How Reviewers Think About Internal and External Validity in Empirical Software Engineering

Janet Siegmund* Norbert Siegmund† Sven Apel‡

Abstract: Empirical methods have grown common in software engineering, but there is no consensus on how to apply them properly. Is practical relevance key? Do in-ternally valid studies have any value? Should we replicate more to address the trade-off between internal and external validity? We asked the key players of software-engineering research, but they do not agree on answers to these questions.

The original paper has been published at the International Conference on Software Engineering 2015 [SSA15]. Empirical research in software engineering came a long way. From being received as a niche science, the awareness of its importance has increased. In 2005, empirical studies were found in about 2% of papers of major venues and conferences, while in recent years, almost all papers of ICSE, ESEC/FSE, and EMSE reported some kind of empirical evaluation, as we found in a literature review. Thus, the amount of empirically investigated claims has increased considerably.

With the rising awareness and usage of empirical studies, the question of where to go with empirical software-engineering research is also emerging. New programming languages, techniques, and paradigms, new tool support to improve debugging and testing, new visualizations to present information emerge almost daily, and claims regarding their merits need to be evaluated—otherwise, they remain claims. But, how should new approaches be evaluated? Do we want observations that we can fully explain, but with a limited generalizability, or do we want results that are applicable to a variety of circumstances, but where we cannot reliably explain underlying factors and relationships? In other words, do researchers focus on internal validity and control every aspect of the experiment setting, so that differences in the outcome can only be caused by the newly introduced technique? Or, do they focus on external validity and observe their technique in the wild, showing a real-world effect, but without knowing which factors actually caused the observed difference?

This tradeoff between internal and external validity is inherent in empirical research. Due to the options' different objectives, we cannot choose both. Deciding for one of these options is not easy, and existing guidelines are too general to assist in making this decision.

With our work, we want to raise the awareness of this problem: *How should we address the tradeoff between internal or external validity?* In the end, every time we are planning an experiment, we must ask ourselves: Do we ask the right questions? Do we want pure,

*University of Passau, siegmunj@fim.uni-passau.de
†University of Passau, siegmunn@fim.uni-passau.de
‡University of Passau, apel@uni-passau.de

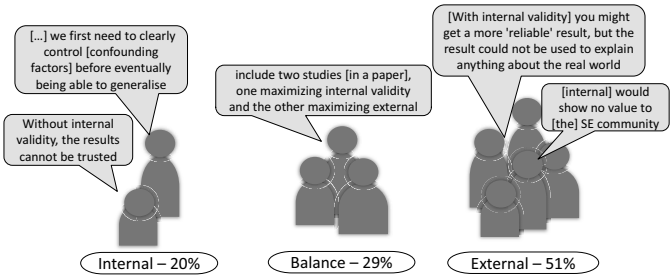

Fig. 1: Preferences for internal vs. external validity among program-committee and editorial-board members.

ground research, or applied research with immediate practical relevance? Is there even a way to design studies such that we can answer both kinds of questions at the same time, or is there no way around replications (i.e., exactly repeated studies or studies that deviate from the original study design only in a few, well-selected factors) in software-engineering research?

To understand how the key players of software-engineering research would address this problem, we conducted a survey among the program-committee members of the major software-engineering venues of the recent years [SSA15]. In essence, we found that there is no agreement and that the opinions of the key players differ considerably (illustrated in Fig. 1). Even worse, we also found a profound lack of awareness regarding the tradeoff between internal and external validity, such that one reviewer would reject a paper that maximizes internal validity, because it "[w]ould show no value at all to SE community". When we asked about replication, many program-committee members admitted that we need more replication in software-engineering research, but also indicated that replications have a difficult stand. One reviewer even states that replications are "a good example of hunting for publications just for the sake of publishing. Come on."

If the key players cannot agree on how to address the tradeoff between internal and external validity (or even do not see this tradeoff), and admit that replication—a well-established technique in other disciplines—would have almost no success in software-engineering research, how should we move forward? In the original paper, we shed light on this question, give insights on the participants' responses, and make suggestions on how we can address the tradeoff between internal and external validity.

References

[SSA15] Janet Siegmund, Norbert Siegmund, and Sven Apel. Views on Internal and External Validity in Empirical Software Engineering. In *Proc. Int'l Conf. Software Engineering (ICSE)*, pages 9–19. IEEE CS, 2015.