The Core Evaluation of a Software Engineering Research Result Must Be Empirical in Nature

Tayfun Elmas and Xiaoyu Chen

Abstract

This paper discusses the role of empirical evaluation in software engineering (SE) research and defends empirical approaches as the core part of the evaluation process for research results. We first describe SE research and the evaluation process with its important aspects. We then focus on empirical evaluation and highlight its crucial role in assessing SE research results.

1 Software engineering research

Research in SE aims to achieve two main goals: 1) To increase the knowledge about what are useful in the SE discipline; 2) To introduce cost effective solutions to problems in SE. To demonstrate the results of SE research from different points of view, a set of questions can be used, ranging from how superior the new techniques are compared with the previous ones, to how they improve the software process [1].

2 Evaluation of software engineering research

The evaluation process for SE research assesses how much the initial claim of the research is achieved and how well the related research questions have been addressed. The quality of these answers is determined by how the evaluation is conducted and is important for assessing the reliability of research results. The evaluation methods depend on the type of a specific study and the claim it makes, varying from theory or method development to system analysis. However, from the SE point of view, the essential part of the evaluation is to convince people the reliability and the effectiveness of the claim, which is mostly related to the usability of the research results in future SE research studies and SE practices. We believe that the ability to evaluate the research results is a main factor in distinguishing good SE research from bad SE research.

3 Empirical evaluations

3.1 Empirical studies in software engineering

The aim of empirical studies in SE is to provide a scientific and thus more rational basis for understanding, evaluating, predicting, controlling, and improving tools, methods, and techniques used in SE [3]. Empirical studies collect and analyze observations about theories, models, and systems, based on instances from either the real world or the models of the real world [6]. Controlled experiments and case studies are examples of empirical studies. Among all empirical studies, empirical evaluation especially focuses on providing evidences that support research results.

3.2 The purposes of empirical evaluation

As any other evaluation method, empirical evaluation depends on the initial claim of the research. It assesses not only the success of the research based on the claim but also the effect of the claim on its applications. Moreover, empirical evaluation obtains a generalized prevision of how successful SE research results would be in the real-world practices. The generality of empirical evaluation is based on statistical theory of sampling, which is widely accepted by many other science disciplines as well. There are statistical guidelines on collecting data and assessing the confidence of the experimental results. For some specific set of problems, sampling is even moved further by having standard benchmarks. Finally, empirical evaluation of the existing techniques may serve as the motivation of developing new hypotheses and techniques.

3.3 The importance of empirical evaluation

First of all, conducting experiments is a general approach to validate any idea in sciences [2]. It has been well known that software is complex, invisible, and difficult to visualize [5]. Due to such essential properties of software, it is hard to assess the success of an existing SE technique in different domains; it is even harder to predict the effect of a new SE technique on real-world applications.

On one hand, non-empirical evaluation methods may be hard to find or even not exist in many cases, where deriving a sufficiently precise model of the problem space is difficult. To prove SE research results in a theoretical manner, we need make assumptions to reduce the complexity and build a simplified model. However, big assumptions are usually problematic

when applied to the real world, and it is hard to map a simplified model to the complex real world. For example, a research work on software analysis claims cost reduction and uses asymptotic bounds in its evaluation. Such theoretical evaluation may not be strong enough for the SE community, because asymptotic bounds are not sufficient to convince people that the method will achieve the asserted cost reduction in real programs. This is due to the difficulty in mapping the evaluation model to the real applications.

On the other hand, based on statistical methods such as sampling, empirical evaluation can access the full complexity of the real world, and give a realistic view of SE research results. Following the claim that the success of SE research is usually determined by its application to the real world, empirical evaluation is especially important in the field of SE. As shown in paper [2], SE research papers do have a higher percentage of empirical work than those from other areas of computer science.

Therefore, based on the nature of software and the goals of SE research, we claim that the core evaluation of a software research must be empirical.

4 Discussion

Notice that we do not claim that empirical evaluation is the only evaluation method for a SE research result. In fact, some theoretical studies may focus on SE problems in a small problem space such as type-checking, and non-empirical evaluation like proofs may fit well. However, the applications of such research results will eventually be in a complex world with a big problem space. In order to assess the effect of those research results on their applications, empirical evaluation need to be performed with a global view of the problem space in the real world.

In addition to above discussion, a specific research work in SE might be hypothetical or theoretical, and it may contain no empirical evaluation by itself. However, in the long term, a research result of impact in the field must have been empirically evaluated by the researchers or the practicers who follow up the research. This indicates that, either in the short term or in the long term, empirical evaluation is eventually the core part in assessing the reliability and value of a research result.

In the real world, the cost of applying a SE research result could be high, and the failure of the application may not be affordable. Therefore, practicers in industry may think that it is risky

to employ new techniques or methods without enough experimental support. This fact again highlights the importance of convincing arguments about a research result, which are usually provided by empirical evaluation.

5 Questions supporting our claim

The following questions can be raised against the claim that empirical evaluation does not have to be the core evaluation for SE research results:

- Given the complexity of the real world in which SE is applied, what non-empirical evaluation methods could be used to evaluate research in SE? How reliable are they?
- Given a SE research study with claims applied to general problems (e.g. a new source code analysis algorithm), suppose there is a non-empirical evaluation for the claim (e.g. the asymptotic runtime bound for the algorithm). How could one map the simplified model to the whole problem space of SE and predict the algorithm will give desirable results in the general case?
- Without empirical evaluation, how can we distinguish good SE research from bad SE research in the terms of real-world applications?
- Given a theoretical study that is not appropriate to be applied directly in practices but need to be improved by subsequent studies. Without performing empirical evaluation on the study, how could we be convinced that the future studies will lead to success in practices?
- When should a SE research result be considered mature enough to be applied in practices (i.e. in real SE development projects)? What are the possible non-empirical criteria for the usability of the research result?

References

[1] Shaw, M. 2003. Writing good software engineering research papers: minitutorial. In Proceedings of the 25th international Conference on Software Engineering (Portland, Oregon, May 03 - 10, 2003). IEEE Computer Society, Washington, DC, 726-736.

[2] P. Lukowicz, E. Heinz, L. Prechelt, and W. Tichy: "Experimental evaluation in computer science: A quantitative study". Technical Report, 17/94, Department of Informatics, University of Karlsruhe, 1994.

[3] V. R. Basili, R. W. Selby, and D. H. Hutchens, "Experimentation in Software Engineering". IEEE Transactions on Software Engineering, vol. SE-12, pp. 733-742, 1986.

[4] hilip Johnson, Readings in Empirical Evaluation for Budding Software Engineering Researchers. Collaborative Software Development Laboratory. University of Hawaii. http://csdl.ics.hawaii.edu/techreports/05-06/05-06.html

[5] Frederick P. Brooks, "No Silver Bullet: Essence and Accidents of Software Engineering". Computer, April 1987.

[6] R. Jeffrey, L. Scott. “Has twenty five years of empirical software engineering made a difference”. Proc. Asia–Pacific Softw. Eng. Conf., 2002, pp 539-546.