From Craft to Science: The Road Ahead for Empirical Software Engineering Research

Matthias Galster
University of Canterbury
Christchurch, New Zealand
mgalster@ieee.org

Danny Weyns
KU Leuven
Leuven, Belgium
danny.weyns@kuleuven.be

Antony Tang
Swinburne University of Technology
Hawthorn, Australia
atang@swin.edu.au

Rick Kazman
University of Hawaii
Honolulu, USA
kazman@hawaii.edu

Mehdi Mirakhorli
Rochester Institute of Technology
Rochester, USA
mehdi@se.rit.edu

ABSTRACT

Empirical software engineering (SE) research is often criticized for poorly designed and reported studies, a lack of replications to build up bodies of knowledge, and little practical relevance. In this paper, we discuss issues in empirical software architecture research as an illustration of these issues in one subfield of SE and as a step towards better understanding empirical research in SE in general. Based on feedback from software architecture researchers and practitioners, we explore why, despite persistent discussions in the SE research community, there are still disagreements about why and how to conduct empirical research. Then, we explore how empirical SE research can progress beyond "one-off" studies and endless "new and exciting" results toward SE research as a mature science. This would allow us to establish foundations for evaluating existing and future empirical research and help researchers design and publish better studies.

CCS CONCEPTS

• **Software and its engineering** → *Software architectures*;
• **General and reference** → ***Empirical studies;***

KEYWORDS

Empirical research, software engineering

ACM Reference format:

M. Galster, D. Weyns, A. Tang, R. Kazman, and M. Mirakhorli. 2018. From Craft to Science: The Road Ahead for Empirical Software Engineering Research. In *Proceedings of 40th International Conference on Software Engineering: New Ideas and Emerging Results Track, Gothenburg, Sweden, May 27-June 3 2018 (ICSE-NIER'18)*, 4 pages.
https://doi.org/10.1145/3183399.3183421

ICSE-NIER'18, May 27-June 3 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5662-6/18/05/$15.00
https://doi.org/10.1145/3183399.3183421

1 INTRODUCTION

Empirical research serves two main goals: 1) to gain well-founded insights about phenomena (e.g., in exploratory studies); 2) to obtain evidence for the validity of new solution proposals (e.g., in evaluation studies). Recently, the software engineering (SE) community has started to critically reflect on the meaning of science and the role of empiricism [11], the relevance of research [7], the gap between research and practice and the worthiness of SE research [10], and the balance between relevance (i.e., problems are identified from and solved for substantial problems of practice) and rigor (i.e., scientific, and often time-consuming, approaches to solving problems are followed). On the other hand, we currently lack an understanding of *why* and *how* those who conduct and evaluate empirical SE research have certain beliefs and perceptions about their research approach. This has led to a lack of consensus on why and how to conduct empirical research and how to build up evidence [9]. Besides hindering the execution of empirical research, this also affects reproducibility and the willingness of researchers to go beyond "one-off" studies and "new and exciting" results. Therefore, the goal of this paper is to break through existing complaints and anecdotal discussions about empirical research in SE and to explore innovative paths to push the field forward towards a mature and relevant science.

Key related work: Siegmund et al. investigated views on empirical research in SE in general [9]. However, applying empirical research in specific SE fields can lead to issues that researchers overlook or even disregard. Therefore, this paper identifies issues specifically in the context of empirical software architecture (SA) research. We chose SA as an important and established subfield of SE since we wanted to obtain a clear and focused view of the problems in one particular field. Such a focused view is relevant as we aim at understanding *why* problems are persistent and *what steps* could be taken to move empirical SE research forward. The insights that emerge from this investigation can then provide a foundation for further investigations and validations in SE in general. Furthermore, SA is one of the most challenging fields in which to conduct empirical research. It relies on diverse concepts related to computer science, modelling, human factors, technologies, etc. and advances are often driven by "talented people and industrial experience" [3]. The dearth of empirical research standards in SA can lead to research approaches that lack formal foundations which affect their relevance and significance to industry. Also, empirical studies in SA often lack details in their description, raising concerns about the reliability of the results, or at least introduce difficulties in replicating and extending them [4].

New and emerging insights: 1) Novel insights about the views on empirical SA research (and how they generalize to SE) based on the perceptions of those who conduct and evaluate research; insights go beyond current insights that show that practitioners do not consume SE research (e.g., [6]). 2) A list of problems that indicate why, despite ongoing discussions about how to conduct empirical research in SE, issues still exist. 3) A set of innovative recommendations to break through the status quo to advance the state of empirical research in SE and in particular in SA.

2 PERSISTENT PROBLEMS

2.1 Method to Identify Problems

To understand perceptions of SA researchers on empirical research, we sent an online questionnaire to 455 PC members of all editions of major SA conferences (CBSE, ECSA, ICSA, QoSA, WICSA). These are experienced actors who shape architecture research by reviewing papers, advising emerging researchers and guiding funding agencies. We received 105 valid responses (i.e., a response rate of ~23%). Twelve percent answered as practitioners. All respondents are also reviewers for top journals and conferences in the broader field of SE and not only SA. 72% have spent more than 10 years in academia and only 7% never spent time in academia (and 22% of respondents never spent time in industry). Only 12% reported that they had never reviewed an empirical study. The questionnaire included closed questions, most of them with the option to provide comments (and from these comments we learned that many responses were not specific to SA, but that respondents often reflected on SE in general). Data were analyzed using descriptive statistics and open coding of textual comments. Percentages reported below refer to the full sample size of 105 respondents. Below we also provide representative quotes to highlight our findings.

2.2 Identified Problems

Problem 1: Disagreement about the role and value of different research methods. Empirical research methods can be categorized into qualitative and quantitative [11]. In our study, 60% of respondents indicated no preference of any type of method over the other. However, 25% preferred quantitative over qualitative methods. As one respondent stated, "*I like the clear nature of quantitative research.*" Others stated, "*Where a quantitative method is possible, valid and realistic, I like to see the data to understand how strong the [...] claims are.*" and "*The only thing that always gets me furious are researchers that 'abuse' one of the methods to claim they showed* [something] *that is out of reach for the specific method.*"

From a researcher's point of view, when reporting on conducting their own studies, 23% believe quantitative research is easier to get accepted while 42% of respondents do not think that way. From the comments we learned that respondents consider quantitative studies easier to conduct/write up, and believe that quantitative studies require less effort compared to qualitative studies. Also, respondents acknowledged that acceptance and appreciation of empirical work depends on the venue. One respondent stated that "*I don't prefer* [quantitative studies] *as a researcher; but I believe that it is easier to get them accepted because it looks more like hard results and it is work to prove they are wrong – I believe many reviewers don't like to spend that effort.*"

When it comes to evaluating empirical studies conducted by others (in the role of reviewers), we found that 21% believe quantitative studies are easier to review, while 50% of respondents do not believe that quantitative studies are easier to review than qualitative studies. These numbers do not correlate with whether or not respondents had actually ever reviewed empirical studies. Based on the textual comments, we found that respondents consider the effort differently for reviewing qualitative and quantitative studies. Also, as reviewers, respondents judge the value of qualitative and quantitative studies differently.

Problem 2: Replications are appreciated but not conducted. One key characteristic of science is reproducibility and replications of studies [1]. 61% of the respondents (strongly) agree that replications could advance SA. However, 76% of respondents had never reviewed a replication and only 17% had seen one or two. Based on comments we found that those who agree also consider replications at the heart of science, e.g., "*I think that replication is a fundamental element of empirical studies.*" On the other hand, those who disagree are mostly doubtful about the (added) value of replications as shown in statements by respondents, e.g., "*Highly doubtful that any useful question in* [SA] *can be addressed in a replicated experiment.*" Similarly, 55% of respondents (strongly) disagree that replicated studies do not have a place in top SA venues. From the comments provided we found that respondents appreciate the difficulties of replications, e.g., "*If someone is willing to go to the trouble of doing a replication, we should pay attention. My guess is that architecture studies are more susceptible to validity threats than most, and you'll get terrible replication results.*" Finally, some of those who agree indicate that reviewers are simply not trained to properly judge

replications. The numbers above do not correlate with whether or not respondents had ever reviewed a replication (76% of respondents had never reviewed a replication). The perceptions of the respondents reveal a paradox since, as reviewers, respondents appreciate replications and agree that we need more; but as authors, respondents are worried about submitting replications. This paradox is even more noteworthy since our survey focused on a subfield of SE where the ones who review research are the ones who conduct research.

Problem 3: Despite ongoing discussions, some researchers still lack appreciation for (and training to judge) empirical work. 40% of respondents indicated that their views on empirical research have changed over time. This number does not depend on how much time a respondent has spent in industry or academia or for how many different venues respondents were reviewers. From textual comments we extracted how views changed: Those who did not change their view were nearly always in favor of empirical research. Those who changed their view can be separated in those who gained appreciation for empirical research and those who lost appreciation. Most comments indicated an increasing appreciation. As one respondent stated, "*You have to become knowledgeable on systematic studies to appreciate them.*" Another one said that "*I learnt more empirical methods, i.e., I also changed my reviews over time. I often see reviews of others which obviously have not learnt empirical methods.*" On the other hand, we also found that appreciation decreased in the view of some respondents. To quote one respondent, "*I am now more biased against such work, because it is usually so contrived. A nice experiment, but a "who cares?" result.*"

Problem 4: There is frustration towards so-called empirical research. By definition, empirical evidence is grounded on and verified by experiences and observations. Whilst many researchers claim that they *do* empirical research, there seems to be an implicit interpretation of what empirical *means*. While some acknowledge the importance of empirical works, they expressed their frustration of what others consider "empirical". As a respondent stated, "*Quantitative stuff is usually bogus. But a lot of it gets published*!" Other respondents expressed frustration about the lack of practical relevance: "*People who do this empirical research should pair up with a person doing *real* work so that they can understand when they are writing useless drivel*" and "*I build real stuff. People use it. Does that qualify according to your criteria? I doubt it.*" Another respondent stated, "[SA] *is much more related to people, the largely automated studies making conclusions about things in software repositories* […] *are just producing plainly wrong/arbitrary results in design/architecture.*"

3 WAYS FORWARD

The problems presented above highlight fundamental mismatches and contradictions in the views of SA and SE researchers. This is alarming, giving that respondents were active researchers and practitioners in the field. As such, these insights can be used as considerations to challenge the status quo and to move forward in empirical research. Below we make four recommendations.

3.1 Different Approaches to Empirical Work

The wide variety of contexts in various SE subfields could be a reason why there are many debates on why and how empirical SE research should be conducted. We found that 20% of respondents did consider empirical research in SA to be different to empirical research in the other fields of SE. 57% considered SA not different (and 23% responded "I don't know"). Based on the respondent comments, we identified three categories of differences: 1) Study context (and objects under study): As one respondent stated, "*the goals, tasks, capabilities, and artifacts in software architecture research differ strongly from other software engineering sub-fields and need an adaptation of the empirical approaches that have worked well in another sub-field of software engineering*". This broader context was also expressed by a respondent who stated that it is "*more difficult to quantify or compare than in other disciplines*" and another one who said that "*trying to have quantitative data about a primarily qualitative discipline* […] *where there is no one right answer is difficult.*" 2) Scope of studies (and types of questions tackled): As stated by a respondent, "… *software architecture is relevant only for large scale problems* […] – [SE] *can also relate to smaller problems/systems.*" Another one stated that "*Good questions in* [SA] *are much bigger and harder than 'reasonable' questions in, say, testing.*" 3) Information sources (and required research subjects): As stated by a respondent, it "… *requires more experience than other areas of software engineering which might require higher involvement of professionals.*" and "*few* [software development projects] *have to tackle significant architectural issues. So the range of possible real subjects is much more narrow.*" These three types of differences show that empirical research approaches in different SE fields can differ. Whether the perceived need for different approaches to empirical work applies to other subfields requires further investigation.

3.2 Towards a Charter for Empirical Research

Members of the research community in principle agree that we need empirical research. However, based on disagreements about empirical methods, the replications paradox and skepticism on empirical research, we suggest that there is a need to clarify what empirical research *is* and how we can ensure that it is done appropriately. Foremost, we propose to aim for an alignment and agreement about the role of empirical work. For that reason, we propose to establish a charter for empirical research. If accepted across SE communities, such charter would help to: a) evaluate the strengths of the empirical method used to solve a research problem, b) evaluate the strengths of the empirical data (and evidence), c) evaluate how well the results solve or explain a defined problem, d) assess the scientific relevance (e.g., fundamental research to understand phenomena), and e) evaluate how close a research problem resembles practice. The charter could define the principles of empirical research, the guidelines to conducting empirical research, conditions and research methods used in replications, etc. More than 50% of our respondents were not sure whether there are any good examples of empirical SA research. An agreed and enforced charter could increase the

accountability for the research we do and serve as a governing instrument, e.g., by providing guidance in debates on whether a theory is right or wrong (which seldom occurs), or by encouraging conferences and journals to explicitly commit to a particular percentage of empirical papers, replications, etc. (dedicated tracks may not be as efficient as these may be considered lower impact contributions). We envision that different parties could then subscribe to the charter, including individual researchers, conferences, journals, funding agencies, educational programs, etc. This goes beyond current initiatives, such as ISERN, which focus on sub-communities interested in empirical research.

3.3 Classification of Empirical Research

We suggest that empirical research contributions are classified, e.g., based on Redwine-Riddle's maturity model [8], which distinguishes basic research, concept formulation, development and extension, enhancement and exploration, popularization. This would help clarify the target of research works, e.g., researchers (e.g., in the case of descriptive studies and basic research) or industry (e.g., in the case of solution proposals). To evaluate research contributions, we put forward six criteria: relevance, practicality, novelty, beneficiaries, cost and benefits, and rigor. Depending on the type of research, criteria can be weighted differently. As part of classifying empirical research, we also need to reflect on the meaning of generalizability in SE research. Is generalizability overrated and should empirical SE research aim for context-driven research [2]? Classification guidelines could be part of a charter for empirical research.

3.4 Training and "Certification" Initiatives

As we have found, the more empirical research is submitted (reviewed, and published), the more likely appreciation for it will increase. However, SE current curricula focus on transferring knowledge and obtaining skills. Education of how such knowledge is built and how useful skills can be evaluated is, to a large extent, missing. The Software Engineering Body of Knowledge (SWEBOK) for example highlights only a few aspects of empirical methods under "engineering foundations." However, these foundations lack methodological aspects and are not well covered in university curricula. In other disciplines (e.g., psychology) knowledge transfer and learning how knowledge is built up are balanced. SE programs may apply a similar approach.

Also, the research community could apply some best practices for new researchers (e.g., conduct a replication in the first year of a PhD). Furthermore, the community may consider "certification programs" organized at main conferences (e.g., with topics such as basics of scientific methods, principles of quantitative and qualitative methods, etc.). These programs could become instruments to assign key roles in our community (e.g., a PC chair requires certain certifications). Training/certification initiatives could again become part of a charter for empirical research.

4 DISCUSSION AND CONCLUSIONS

We identified a set of persistent problems that hamper the maturation of empirical research in SA and SE. From these problems we identified recommendations to pave the way to a more mature scientific discipline so that empirical SE research can serve its two main goals: gain well founded insights about phenomena; obtain evidence for the validity of new solution proposals. Where do we go from here? First, we need to explore other fields within SE in more detail. Second, we need to bring findings from the different fields of SE together so they can learn from each other. For example, some conferences have already rethought how research is conducted, e.g., by introducing technical, scientific and empirical tracks, and by establishing concrete (and enforced) criteria to evaluate papers. Third, we need to implement the aforementioned path and in particular the charter for empirical research. We suggest a step-wise bottom-up approach where we first open the debate in SE sub-communities and based on the outcome open a broad debate in the general SE community (e.g., through dedicated sessions at major venues).

We need to acknowledge that we cannot overcome all issues. Several "corollaries in research" [5] are relevant to "applied" and industry-sponsored/-relevant research as done in SE: For example, the greater the financial and other personal interests and prejudices in a scientific field, the less likely the findings are to be true and independent. Also, the greater the number and fewer the tested relationships in a scientific field, and the greater the flexibility in designs and definitions, the less likely the research findings are to be generalizable.

ACKNOWLEDGMENTS

We thank all researchers and practitioners who provided feedback and the anonymous reviewers for their comments.

REFERENCES

[1] M. Baker. 2016. Is there a Reproducability Crisis? *Nature*, 33: 452-454.
[2] L. Briand, D. Bianculli, S. Nejati, F. Pastore and M. Sabetzadeh. 2017. The Case for Context-driven Software Engineering Research. *IEEE Software*, 34, 5: 72-75.
[3] D. Falessi, M. A. Babar, G. Cantone and P. Kruchten. 2010. Applying Empirical Software Engineering to Software Architecture: Challenges and Lessons Learned. *Empirical Software Engineering*, 15, 3: 250-276.
[4] M. Galster and D. Weyns. 2016. Empirical Research in Software Architecture - How far have we come? In *Proceedings of the 13th Working IEEE/IFIP Conference on Software Architecture (WICSA)*. Venice, Italy, 2016, 11-20.
[5] J. Ioannidis. 2005. Why Most Published Research Findings are False. *PLoS Medicine*, 2, 8: 696-701.
[6] V. Ivanov, A. Rogers, G. Succi, J. Yi and V. Zorin. 2017. What Do Software Engineers Care About? Gaps Between Research and Practice. In *Proceedings of the 11th Joint Meeting on Foundations of Software Engineering (FSE)*. Paderborn, Germany, 2017, 890-895.
[7] D. Lo, N. Nagappan and T. Zimmermann. 2015. How Practitioners Perceive the Relevance of Software Engineering Research. In *Proceedings of the 10th Joint Meeting on Foundations of Software Engineering (FSE)*. Bergamo, Italy, 2015, 415-425.
[8] S. T. Redwine and W. E. Riddle. 1985. Software Technology Maturation. In *Proceedings of the 8th International Conference on Software Engineering (ICSE)*. London, England, 1985, 189-200.
[9] J. Siegmund, N. Siegmund and S. Apel. 2015. Views on Internal and External Validity in Empirical Software Engineering. In *Proceedings of the 37th International Conference on Software Engineering (ICSE)*. Florence, Italy, 2015, 9-19.
[10] A. Tang and R. Kazman. 2017. On the Worthiness of Software Engineering Research.
[11] C. Wohlin and A. Aurum. 2015. Towards a Decision-making Structure for Selecting a Research Design in Empirical Software Engineering. *Empirical Software Engineering*, 20, 6: 1427-1455.