Evaluation Methods and Replicability of Software Architecture Research Objects

Marco Konersmann,[1] Angelika Kaplan,[2] Thomas Kühn,[2] Robert Heinrich,[2] Anne Koziolek,[2] Ralf Reussner,[2] Jan Jürjens[3,4], Mahmood al-Doori,[3] Nicolas Boltz,[2] Marco Ehl,[3] Dominik Fuchß,[2] Katharina Großer,[3] Sebastian Hahner,[2] Jan Keim,[2] Matthias Lohr,[3] Timur Sağlam,[2] Sophie Schulz,[2] Jan-Philipp Töberg[2, §]

Abstract: Our paper at the 19th IEEE International Conference on Software Architecture (ICSA 2022) started by noticing that Software Architecture (SA) as research area experienced an increase in empirical research. Empirical research builds a sound foundation for validity and comparability. A current overview of the evaluation and replicability of SA research objects could help to discuss our empirical standards as a community. However, no current overview existed. We assessed the current state of practice of evaluating SA research objects and replication artifact provision in full technical conference papers from ICSA and the European Conference on Software Architecture (ECSA) 2017–2021. We first developed a categorization schema for SA research object evaluation and artifact provisioning. In a systematic literature review with 153 papers, we then classified the papers according to that schema. From our findings we derive and describe four proposals for improving the state of practice in evaluating SA research objects.

Keywords: software architecture research, meta-research, systematic literature review, evaluation

Motivation: Software Architecture (SA) as research area experienced an increase in empirical research [GW16], which can be considered important for the validity and comparability. Our paper [Ko22] creates an overview of the evaluation and replicability of SA research objects to help discussing our empirical standards as a community.

Research Method: We categorized SA research w.r.t. their evaluation and replicability and created an overview of the current state of practice in evaluating SA research. Therefore, we created a classification schema for the validation of SA research evaluations and conducted a systematic literature review (SLR) of 153 full technical papers published at ECSA and ICSA from 2017 to 2021. We discussed our findings and presented proposals for improvement.

Findings: Although there are valid reasons for not publishing replication packages, our results indicate that improvements of generalizability and repeatability of evaluations could enhance the field's maturity. We summarize the answers to our research questions as follows:

RQ 1: What is the distribution of research objects and their evaluation and how did their proportions change over time? SA research at the ECSA and ICSA is quite diverse w.r.t. research objects with a focus on analysis and design methods (33% of research objects).

[1] Software Engineering, RWTH Aachen University, Germany, konersmann@se-rwth.de
[2] Karlsruhe Institute of Technology, Germany, {firstname.lastname}@kit.edu, [§]uexdy@student.kit.edu
[3] University of Koblenz-Landau, Germany, {lastname,mahmoodaldoori,mehl,matthiaslohr}@uni-koblenz.de
[4] Fraunhofer Institute for Software and Systems Engineering, ISST, Germany

Case studies and technical experiments are the dominating evaluation methods. Most (58%) evaluation methods are used to measure exactly one quality. We see that neither research objects nor evaluation methods heavily changed in the past five years with a trend to more artifact provisioning since 2019.

RQ 2: How are specific research objects evaluated and how accessible are their evaluation artifacts? The most prominent way of evaluation in the investigate papers is to measure the functional suitability and performance using technical experiments and case studies. The human-centered practice architecture decision making is mostly evaluated with human-centered evaluation methods: interviews and focus groups. Few comparative methods, like benchmarks, are used. Overall, we see no clear agreement on which properties should be evaluated for specific research objects or which methods to use for specific properties.

RQ 3: Which guidelines are used for evaluation? 17% of the papers reference evaluation guidelines. 14% reference guidelines for threats to validity. The two most-referenced guidelines in both categories describe how to conduct and report case studies and how to describe their threats to validity. Overall, we can observe that guidelines are not systematically referenced in the investigated papers.

Conclusion: We derive and describe four proposals for improving the state of practice in evaluating SA research objects: (P1) to foster the generalizability of evaluation results, (P2) to develop more benchmarks to compare approaches, (P3) to foster the provision of replication packages, and (P4) to build guidelines for what and how to evaluate, and which threats to validity should be discussed for these methods. Researchers can use our results to identify recommendations on relevant properties and methods for evaluation and to find reusable artifacts to compare their approaches with existing research. Reviewers can use our results to compare the evaluation and replicability of submissions with the state of the practice.

Data Availability: We provide a replication package[5] with the tabulated and visualized review data, a BibTeX file with all papers considered, scripts for summarizing and visualizing, and a copy of a wiki that was used for collaboration of the reviewers in our SLR. All investigated papers are listed online[6].

References

[GW16] Galster, Matthias; Weyns, Danny: Empirical Research in Software Architecture: How Far have We Come? In: 13th Working IEEE/IFIP Conference on Software Architecture, WICSA 2016, Venice, Italy, April 5-8, 2016. IEEE Computer Society, pp. 11–20, 2016.

[Ko22] Konersmann, Marco; Kaplan, Angelika; Kühn, Thomas; Heinrich, Robert; Koziolek, Anne; Reussner, Ralf; Jürjens, Jan; Al-Doori, Mahmood; Boltz, Nicolas; Ehl, Marco; Fuch, Dominik; Großer, Katharina; Hahner, Sebastian; Keim, Jan; Lohr, Matthias; Sağlam, Timur; Schulz, Sophie; Töberg, Jan-Philipp: Evaluation Methods and Replicability of Software Architecture Research Objects. In: 2022 IEEE 19th International Conference on Software Architecture (ICSA). IEEE, Los Alamitos, CA, USA, pp. 157–168, March 2022.

[5] Replication Package: `https://doi.org/10.5281/zenodo.6044059`

[6] Wiki: `https://gitlab.com/SoftwareArchitectureResearch/StateOfPractice/-/wikis/Results`