinct (i.e., there is minimal overlap among several readers). We also provide in this section a description of other reading techniques and the previous research that has already been focused on them.

Since the technique against which PBR is evaluated is one of the facets that should be customized, by providing these subsections we hope to make people think harder about the technique that they supply for comparison. In particular, we find it beneficial if the technique can be placed within a common research framework, and described using a common classification scheme. However, on future lab packages, stronger measures should be taken to ensure a consistent terminology.

Finally, we provide a more in-depth description of the PBR technique itself. This is necessary both to convince other researchers of the importance of this series of experiments, and to allow them sufficient background to be comfortable teaching the new technique to the subjects of their experiment.

The idea behind Perspective-Based Reading is that various customers of a product should read the document from a particular point of view. For PBR we use the different roles within the software development process as a model, with each role representing a different point of view. We chose the roles of tester, designer and user. To support the reader throughout the reading process we have developed operational descriptions (called scenarios) for each role. We provide these scenarios as part of the lab package.

2.2. Experimental Design

This section addresses the main factors in the experimental design: constraints on the participants, documentation to be reviewed, time available for review, and place for conducting the experiment. The design schema of the experiment is also described.

The design which we present has evolved through our experiences with a pilot study of this experiment, and has been successfully used in a full run of the experiment in 1995. We specify it at a level of detail such that it can be used "as is" by other researchers, or serve as the basis for modifications which adapt it to a particular set of constraints. However, we acknowledge that the further the design is modified the less chance there is that a meaningful comparison can be made between replications. This is a topic that is deserving of serious consideration, since there exist few practical guidelines that assist the researcher in determining how much of a change is permitted between two experimental conditions before the results are no longer usefully comparable.

2.3. Analysis

In this section we provide some details on the statistical tests which we used in analyzing our data. Copies of the programs we have written to analyze our own data are provided as a model which can be adapted to the needs of other experimenters. The analysis is composed of two main parts: testing team performance and testing individual performance.

The former, testing team performance, describes how we applied a permutation test in order to validate the performance of PBR as compared to the usual NASA technique when the participants are combined into teams. Since our subjects read on an individual basis, team performance must be simulated during the analysis. Since the permutation test has not been extensively used in software engineering literature, we present a detailed motivation for the test, along with both the algorithm and a C program which performs the computation.

The latter part, testing individual performance, describes how we applied the analysis of variance in order to test the effect of PBR (in comparison to the usual technique) on individual detection rates. We describe the necessary analysis, given our experimental design. We illustrate the approach to solving this statistical problem through a program, included in the package, which we have written for use with the SAS statistical package.

2.4. Experimental Materials

This section contains the various materials necessary for running the experiment. All of them can be downloaded in postscript format. Most can be used as is, though a few may have to be altered if used in a replication of the experiment - these are provided as examples only. Suggested schedules are provided for each day of the experiment. Each schedule lists the packets of experimental materials that will be necessary for each activity (presentation, training, or review) in the experiment. The materials include documents which are used in the experiment itself (e.g. sample documents, copies of forms, slides used in training) as well as additional resources for the experimenter (e.g. resources for evaluating subjects' work, technical reports).

2.5. Experiences and Change History

This section contains a list of replications of this experiment (currently, two) and the experiences that have been gained from them. A history of the changes made from replication to replication is also included. By collecting this information in one place, we intend to strengthen the ability to draw conclusions about the effectiveness of PBR in multiple environments, and also to learn how an experiment must be tailored to different environments.

This section represents one of the most interesting open questions in developing an effective lab package. The ongoing research into software metrics reflects the fact that the relevant parameters and effects in software development are not yet completely understood. It is therefore difficult to know exactly how to specify the experimental context of each replication, in order to allow for meaningful comparisons between them.

Currently, we summarize context in terms of several factors (e.g. participant experience and the setting in which the experiment took place) as well as asking researchers to specify how they customized the experiment (e.g. the reading technique against which PBR was compared, and a brief description of the domain-specific documents used). We note that other changes may be introduced by experimenters, but should be documented. For example, while we provide three PBR perspectives with this lab package, others may be developed and tested in other environments.

2.6. Bibliography

Finally, we provide a list of references which can be useful to other researchers intending to replicate the empirical investigation of Perspective-Based Reading.

References

[BR88] V. R. Basili, and H. D. Rombach, The TAME Project: Towards Improvement-Oriented Software Environments, *IEEE Trans. on Softw. Eng.*, Vol.SE-14, No.6 (1988), pp.758-773.

[BGLLSSZ96] V. R. Basili, S. Green , O. Laitenberger, F. Lanubile, F. Shull, S. Sorumgard, and M. Zelkowitz, The Empirical Investigation of Perspective-Based Reading, International Software Engineering Research Network, Technical Report ISERN-96-06, 1996.